Combining Granularity-based Topic-Dependent and Topic-Independent Evidences for Opinion Detection

Malik Muhammad Saad Missen

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Informatique*

Présentée et soutenue par *Malik Muhammad Saad MISSEN*
Le *07 Juin, 2011*

Titre : *Combining Granularity-Based Topic-Dependent and Topic-Independent Evidences for Opinion Detection*

JURY
Mme Brigitte Grau, LIMSI Paris, France - Rapportrice
Mme Marie Francine Moens, University of Leuvain, Belgium - Rapportrice
M Fabio Crestani, Université of Logano, Switzerland - Rapporteur
Mme Lynda Tamine-Lechani, Maître de Conférences HDR, UPS - Examinatrice
M Mohand Boughanem, Professeur UPS, France - Directeur de Thèse
Guillaume Cabanac, Maître de Conférences UPS, France - Co-encadrant
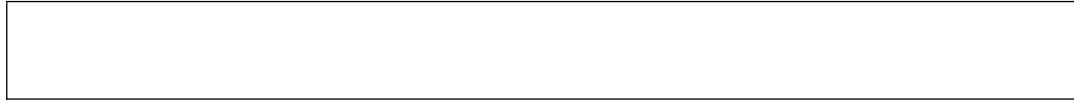

Ecole doctorale : *MITT*
Unité de recherche : *SIG-IRIT*
Directeur(s) de Thèse : *Professeur Mohand Boughanem*
Rapporteurs : *Mme Brigitte Grau (LIMSI Paris, France)*
Mme Marie Francine Moens (University of Leuvain, Belgium)
M Fabio Crestani (Université of Logano, Switzerland)

Dedication

Dedicated to my parents...

Acknowledgment

I am deeply indebted to my supervisor Prof. Mohand Boughanem, who provided me an opportunity to perform this work and for his constant support, guidance and fellowship to carry out this Ph.D thesis in his supervison, and also for helping me to have better perspective in my scientific thinking. His jokes and lively attitude left smiles on my face even during hard times during our meetings.

M. Guillaume Cabanac is especially thanked for his valuable discussion to the experimental concepts and for his constructive criticisms during the seminars. I got oppurtunity to learn a lot from your professional attitude.

I am really thankful to M. Ricardo Baeza-Yates and M. Hugo Zaragoza for giving me a chance to work at prestigious Yahoo! Labs, Barcelona for my internship. Special thanks to our team leaders M. Hugo Zaragoza and M. Roi Blanco who delivered their best to make me profit from their experience and skills. Thanks to my research fellow Gianluca Demartini for his useful cooperation.

I am really grateful to Mme Brigitte Grau (LIMSI Paris, France), Mme Marie Francine Moens (University of Leuvain, Belgium) and M. Fabio Crestani (Université of Lugano, Switzerland) for honoring me with their precious time to read this manuscript and giving their valuable feedback and also for being part of the jury. I am equally thankful to M. Claude Chrisment (Professor at University of Toulouse III, France) for being part of the jury. I really appriciate your comments, feedback and criticism while evaluating my work and I hope these will be helpful for me for my future work.

A lot of thanks to all members of my team SIG at lab IRIT for their instant help and kindness. Particularly, I would like to thank Cecile Laffaire, who has always been willing to help me to solve experimentation problems. A lot of encouragement from Mme Lynda Lechani-Tamine and Mme Karen Pinel-Sauvagnat is also appreciated. The working hours that I shared with Faiza Belbachir who worked with me in collaboration during her internship in our lab have their own worth. Thanks a lot to Faiza for her cooperation. Similarly, a bundle of thanks to Mariam Daoud (previous PhD Student) for her kind and useful suggestions that helped me to give right directions

during my thesis progress. I will never forget the beautiful moments I shared with my friends at IRIT during these 3 years especially useful discussions with Mariam, Mouna, Arlind, Lamjed, Dana, Ourdia and Dui, sharing jokes with Hussaim, Ihab, Fatima, Karim, and Najeh, useful guidance from Hamdi for . All of you are great.

My friends in Toulouse played a great role to help me get myself out of the worries caused by PhD studies or personal life. I passed unforgetful moments of my life with all of you. Good luck with your studies and life ahead.

A lot of thanks to my friends and family back in my country. My mother has been counting days for many years for my return to my home. My father (RIP), if was alive, would have been very happy for realization of his dream for his son. My sisters are waiting for me to share some time with their elder brother. My younger brother, who used to be a bit irresponsible when I was there, stood strong during my absence to support all family. My petites nieces who have always been asking me the date of my return with thier cute accent. Thanks all of you for your moral encouragement. I love you all.

I take this opportunity to express my profound gratitude to all my teachers because of whose blessings I have come so far. Last but not the least, I thank my dear parents, brother and sisters without whose encouragement and support, my PhD would have been an unfulfilled dream.

Toulouse,   17.02.2011   Malik Muhammad Saad Missen

Abstract

Opinion mining is a subdiscipline within Information Retrieval (IR) and Computational Linguistics. It refers to the computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online sources like news articles, social media comments, and other user-generated content. It is also known by many other terms like *opinion finding, opinion detection, sentiment analysis, sentiment classification, polarity detection,* etc. Defining in more specific and simpler context, opinion mining is the task of retrieving opinions on an issue as expressed by the user in the form of a query. There are many problems and challenges associated with the field of opinion mining. In this thesis, we focus on some major problems of opinion mining.

One of the foremost and major challenges of opinion mining is to find opinions specifically relevant to the given topic (query). A document can contain information about many topics at a time and it is possible that it contains opinionated text about each of the topic being discussed or about only few of them. Therefore, it becomes very important to choose topic-relevant document segments with their corresponding opinions. We approach this problem on two granularity levels, sentences and passages. In our first approach for sentence-level, we use semantic relations of WordNet to find this opinion-topic association. In our second approach for passage-level, we use more robust IR model (i.e., language model) to focus on this problem. Basic idea behind both contributions for opinion-topic association is that if a document contains more opinionated topic-relevant textual segments (i.e., sentences or passages) then it is more opinionated than a document with less opinionated topic-relevant textual segments.

Most of the machine-learning based approaches for opinion mining are domain-dependent (i.e., their performance vary from domain to domain). On the other hand, a domain or topic-independent approach is more generalized and can sustain its effectiveness across different domains. However, topic-independent approaches suffer from poor performance generally. It is a big challenge in the field of opinion mining to develop an approach which is both effective and generalized at the same time. Our contribu-

tions for this thesis include the development of such approach which combines simple heuristics-based topic-independent and topic-dependent features to find opinionated documents.

Entity-based opinion mining aims at identifying the relevant entities for a given topic and extract the opinions associated to them from a set of textual documents. However, identifying and determining the relevancy of entities is itself a big challenge for this task. In this thesis, we focus on this challenge by proposing an approach which takes into account both information from the current news article as well as from the past relevant articles in order to detect the most important entities in the current news. We look at different features at both local (document) and global (data collection) level to analyse their importance to assess the relevance of an entity. Experimentation with a machine learning algorithm shows the effectiveness of our approach by giving significant improvements over baseline.

In addition to this, we also present idea of a framework for opinion mining related tasks. This framework exploits content and social evidences of blogosphere for the tasks of opinion finding, opinion prediction and multidimensional ranking. This premature contribution lays foundations for our future work.

Evaluation of our approaches include the use of TREC Blog 2006 data collection and TREC Novelty track data collection 2004. Most of the evaluations were performed under the framework of TREC Blog track.

Resumé

Fouille des opinion, une sous-discipline dans la recherche d'information (IR) et la linguistique computationnelle, fait référence aux techniques de calcul pour l'extraction, la classification, la compréhension et l'évaluation des opinions exprimées par diverses sources de nouvelles en ligne, social commentaires des médias, et tout autre contenu généré par l'utilisateur. Il est également connu par de nombreux autres termes comme trouver l'opinion, la détection d'opinion, l'analyse des sentiments, la classification sentiment, de détection de polarité, etc. Définition dans le contexte plus spécifique et plus simple, fouille des opinion est la tâche de récupération des opinions contre son besoin aussi exprimé par l'utilisateur sous la forme d'une requête. Il ya de nombreux problèmes et défis liés à l'activité fouille des opinion. Dans cette thèse, nous nous concentrons sur quelques problèmes d'analyse d'opinion.
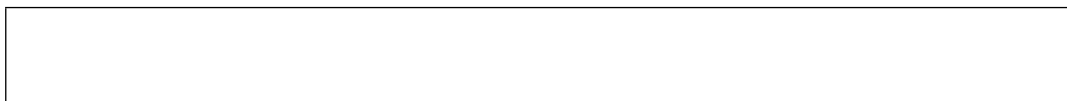
L'un des défis majeurs de fouille des opinion est de trouver des opinions concernant spécifiquement le sujet donné (requête). Un document peut contenir des informations sur de nombreux sujets à la fois et il est possible qu'elle contienne opiniâtre texte sur chacun des sujet ou sur seulement quelques-uns. Par conséquent, il devient très important de choisir les segments du document pertinentes à sujet avec leurs opinions correspondantes. Nous abordons ce problème sur deux niveaux de granularité, des phrases et des passages. Dans notre première approche de niveau de phrase, nous utilisons des relations sémantiques de WordNet pour trouver cette association entre sujet et opinion. Dans notre deuxième approche pour le niveau de passage, nous utilisons plus robuste modèle de RI i.e. la language modèle de se concentrer sur ce problème. L'idée de base derrière les deux contributions pour l'association d'opinion-sujet est que si un document contient plus segments textuels (phrases ou passages) opiniâtre et pertinentes à sujet, il est plus opiniâtre qu'un document avec moins segments textuels opiniâtre et pertinentes.

La plupart des approches d'apprentissage-machine basée à fouille des opinion sont dépendants du domaine i.e. leurs performances varient d'un domaine à d'autre. D'autre part, une approche indépendant de domaine ou un sujet est plus généralisée et peut maintenir son efficacité dans différents domaines. Cependant, les approches indépendant de domaine souffrent de mauvaises performances en général. C'est un grand défi dans le domaine de fouille des opinion à développer une approche qui est

3

plus efficace et généralisé. Nos contributions de cette thèse incluent le développement d'une approche qui utilise de simples fonctions heuristiques pour trouver des documents opiniâtre.

Fouille des opinion basée entité devient très populaire parmi les chercheurs de la communauté IR. Il vise à identifier les entités pertinentes pour un sujet donné et d'en extraire les opinions qui leur sont associées à partir d'un ensemble de documents textuels. Toutefois, l'identification et la détermination de la pertinence des entités est déjà une tâche difficile. Nous proposons un système qui prend en compte à la fois l'information de l'article de nouvelles en cours ainsi que des articles antérieurs pertinents afin de détecter les entités les plus importantes dans les nouvelles actuelles.

En plus de cela, nous présentons également notre cadre d'analyse d'opinion et tâches relieés. Ce cadre est basée sur les évidences contents et les évidences sociales de la blogosphère pour les tâches de trouver des opinions, de prévision et d'avis de classement multidimensionnel. Cette contribution d'prématurée pose les bases pour nos travaux futurs.

L'évaluation de nos méthodes comprennent l'utilisation de TREC 2006 Blog collection et de TREC Novelty track 2004 collection. La plupart des évaluations ont été réalisées dans le cadre de TREC Blog track.

Contents

List of Tables

List of Figures

15

Chapter 1

Introduction

Applying computer technology is simply finding the right wrench to pound in the correct screw.
Anonymous

With the passage of time, the World Wide Web (WWW) has been evolving in number of its users, data volume, data genre and type of services. Having reached to 1,966 millions users[1] (28.7% of the world's population) by June, 2010, the growth rate of Internet (figure 1.1) continues to increase beyond expectations [286] especially after the introduction of Web's second phase (i.e. Web 2.0) [330]. In the framework of Web 2.0, data consuming users have taken the role of data producers and daily, they are uploading huge volumes of data on the web. According to Maurice de Kunder[2], there are almost 14.87 billion pages (by August, 2010) in the total indexed web data. With this huge amount of data available to users, it becomes impossible for a user to browse through all the data collection to find some information satisfying his information need. Let us take a scenario where a computer science student is looking for information about the topic *Software Quality* on the web. Knowing the size of web, it is obvious that he will have to spend many days (if not months) to find some useful information. This does not mean that web does not contain any useful information on *Software Quality* but it is sheer volume of web data and limited capability of human being which makes this simple search task a daunting task. This is where the field of *Information Retrieval (IR)* provides its services to the users.

---

[1] http://www.internetworldstats.com/emarketing.htm
[2] http://www.worldwidewebsize.com

1.1   Information Retrieval (IR)

IR has been defined as:

Information retrieval (IR) is the process of finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [216].

It is apparent from its definition that IR is not limited to web only but it deals with the unstructured data from any data source like a company's business files, a database of published articles, or medical records of patients in a hospital, etc. The term *information need* present in this definition of IR is actually the user requirement which is expressed by the user in the textual form and it is often a combination of few words. In IR framework it is usually called as a *Query*. The systems performing the job of IR are called *IR systems*.

Generally, an IR system takes a query from the user and uses an IR model to estimate the relevancy of documents of the given collection against the provided query. At the end of the process, the user is provided with an ordered list of relevant documents ranked from most relevant document on top to least relevant documents on bottom. The best examples of IR systems available today are *Search Engines* like Google[3], Yahoo[4], etc.

Search Engines help to locate a set of relevant documents in a large document collection. With the increasing size of data on the web, it is impossible to search for relevant information without such IR systems. However it is very important to note that it is not only the size of the data that has been changing with time but a similar change has been observed in type and genre of the web data which eventually affects the associated information needs of the users. In start, Web IR search was limited only to textual data but with the availability of multimedia content (like photos, audios and videos) on the web has blatantly changed the dynamics of web search. New multimedia search engines have been developed to meet this updated information need of the users. Similarly, genre of the data available on the Web is also evolving. Earlier, web used to contain only factual information (or objective data) but now a sudden increase in opinionated information (or subjective data) has been noted and consequently user desire to access, retrieve and analyze this information is on boost. Therefore, a strong motivation is needed to satisfy the information needs targeting

---

[3]http://www.google.com/
[4]http://www.yahoo.com/
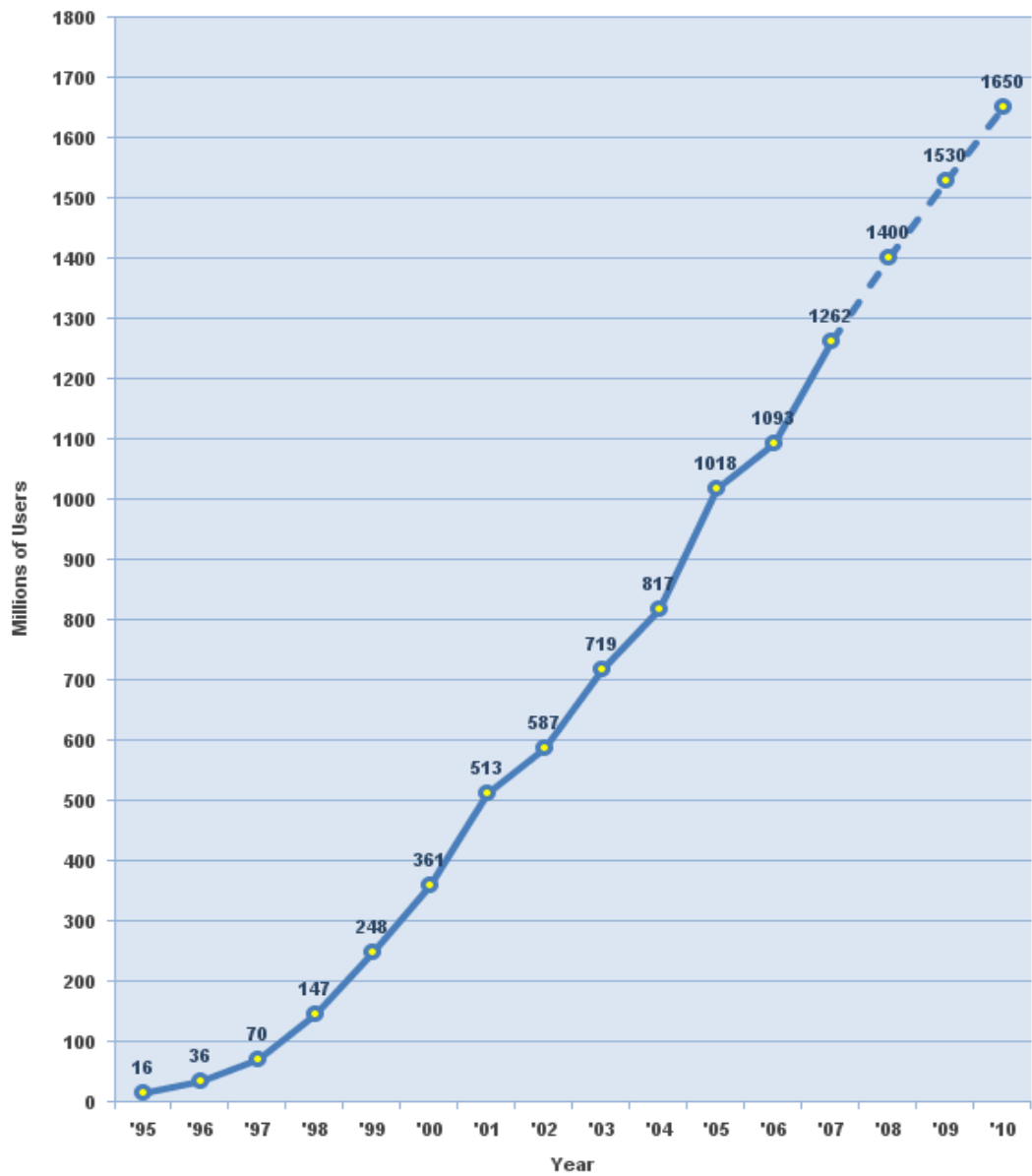
Figure 1.1: *Internet growth per year (figure from* http://www.internetworldstats.com/emarketing.htm*)*

opinionated information.

1.2   Research Problem Context

Opinionated data is one of those data genres that are increasing their presence on the web especially after the start of online social networking on web (like Facebook, MySpace, Orkut, etc.). For example, blogosphere[5] is one of the most influential social networks on the web today [272] which shares the responsibility of holding real world opinions on a variety of issues with other online social networks. Opinionated data basically represent the inner feelings or emotions or sentiments of a person about something. Opinions are dynamic (i.e., they are subjected to changes with respect to time and different individuals). A very common example of this opinion change is election results in a democratic country where win of opposition proves that public opinion for current government has changed with the passage of time and it has become more positive about its opposition.

A huge increase in popularity of online social networks has been observed over the last few years. For instance, blogs (web logs) have gained massive popularity and have become one of the most influential web social media in our times [272]. According to the Technorati[6], the blogosphere doubling its size every six months. Technorati reports that from July 2006 onwards, each passing day sees the birth of approximately 175,000 new weblogs, which means that 7,200 new blogs per hour or more than 2 blogs per second. The number of blogposts is increasing to the 1.6 million blogposts per day, meaning 18.6 posts per second.

People writing blogs (i.e., bloggers) express their opinions, thoughts and ideas about different issues (like sports, politics, etc). The readers of the blogs cannot only read the blogs but also can agree or disagree with the author by posting their comments. Blogosphere is a social network of bloggers that not only propagates factual information but a wave of opinionated information which may influence opinion of others. This also suggests that if we want to extract just opinionated information from the blogs then we have to separate the factual information from opinionated information. The opinionated and people-centric nature of the blogosphere data [133] with its continuously growing size makes it the largest source of opinions on the web and an ideal choice for searching opinions and for data analysis [159].

---

[5]Blogosphere is the collection of all blogs on Internet [212]. Blog is actually originated from the term *Web Log*. It is commonly described as a collection of web pages and is frequently updated [6, 44]. The entries of a blog are displayed in a reverse chronological order on the home page of the blog. The phenomenon of writing blogs is called *Blogging*.

[6]Technorati annual report on blogosphere's growth available at http://www.sifry.com/alerts/archives/000436.html.

With the availability of such a large collection of opinionated data on the web, what is needed is a system which takes a query from a user and retrieves him/her a list of documents containing relevant opinions. The question arises here that, *are the existing search engines not enough to satisfy the task of opinion search?*.

To search the answer for this question, let us suppose a scenario where a user intends to search for opinions of others about *President Clinton Sex Scandal*. He types this query *What People think about Clinton Scandal* in the popular search engine *Google*. What he expects from Google is a list of documents containing people's opinion about *Clinton Scandal* whether negative or positive or neutral. A more suitable presentation would be to have the results sorted by their polarity (i.e., documents with positive opinions are sorted apart from documents with negative opinions). The results returned by Google are shown in figure 1.2.

As shown in figure 1.2, almost all top most returned documents seem to contain the relevant information only (as shown by their snippets) and not the opinionated information. It shows the deficiency current search engines have for searching opinions. This example also shows that conventional search engines cannot differentiate between factual and opinionated information and therefore are not qualified for retrieving opinionated information [133, 212]. However, there are several blog search engines like *BlogPulse, BlogDigger, Technorati,*[7] etc. to search and analyze the blogosphere content but their selection criterion is also based only on data relevancy and they do not take into account the opinionated nature of the blogosphere which is one of the most important features of blogosphere data [133]. Therefore, still there is need for such IR systems that do not only take into account the relevancy of the data but also exploit its opinionated nature. The task of retrieving such documents which are not only relevant but also contain opinionated information about the subject expressed in the given topic is called *Opinion Mining* [260].

1.3 Opinion Mining

This thesis presents our work for opinion mining (also known as "opinion detection" or "opinion finding"). The basic principle behind the task of opinion mining is to differentiate between factual and opinionated information. Various evidences have been proposed by many researchers to highlight this difference and details of these evidences and approaches will be discussed in coming chapters. Opinion mining is not

---

[7]http://www.blogpulse.com/, http://www.blogdigger.com/, http://www.technorati.com/

Figure 1.2: *Google results for query "What People think about Clinton Scandal"*

a simple task though and researchers working in this domain have been facing a lot of problems that are yet to be dealt with more effective techniques and algorithms. In this section, we briefly discuss the challenges of opinion mining that our work deals with; however a detailed description can be found in chapter 4.

- A document may contain information relevant to more than one query at a time. For example, in a relevant document for a query *US Elections 2008*, we can find few sentences or even a passage about issues like war on terrorism, energy crisis, gay marriages, etc. Similarly, we can expect that this document

can have opinions about all of these issues (i.e., US elections 2008, war on terrorism, energy crisis) or may be just on few of these. In this situation, it becomes a big problem for a computational approach to find opinions on the issue as expressed by the given query. We believe that this problem can better be handled on smaller granularity levels (i.e. *sentence and passage levels*) rather than taking the whole document as a processing unit. In this thesis, we propose sentence-level and passage-level approaches for coping with this problem.

- Generally, two kinds of approaches have been proposed for opinion detection so far, first are the approaches that use query or query-based techniques (like query expansion) for detecting opinions from textual documents; second are the approaches that propose query-independent evidences for detecting opinions. It has been observed that query-dependent approaches perform better but these approaches lose their generalization in this process. In other words, performance of query-dependent approaches differ from domain to domain. On the other hand, a query-independent approach is more generalized but suffers from poor performance. In this context, an ideal situation is to have an opinion finding approach which gives better performance as well as retain its generalization. Proposing an approach which combines positive aspects of both types (i.e. performance and generalization) is a real challenge. In this thesis, we propose an approach for opinion detection that performs better than previous query-dependent or query-independent approaches.

- Entity-based opinion detection is relatively a new subject in opinion detection. It deals with the task of finding opinions about all entities relevant to the issue as expressed in the given query. For example, if the given query is *Tiger Woods Scandal* then possible candidates for relevant entities are Tiger Woods, Elin Nordegren (his wife), Rachel Uchitel, Jaimee Grubbs and finding people's opinions about these entities is the major task of entity-based opinion detection. Generally, this task is performed in three steps: 1) finding relevant entities for the given query, 2) ranking the identified relevant entities according to their relevance, and 3) finding the opinions related to those entities. Identifying a set of relevant entities is a big challenge for this task. In this thesis, we propose an effective approach for ranking entities in a news corpus yielding a significant improvement over its baseline.

- Most of the work for opinion detection have been using content-based evidences so far. Online social networks, being rich with opinionated data, are equipped with many social features that can be helpful for opinion detection task. There-

fore, an approach is needed which combines both content-based evidences and social evidences for opinion detection. However, this remains an open challenge for research community to identify and combine such features. In this thesis, we propose a preliminary framework that exploits the networked structure of blogosphere for detection and prediction of opinions in blogosphere.

Our contributions for this thesis include the approaches that focus on the challenges discussed above. In next section, we highlight our major contributions for this thesis.

1.4   Contributions

The research in this dissertation contributes to an on-going line of research in opinion mining. This thesis is a composition of some approaches that focus on different challenges of the field of opinion mining. However, our work can also be viewed in perspective of different levels of granularity (i.e., entity, sentence, passages and document level) with the objective of analyzing the effectiveness of different opinion finding techniques on different processing units.

As part of this dissertation, we propose two approaches (sentence-level and passage-level) for finding opinion-topic associations (OTA) in documents. The aim of finding opinion-topic associations is to give more importance to documents with more opinionated textual segments on the given topic. Both of these approaches take support of a novel way of query expansion wherein the given query is expanded with two types of terms (i.e., relevant and opinionated terms). However, the process of query expansion in sentence-level approach is different from passage-level approach because later one also uses proximity technique.

In our sentence-level approach, we use some simple heuristics-based features and semantic relations of WordNet [230] to compute opinion-topic associations between expanded query and sentences of a document. Each sentence is assigned an OTA score and OTA score of a document is computed on behalf of OTA scores of the sentences within it. Second contribution in this regard is the passage-based language modeling approach where we used the passages as our basic processing unit for finding opinion-topic associations in documents. In this approach, we propose to use opinion score of a term as part of the language model. The results of the experiments proved the effectiveness of passage-based

language models for the task of opinion detection.

Another useful contribution for our work related to opinion detection is the experimentation with document level topic-dependent and topic-independent features that are combined to find a combination where we can get optimal results. The use of machine learning techniques makes it possible for us successfully improving the results of various baselines.

One of the most important contributions of this work is our approach for entity ranking in news articles. The basic idea is to exploit the occurence history of entities over the timeline to estimate their relevancy status in a document. For this purpose, we annotate a data collection to extract entities of several types. This work also reports some interesting findings revealed after the analysis of the annotated data collection. This work develops a foundation for our future work for entity-based opinion detection. In addition, this work also creates a bridge with our work for opinion detection where we used a manual procedure for entity selection from Wikipedia.

We also test the usefulness of social network based evidences of the blogosphere for the task of opinion detection in our preliminary work which also forms the foundations for our future work. This work aims at testing social network based evidences of the blogosphere for various tasks like opinion detection, opinion prediction and multidimensional ranking of blog documents.

It is to be noted that even most of the approaches we propose use blogs as test data collection for experimentation; they can be easily customized for other data collections because we do not use any blog specific features in our work.

1.5   Origins of the Materials

The material that form parts of this thesis have found their origins in various conference papers we have published during the course of PhD research. In particular

⋄ Sentence-level opinion detection approach is based on the work published in [236] as a poster paper, in [235] as a regular paper and with its extension published in [238].

⋄ The challenges for sentence level opinion detection in blogs find their origin in article published in [237].

⋄ Article [240] represents the work of passage-based opinion detection in blogs.

⋄ The work related to entity ranking was published in SIGIR-2010[8] as a poster [89] and with its extension accepted at CIKM-2010[9] [90].

⋄ The work laying foundation of our future work (i.e. the discussion of social network evidences for opinion detection in blogs) can be consulted in published article [239].

⋄ Our opinion finding approach combining topic-dependent and topic-independent evidences can be consulted in article [234].

1.6   Thesis outline

This thesis has two major parts: *State of the Art* part and *Contributions* part. State of the Art part describes a detailed context of our work by giving an overview of the field of Information Retrieval (IR) (chapter 2) and Opinion Detection (chapter 3) while state of the art for opinion detection is discussed in chapter 4. Similarly, chapter 5 gives a detailed introduction about the field of Entity Ranking. In the second part of Contributions, we discuss our proposed approaches. If we further divide the contribution part then it can have four sections with each section focusing on a different problem of opinion detection as discussed above. These four sections are:

○ Section I: Opinion-Topic Association (chapter 6 & 7)

○ Section II: Dealing with Topic Dependencies (chapter 8)

○ Section III: Use of Social Evidences (chapter 9)

○ Section IV: Entity Ranking (chapter 10)

Thesis ends with a chapter about future work directions. Below we give the chapter-wise details of the thesis outline.

○ **Chapter 1 - Introduction: Current Chapter**
This chapter gives an overview of our work for this thesis by introducing our field and giving a brief description of our contributions.

○ **Part 1 - State of the Art**

○ **Chapter 2 - Information Retrieval**
This chapter presents a brief introduction to the field of Information

---

[8] http://www.sigir2010.org/
[9] http://www.yorku.ca/cikm10/

Retrieval by giving an overview of the basic concepts and technologies used to perform different IR tasks.

○ **Chapter 3 - Opinion Detection**
Chapter 3 describes the importance of opinion in a society and how focus is shifting from traditional media to social media. It also gives a detailed overview of the task of opinion detection and its applications. In chapter 3, we have also justified the selection of blogs as our source of opinions by giving statistical figures. In addition, we have discussed the TREC (Text Retrieval Conference) Blog track in detail by describing the tasks and topics and their evaluation framework.

○ **Chapter 4 - Opinion Detection: From Word to Document Level**
The importance of chapter 4 can be estimated by the fact that it discusses the related work of opinion detection in its contents. It categorizes the work and discuss it with respect to different techniques used.

○ **Chapter 5 - Entity Ranking**
This chapter gives an overview of the field of Entity Ranking and describes the related work in general. It gives details about TREC and INEX entity ranking tracks.

○ **Section I - Opinion-Topic Association**

○ **Chapter 6- Sentence Level Opinion-Topic Association in Blogs**
This chapter describes our approach for sentence-level opinion detection in detail in which we benefit from the semantic relations of Word-Net to find the association between opinion and topic. In addition, we discuss the results of a sentence-level annotation study which proves that polarity detection is a very complex task even for human beings.

○ **Chapter 7 - Passage-Based Opinion Detection**
Our passage-based approach for opinion detection in blogs is presented in this chapter. This approach basically takes advantage of the baseline provided by TREC and focus on passage level rather than whole document for opinion detection task. It demonstrates that the use of a subjectivity lexical resource and proximity approach can be beneficial for opinion mining task.

○ **Section II - Dealing with Topic Dependencies**

○ **Chapter 8 - Combining Topic-Dependent and Topic-Independent Evidences For Opinion Detection**

In this chapter, we present a mixed approach which combines topic-dependent and topic-independent evidences for opinion finding task. The experimentations with TREC Blog06 collection show that our approach gives a significant improvement over TREC provided strongest baseline.

○ **Section III - Use of Social Evidences**

○ **Chapter 9 - Social Network Exploitation for Opinion Detection in Blogs**

Chapter 9 details our work for proposal of a social network based framework for the task of opinion detection. The objective of this work is to focus on Social Network based evidences that can be exploited for the task of Opinion Detection. We propose a framework that makes use of the major elements of the blogosphere for extracting opinions from blogs. Besides this, we highlight the tasks of opinion prediction and multidimensional ranking. In addition, we also discuss the challenges that researchers might face while realizing the proposed framework. At the end, we demonstrate the importance of social networking evidences by performing experimentation.

○ **Section IV - Entity Ranking**

○ **Chapter 10 - Time-Aware Entity Retrieval**

In our work for entity ranking, we analyze and discuss some statistics about entities in news trails, unveiling some unknown findings such as the persistence of relevance over time. We focus on the task of query dependent entity retrieval over time. For this task we evaluate several features, and show that their combination significantly improves performance.

○ **Chapter 11 - Conclusions and Directions for Future Work**

In this chapter, we conclude our thesis and give a general overview of our work. In addition to this, we also give an overview of our future work directions.

Part 1

State of the Art

Chapter 2

Information Retrieval

> *An information retrieval system will tend not to be*
> *used whenever it is more painful and troublesome for*
> *a customer to have information than for him not to have it.*
>
> *Calvin N. Mooers, 1959*

2.1 Introduction

In the literature, *Information Retrieval (IR)* has been defined in many ways:

Information Retrieval (IR) is a field at the intersection of information science and computer science. It concerns itself with the indexing and retrieval of information from heterogeneous and mostly-textual information resources [135].

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [216].

Information retrieval (IR) is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information [302].

If we summarize all definitions given above then IR sums up to a process of searching and retrieving relevant documents for an information need

provided by a user. This is the most standard form of IR and is called *Ad-hoc Information Retrieval*. The information need is usually transformed to a query (or topic) which is a textual form of information need. A document is relevant if the user thinks that it contains valuable related information for his information need.

This chapter introduces the basic concepts of the IR field. Section 2.2 presents a brief history of IR by highlighting major developments in its early days while section 2.3 describes the basic working of an IR system. In section 2.4, we briefly overview major IR models like Boolean model, Vector Space Model (VSM), etc. Relevance feedback plays a vital role in the IR field and has been discussed in section 2.5. We discuss basic evaluation measures and major test data collections in section 2.6 and conclude the chapter in last section.

2.2   A Brief History of IR

Going into the history of IR system reveals the importance of work done by the famous IR researchers like Gerard Salton, Cyril W. Cleverdon, and Allen Kent, etc. in the 1960s. The *SMART (System for the Mechanical Analysis and Retrieval of Text)* Information Retrieval System [296] was one of the earliest information retrieval systems developed by Gerard Salton and his team at Cornell University in the 1960s. Many important concepts in information retrieval were developed as part of research on the SMART system, including the *Vector Space Model (VSM), Relevance Feedback, and Rocchio Classification*. Similarly, Cyril W. Cleverdon proposed a model for IR system evaluation in 1962 [69]. Later on, Salton (and McGill) introduced the concept of *Modern Information Retrieval* with heavy emphasis on vector models. Research in IR took a big turn in 1991 with the proposal of World Wide Web (WWW) by Tim Berniers-Lee at CERN (Conseil Europeen pour la Recherche Nucleaire), hence forming the foundation for *Web Information Retrieval (WIR)*.

With the introduction of the Web, many existing IR systems went obsolete because of their inability to deal with larger volume of data. A rapid progress was being seen in the Web technologies with the development of new and more powerful web protocols like Hypertext Transfer Protocol (HTTP). This rapid progress also triggered the research process in IR

field to cope with the new challenges brought up by a new and relatively bigger data collection. Existing IR systems were being improved with new techniques and algorithms. This advancement in IR research also called on prestigious IR events like *ACM SIGIR (Special Interest Group on Information Retrieval)* and *TREC (Text Retrieval Conference)*. Although the results of this IR research started to twinkle with few commercial IR systems in mid 1980s but real progress in this regard was noted in late 1990s with the beginning of development of Web Search Engines.

2.3   How IR System Works?

Generally, an IR system is supposed to support three processes [112] (i.e., indexing of the data collection, query processing, and matching of each document in the collection with the query). All these processes are represented in the figure 2.1:
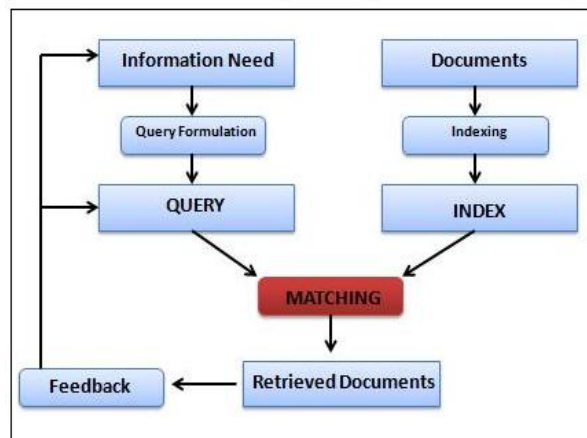
Figure 2.1: *Information Retrieval Processes [112]*

2.3.1   Indexing

Indexing is done to create the data structure called *Index* which is actual representation of the documents in IR systems and is searched for each query. More complete the index is, better are the search results. The process of indexing is the composition of many sub-processes as explained below:

- ◦ Generally in the first step, all documents in the data collection are transformed in a consistent format if they are not already and then tokenization of the documents is done. Tokenization of the document identifies the potential indexble items. A by default option can be one word but when dealing with compound words, single words lose their meanings like word *database* and words *data* & *base* convey different meanings.

- ◦ In this step, a list of very common words is prepared which is called *stopword list*. Stopword list contains all words that are not useful for retrieving the relevant documents for a given query and are removed from the documents. It includes the such as articles (a, the), conjunctions (and, but), interjections (oh, but), prepositions (in, over), pronouns (he, it), and forms of the *to be* verb (is, are). This step saves system resources like memory and processing power.

- ◦ Stemming attempts to reduce a word to its stem or root word. It reduces size of the index and improves recall by reducing all forms of the word to a base or stemmed form. For example, if a user searches for word W, he may also want all those documents that contain the variants of word W i.e. W-ing, W-es, W-ed, W-ness, etc. Therefore, the related terms are stemmed to their root (as shown in table 2.1[1]) so that documents which include various forms of W will have equal likelihood of being retrieved. Few examples of words and their roots are given below. Different stemming algorithms like Porter stemming algorithm or Lovins stemming algorithm, are available to be used [206, 277].

Word	Stem
Played	Plai
Motoring	Motor
Cats	Cat
Relational	Relat

Table 2.1: *Few examples of stemming using Porter Stemmer [277]*

- ◦ Once the stemming has been done, the indexing data structure is created which represents each document by a set terms considered important for a document. The importance of these terms is calculated by different measures which can be as simple as Boolean measure (i.e. 1 if

---

[1]http://qaa.ath.cx/porter_js_demo.html

the term is present in a document and 0 if not) and as complicated as TF/IDF (TF: Term Frequency, IDF: Inverse Document Frequency). TF of a term t is calculated by counting its occurrences in a document d and is generally represented as $tf_{t,d}$ [158]. The drawback of TF is to give equal importance to all the terms in a document as far as their relevance is concerned. However, it is a fact that certain terms are more discriminating than others to determine the relevancy of a document. For example, the term *Obama* is more discriminative than the term *elections* in a collection of documents on US elections and should be given higher score. To take this fact into account, another measure IDF was proposed which is computed as shown in equation 2.1.

$$idf_t = \log \frac{N}{df_t} \tag{2.1}$$

where N is the total number of documents in the given data collection while df_t is the document frequency of the term t which represents the number of documents in the collection that contain term t. From equation 2.1 it is obvious that IDF of a rare term is will be high whereas the IDF of a frequent term is low [217]. The combination of TF and IDF provides another very effective weighting scheme as given by equation 2.2.

$$TF/IDF = tf_{t,d} \times idf_t \tag{2.2}$$

According to Karen Sparck Jones [158], TF/IDF assigns a weight to term t in document d which is

1. highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term t occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents.

2.3.2   Query Processing

○ For query processing, same steps are repeated as for documents till the stemming step; however no index is created for queries.

○ Once the stemming of query terms is done, a proper representation of the query is prepared according to the IR model to be used for matching the query and documents. For example, if a *Vector Space Model* (discussed in next section) is to be used then query will take the form of a vector or if a Boolean matcher will be used then the system must create logical sets of the query terms connected by AND, OR, or NOT. At this point, an IR system may take the query representation and perform the search against the index. However more advanced IR system may go for further processing of the query like *query expansion* and *query term weighting*.

2.3.3   Document Query Matching

Once the indexing and query processing is finished, an IR model is used that compares the query representation and each document in the collection for relevancy between both and then a relevance score is assigned to each document. At the end of this process, user is presented with a ranked list of relevant documents on behalf of their relevance scores. In other words, an IR Model predicts what a user will find relevant given the user query [112]. In this section, we will discuss major models of information retrieval.

2.4   IR Models

IR models can be divided into two categories in context of their implementation, *Exact Match Models* and *Best Match Models*.

2.4.1   Exact Match Models

In Exact Match Models, queries are formulated using precise criteria and only the documents exactly matching the query are selected. Documents are not ranked in Exact Match Models. Here we will discuss the most popular Exact Match IR Model i.e. Boolean Model.

The Boolean Model

In Boolean Model, queries are formulated using combinations of standard Boolean operators and documents matching the query criteria are retrieved. A more formal definition of the Boolean Model goes below:

The Boolean retrieval model is a model for information retrieval in which we can pose any query which is in the form of a Boolean expression of terms, that is, in which terms are combined with the operators AND, OR, NOT. The model views each document as just a set of words [216].

Let us suppose a user who is looking for information about *Java Islands*. A Boolean model will expect this information need to be expressed by the following Boolean query to not to confuse it with *Java Programing Language*:

Java AND (Island OR Indonesia) AND NOT Programming

Only three set operations are needed to have a set of relevant documents for this query. In the first step, all those documents having the keywords *Island* or *Indonesia* are retrieved. Then this set of documents is filtered to have only those documents that have a keyword *Java* inside. Finally, the documents containing the keyword *programming* are removed from the set of documents received from previous step. The result is a set of relevant documents achieved using a standard Boolean model.

2.4.2 Best Match Models

In Best Match Models, it is not necessary for a relevant document to contain all the query terms. Instead the documents are ranked according to degree of their relevancy with the query. In this section, we discuss prominent Best Match Models.

Vector Space Model

Vector Space Model (VSM) was developed by Salton et al. [301]. In VSM, query and document both are represented in the form of vectors $\vec{q}$ and $\vec{d}$ respectively where each query term is assigned a separate dimension. The relevancy of a document d for a given query q is measured as the similarity between their associated vectors. The similarity between document and

query vectors is usually measured through the cosine of the angle between both vectors. The cosine of an angle is 0 if the vectors are orthogonal in the multidimensional space and 1 if the angle is 0 degree.

The cosine similarity of $\vec{q}$ and $\vec{d}$ is given below:

$$Cos(\vec{q}, \vec{d}) = Sim(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| \left|\vec{d}\right|} = \frac{\sum_{i=1}^{|V|} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}} \qquad (2.3)$$

where:

- q_i is the weight of the term i in the query,
- d_i is the weight of the term i in the document,
- $|\vec{q}|$ and $\left|\vec{d}\right|$ are the lengths of $\vec{q}$ and $\vec{d}$,
- $|V|$ is the total number of terms in the query.

For normalized vectors, the cosine is equivalent to the dot product or scalar product.

$$Cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_{i}^{|V| q_i, d_i} (2.4)$$

The Probabilistic Retrieval Model

Stephen Robertson and Karen Spärck-Jones based their probabilistic retrieval model on the calculation of probability of the relevancy of the document for a query [220, 294, 304]. The basic principle behind is to find such documents that have strong probability of relevancy with the query and at the same time show a weak probability of being non-relevant (i.e., given a query q and a document d, we need to calculate the probability of relevancy of this document d for the q). There are two possibilities:

- R that is the document d is relevant for the query q
- $\bar{R}$ that is the document d is not-relevant for the query q

The documents and queries are represented by Boolean vectors in n-dimensional space. An example of a document d_j and a query q is given below:

$d_j = (w_{1,j}, w_{2,j}, w_{3,j}, ..., w_{n,j})$ and $q = (w_{1,q}, w_{2,q}, w_{3,q}, ..., w_{n,q})$ with $w_{k,j} \in [0,1]$ and $w_{k,q} \in [0,1]$. The value of $w_{k,j}$ (and $w_{k,q}$) shows that if a term t_k appears in a document d_j (or q) or not. The probabilistic model helps to evaluate the probability of relevance for the document d_j for query q. A document is selected if probability of relevance of the document (denoted

as $p(R|D)$) is greater than its probability of being non-relevant (denoted as $p(\bar{R}|D)$). The similarity score RSV (i.e., Retrieval Status Value) between the document D and query q is given by equation 2.5:

$$RSV(q, D) = \frac{p(R|D)}{p(\bar{R}|D)} \tag{2.5}$$

These probabilities are calculated using conditional probabilities according to which a query term is present in a relevant document or a non-relevant document. This similarity measure can be calculated using different formulas.

Robertson et al. [293] combined the 2-Poisson model [**?**] with the probabilistic model for retrieval, to form a series of Best Match (BM) weighting models. In particular, the weight of a term t in a document is computed based on the number of documents in the collection (denoted N), the number of documents the term appears in (N_t), the number of relevant documents containing the term (r) and the number of relevant documents for the query (R):

$$w = \log \frac{(r + 0.5)/(R - r + 0.5)}{(N_t - r + 0.5)/(N - N_t - R + 0.5)} \tag{2.6}$$

However, this expression can be simplified when there is no relevance information available [81]:

$$w^{(1)} = \log \frac{N - N_t + 0.5}{N_t + 0.5} \tag{2.7}$$

which is similar to the inverse document frequency.

Language Modeling (LM)

The term "language model" refers to a probabilistic model of text (i.e., it defines a probability distribution over sequences of words). Language models are very popular and have been successfully applied in many research areas like speech recognition [152] and machine translation [48].

The use of Language Models in Information Retrieval was suggested by Ponte and Croft [274, 324]. They proposed a *query likelihood scoring method*, a new way to score a document. The basic idea behind this new method was very simple: Estimate a language model for each document in first step and then in the second step, rank documents by the likelihood

of the query according to the estimated language model of each document $P(Q|d)$. In essence, the ranking of documents is based on $P(d|Q)$. Bayes rule can be employed, such that:

$$p(d|Q) = \frac{p(Q|d)p(d)}{p(Q)} \tag{2.8}$$

In the equation 2.8, $p(Q)$ has no influence on the ranking of documents, and hence can be safely ignored. $p(d)$ is the prior belief that d is relevant to any query, and $p(Q|d)$ is the query likelihood given the document, which captures how well the document fits the particular query [35]. It is of note that instead of setting $p(d)$ to be uniform, it can be used to incorporate various query-independent document priors. However, with a uniform prior, documents are scored as $p(d|Q) \propto p(Q|d)$, hence with query Q as input, the retrieved documents are ranked based on the probability that the documents language model would generate the terms of the query, $P(Q|d)$. To estimate $p(Q|d)$, term independence is assumed, i.e. query terms are drawn identically and independently from a document:

$$p(Q|d) = \prod_{t \in Q} p(t|d)^{n(t,Q)} \tag{2.9}$$

where $n(t,Q)$ represents the number of occurrences of the term t in the query Q and is used to emphasize frequent terms in long queries. Various models can then be employed to calculate $p(t|d)$, however, it is of note that there is a sparseness problem, as a term t in the query may not be present in the document model d. To prevent this, in language modeling, the weighting models supplement and combine the document model with the collection model (the knowledge of the occurrences of a term in the entire collection) [80]. In doing so, the zero probabilities are removed, known as *smoothing*. Without this smoothing, any document not containing a query term will not be retrieved. Zhai and Lafferty [398] showed how various language models could be derived by the application of various smoothing methods, such as *Jelinek-Mercer, Dirichlet* and *Absolute discounting.*

2.5 Relevance Feedback

It is very difficult for users to formulate a query in the best possible way especially when they do not have any knowledge about the collection. It is

not easy to transform one's information need to a real world query. User
may lack vocabulary, the contexts or even the ordering of words sometimes,
etc. Like Beaulieu et al. [29] noted, *It seems that most users are not aware
of formulating their query in any particular way or able to articulate why
they have typed in particular terms. . . . . . . The majority of users tended to
start with a simple query and then react to what the system did.*

The idea of Relevance Feedback involves the user to compensate for these
lacks by re-formulating the query. The reformulation is aimed to improve
the final document ranking. The basic procedure of relevance feedback is
composed of following steps:

1. The user issues a query to IR system

2. The system returns a set ranked documents

3. The user identifies some of these returned documents as relevant or
   irrelevant (feedback)

4. The system re-computes a better representation of query on behalf of
   this feedback

5. The system creates a new ranking of documents using the newly for-
   mulated query

All above steps have been demonstrated in pictorial form in figure 2.2.
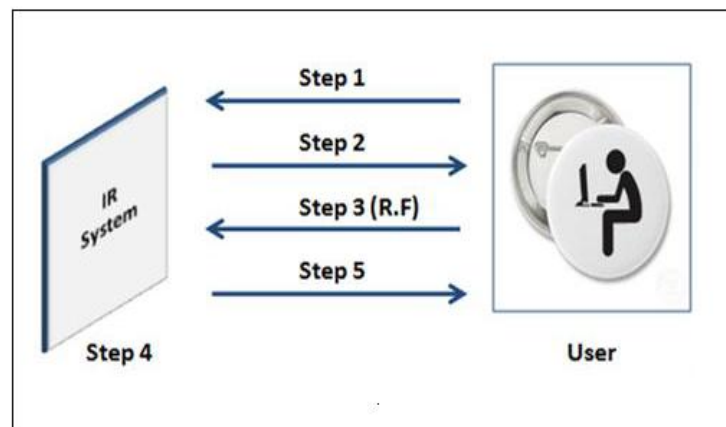
Figure 2.2: *Explaining the process of Relevance Feedback*

2.5.1   The Rocchio Algorithm for Relevance Feedback

The Rocchio algorithm is a relevance feedback algorithm which was de-
veloped using Vector Space Model (VSM) [296]. The algorithm assumes

that most users have a general conception of which documents should be denoted as relevant or non-relevant. Therefore, the user's search query is revised to include an arbitrary percentage of relevant and non-relevant documents as a means of increasing the search engine's recall, and possibly the precision as well. The formula for Rocchio Algorithm is given below:

$$Q_n = \alpha Q_i + \frac{\beta}{n_p} \sum_{n_p} D_p - \frac{\gamma}{n_{np}} \sum_{n_{np}} D_{np} \qquad (2.10)$$

where
Q_n is the vector of the new revised query
Q_i is the vector of the original query
D_p (resp.D_{np}) is the vector of a relevant document (resp. non-relevant)
n_p (resp.n_{np}) is the number of relevant documents (resp. non-relevant)
α, β and γ are the constants such that $\alpha + \beta + \gamma = 1$

A limitation of the Rocchio algorithm is that it often fails to classify multimodal classes and relationships. For example, the county of Burma was renamed to Myanmar in 1989. Therefore the two queries of *Burma* and *Myanmar* will appear much farther apart in the VSM though they both contain similar origins.

The Relevance Feedback described above is the Explicit Relevance Feedback because it required the explicit feedback from the user. However there are two variations of Relevance Feedback discussed below.

Pseudo relevance feedback

It is also known as "Blind Relevance Feedback". In this kind of relevance feedback, the system assumed that top K documents are relevant to the query and these are used to expand the query. This technique has given good performance but suffers from the problem of topic drift when top feedback documents are not relevant to the query.

Indirect relevance feedback

This type of relevance feedback is often called *Implicit Relevance Feedback*. The techniques of implicit relevance feedback unobtrusively obtain information about users by watching their natural interactions with the system [257], for example, their interaction like reading time, actions like

saving a document (or images, etc.), printing and selecting, etc. The primary advantage to using implicit techniques is that such techniques remove the cost to the user of providing feedback. Implicit feedback is less reliable than explicit feedback [253], but is more useful than pseudo relevance feedback, which contains no evidence of user judgments.

2.5.2   Relevance Feedback on the Web

The concept of Relevance Feedback has not got too much attention in Web Search market. It has been observed that very few people like to go for relevance feedback option [36]. Marchionini [219] seems true here by talking about users search behavior as:

They want answers rather than pointers . . . . . . [and they] want to achieve their goals with a minimum of cognitive load and a maximum of enjoyment.

The two major possible reasons behind this user behavior for relevance feedback are:

- The success rate of relevance feedback (i.e., how much successful it is to improve the final ranking of documents). Spink et al. [36] state: *Although it is successful 63 percent of the time, this implies a 37 percent failure rate or at least a not totally successful rate of 37 percent. . . . . . . It points to the need for an extremely high success rate before Web users consider it beneficial (326-327).*

- The design of Web IR systems to support the relevance feedback (i.e., design should be supportive instead of a burden on the user) like Jansen et al. [149] suggest, *At the very least it points to the need to tailor the interface to support these patterns if the goal is to increase the use of relevance feedback.*

2.6   Evaluation

Evaluation of an IR approach is necessary to evaluate its performance or to compare it with another IR approach. A good IR system should satisfy the needs of a user. The quality of the results with respect to the information need, system speed and the user interface are major dimensions that need to be evaluated [215]. Cranfield paradigm defines the evaluation methodology for IR systems and is based on the first information retrieval

system evaluation in the 1960s [70]. This is still the evaluation model for modern evaluation initiatives. Basically there are three basic components of the evaluation framework for IR systems [306]:

- ○ Test Data collection
- ○ A set of queries
- ○ A set of relevance assessments

In the relevance judgments provided, all documents are marked as relevant or non-relevant for each given query to create the gold standard or ground truth judgment of relevance. The size of the data collection and number of queries must be large enough because the results are highly variable over different documents and information needs. At least 25 queries are considered to be sufficient enough to make the evaluation task reliable [51]. In the next sub-section, we discuss the major test data collections available on different IR research platforms which are being used in IR research for evaluation of IR methods and techniques.

2.6.1   Test Data Collections

The Cranfield collection

Cleverdon et al. [69] work emphasized the importance of test collections with the release of a pioneering test collection (i.e., the Cranfield collection). It contained 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and relevance judgments of all (query, document) pairs.

Text REtrieval Conference (TREC)

TREC [360] was started in year 1992 with the sponsor of U.S. Department of Defense and U.S. National Institute of Standards and Technology (NIST). The objective of the TREC is to support and encourage IR by providing the necessary infrastructure for large-scale evaluation of text retrieval methodologies. By the year 2010, TREC has provided several data collections with their relevance judgments for different IR tasks. The large data collections like TREC Blog06, TREC Blog08, TREC GOV2 collection, ClueWeb09, etc have been largely used in IR research. Among these data collections, ClueWeb09 is the largest TREC data collection (25 Terabytes in size) to date.

NII Test Collections for IR Systems (NTCIR)

The NTCIR project [164, 165] almost works the same way as TREC with more focus on East Asian language and cross-language information retrieval, where the queries are made in one language against a data collection containing documents in different languages.

Cross Language Evaluation Forum (CLEF)

The objective of the Cross-Language Evaluation Forum (CLEF) [103, 221] is to promote research in the field of multilingual information access (MLIA). Different tasks have been proposed for the participants to test different aspects of mono- and cross-language information retrieval systems. The aim is to promote the research work for future multilingual multimodal information retrieval (IR) systems.

2.6.2   Evaluation Measures

The two most commonly used evaluation measures for IR are *Precision* and *Recall* [359]. *Precision* is the percentage of retrieved documents that are relevant to the given information need i.e.

$$\text{Precision} = \frac{|\text{Relevant Documents Retrieved}|}{|\text{Total Documents Retrieved}|} \qquad (2.11)$$

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved and can be expressed as given in equation 2.12:

$$\text{Recall} = \frac{|\text{Relevant Documents Retrieved}|}{|\text{Total Relevant Documents}|} \qquad (2.12)$$

The following contingency table will help to further understand these metrics.

	Relevant	Non-Relevant
Retrieved	True Positive (TP)	False Positive (FP)
Not Retrieved	False Negative (FP)	True Negative (TN)

Table 2.2: *Contingency Table*

In light of this contingency table, the above equations for Precision (P) and Recall (R) can be re-written as:

$$P = \frac{TP}{TP + FP} \qquad (2.13)$$

$$R = \frac{TP}{TP + FN} \qquad (2.14)$$

Ideally, one would like to have an IR system which gives good Precision and Recall values simultaneously. However, in practice, it is very hard to have a system which gives 100% of Precision and Recall values. Knowing that Precision and Recall values are not independent, the behavior of a system may favor the measure Precision at one time and Recall at other times. Generally, Precision-Recall curve for a system takes the shape as shown in figure 2.3.
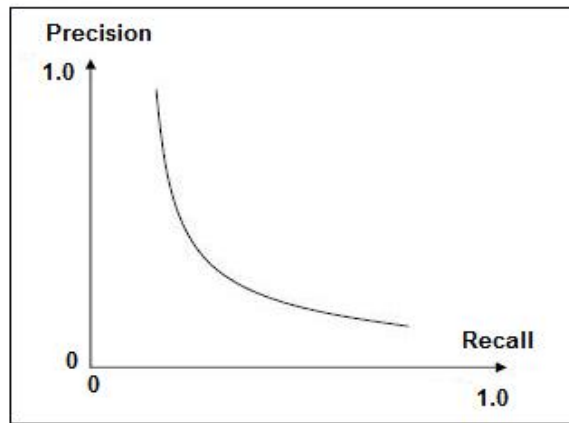
Figure 2.3: *General shape of Precision-Recall curve for an IR system*

There is another measure that actually takes advantages of P and R and combines them to give a new measure called *F-Measure(F)* [288] and is given below:

$$F = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (2.15)$$

i.e. *F*-measure is basically the harmonic mean of Precision and Recall. Precision, Recall, and the F measure are based on the set of unordered documents [216]. Therefore, there was a need of such measures that could help to evaluate a set of ranked documents (like in case of Web Search Engines). Major metrics proposed in this regards are *Average Precision (AP), Mean Average Precision (MAP)*, etc. *AP* is the average of the precision values after each relevant document is retrieved. It focuses on

relevant documents ranked higher.

$$AP = \frac{\sum_{r=1}^{N} \left(P(r) * rel(r) \right)}{}$$

$$P(r) = \frac{|\text{Relevant Retrieved Documents of rank } r \text{ or less}|}{r} \qquad (2.17)$$

Where r is the rank, N is the number of documents retrieved, $rel(r)$ is a binary function on the relevance of a given rank r and $P(r)$ is the precision at given cut-off point. Another very popular variation of this measure is *Mean Average Precision (MAP)*. For a set of test queries, MAP is the mean of the average precisions over all the test queries, is used to evaluate the overall retrieval performance of an IR system. MAP is very common evaluation measure in TREC evaluations (like in TREC Blog track for opinion finding task).

$$MAP = \frac{\sum_{q \in Q} AP_q}{|Q|} \qquad (2.18)$$

where AP_q is the average precision of query q, Q is the set of queries and $|Q|$ is the total number of queries in set Q. The evaluation measures like the *binary preference (bpref)* measure [358], *normalized Discounting Cumulative Gain (nDCG)* [150] and *inferred Average Precision (infAP)* [391], etc. can also very useful and reliable when the relevance judgments are incomplete.

2.6.3   Relevance Assessments

To evaluate an IR system, generally the results of the system are compared with relevance assessments using some evaluation tool[2] and appropriate evaluation measures. Preparing relevance assessments is a very hard and time consuming job. It involves the hiring of human beings (judges) who, given documents and queries, have to judge and label the documents with labels as defined for a particular task (like Relevant or Non-Relevant for Topical Relevance Retrieval task). For a huge collection, it becomes almost impossible to perform this job. Therefore, approach of pooling [327] is adopted in this scenario. In pooling, assessment is done over a subset of the collection that is formed from the top documents returned by a number of different IR systems (usually the ones to be evaluated), and

---

[2]For instance, evaluation tool trec_eval: information about it is available on `http://ir.iit.edu/~dagr/cs529/files/project_files/trec_eval_desc.htm`

perhaps other sources such as the results of Boolean keyword searches or documents found by expert searchers in an interactive process [216].

However, these relevance judgments are not 100 percent reliable. Different judges can judge the same document differently (i.e., one may label a document d as relevant and other as non-relevant for the same query q). However, it is interesting to consider and measure how much agreement between judges exists on judgments. In IR domain, a common measure for checking the degree of agreement between judges is the kappa measure [72, 185]. Kappa measure is a statistical measure of inter-annotators agreement for qualitative (categorical) items and mathematically can be written as:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{2.19}$$

where $P(A)$ is the proportion of the times the judges agreed, and $P(E)$ is the proportion of the times they would be expected to agree by chance. The interpretation of different values of kappa falling in the interval $[-1, 1]$ is shown in table 2.3.

Table 2.3: *Kappa Value Interpretations [185]*

Kappa Value (κ)	Interpretation
Below 0.0	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

2.7   Chapter Summary

This chapter provides a brief introduction to the field of *Information Retrieval (IR)*. Starting with the definitions of IR, it gives an overview of it's history and basic setup in which IR system use to work. Later on, we discuss the most important elements of an IR system (i.e., IR Models). The chapter ends with the introduction of few basic IR concepts and IR evaluation framework. In the next chapter, we will explore the problem of opinion detection in detail.

Chapter 3

Opinion Detection

Public opinion, though often formed upon a wrong basis,
yet generally has a strong underlying sense of justice.
Abraham Lincoln

3.1   Introduction

Opinions are very important in our daily lives. They help an individual to analyze a situation from many aspects and take an appropriate decision. The opinion of one individual may influence another individual's opinion and if this process continues can give birth to an opinion on mass-level (i.e., public opinion).

Public opinion is considered equally important in every domain. In democratic societies, governments keep public opinions on top while planning their policies for different national or global issues. Similarly, commercial product manufacturers keep their eyes on the opinion of general public while marketing and de-marketing their products. On the other hand, the customers also inquire others for the products that others have already used to help them decide which product to buy. Another example can be a person who is planning to spend some time in cinema on incoming weekend seeks for other's opinion or suggestions to choose a good movie.

This chapter focuses on the problem of extracting opinions from documents. It highlights the importance of opinions in a society and explains that how Internet is valuable for expression of opinions (section 3.2). In section 3.3, we give a detailed overview of the process of opinion detection.

Section 3.4 described the importance of opinion detection by discussing
its major applications. In section 3.5, we discuss few characteristics of
blogs that make blogs an ideal data collection for performing opinion re-
lated tasks like opinion detection and analysis. Evaluation framework for
opinion detection approaches has been discussed in section 3.6.

3.2 Internet: A Medium of Opinion Expression

Before Internet, people used to conduct surveys and ask face-to-face ques-
tions to collect and analyze other's opinion about something but Internet
has made the task of collecting opinions much easier than it used to be.
Anyone having Internet access can give his/her opinion about any subject
he/she is interested in; hence contributing to the large volume of informa-
tion about online products, political issues, movie reviews, etc., present in
the form of facts and opinions. This availability of useful information is a
motivation behind many interesting search and browsing behaviors like an
increasing trend in online shopping. The *Office of Fair Trading (OFT)* is a
UK's consumer and competition authority. It reported a comparison of its
two telephonic surveys conducted in Nov, 2006 (on 1,003 UK consumers)
and Jan, 2009 (on 1,001 UK consumers) [258]. It describes the reasons
(shown in table 3.1) UK consumer describe for buying products online and
one of the major reasons reported is the availability of enough information
about the products on online shopping sites. Generally, this information is
present in the form of customer opinions/reviews on online shopping sites
(as shown the number of customer reviews in figure 3.1).

Similar to online shopping, an increase has also been observed in people's
interest to express their opinions about political issues and this is what
has motivated politicians to adopt Internet communication in addition
to the conventional methods of performing political activities. *Arianna
Huffington*, editor-in-chief of *The Huffington Post*, says about President
Obama's election campaign [228]:

*Were it not for the Internet, Barack Obama would not be president. Were
it not for the Internet, Barack Obama would not have been the nominee*

Similarly Claire Cain Miller [228] reports that Mr. Trippi, a campaigner
for President Obama during elections, states about Howard Deans 2004
campaign:

Table 3.1: *Motivating factors for online shopping [258]*

	Nov, 2006	Jan, 2009
Wider choice / can compare prices	74%	85%
Find what you want more quickly / saves time / quick and easy	80%	84%
Shop in comfort / can stay at home	78%	81%
More product information to help make decisions	61%	72%
Can buy products not available in the UK	46%	56%

Figure 3.1: *Image showing link to Customer Reviews for a Digital Camera on www.amazon.com*

> *The campaigns official stuff they created for YouTube[1] was watched for*
> *14.5 million hours, to buy 14.5 million hours on broadcast TV is 47*
> *million Dollars.*

From the above discussion we can conclude that Internet has become not only an effective but a very economical medium for opinion expression especially after the launch of *Online Social Networking Services*.

3.2.1   Opinions and Online Social Networking

Since their launch, Online Social Networking Services have attracted millions of users. These social networks help strangers connect based on their shared interests, political views, or other activities. Some sites offer their services to diverse audiences, while others restrict their members based on factors like language, race, gender, religion, or nationalities. Generally, a user creates her/his profile by adding information about his education, location, and gender, etc. Later on, she/he can add friends and share her/his interests or hobbies with other people in her/his network through services like private messaging or instant messaging. Facebook, Twitter, myspace, blogger[2] are the most popular social networking sites available on the Web. Grunwald Associates LLC [13] conducted a survey with the support of Microsoft, News Corporation and Verizon on 1,277 students (9 to 17-year old) to reveal interesting information about the usage of social networking services. Few of the reported findings reported are:

- 96% of the students with online access report that they have ever used any social networking technology, such as chatting, text-messaging, blogging and visiting online communities, such as Facebook, MySpace, etc.

- 81% say they have visited a social networking Web site within the past three months

- 71% say they use social networking tools at least weekly

- 39% of the non-conformists students (who do not respect safety rules) who recommend products frequently and keep up with the latest brands compared to 27% of others (who respect rules).

---

[1] http://www.youtube.com/
[2] http://www.facebook.com/,   http://www.twitter.com/,   http://www.myspace.com/, www.blogger.com/

Note: This study [13] was conducted in July 2007 and figures must have been increased by now with the increase in Internet access with time. Like The *CIA World Factbook* [4] estimates that, worldwide, in 2005 over one billion people had Internet access and The Computer Industry Almanac [71] suggests that by 2010 there will 1.8 billion.

Public has started preferring online social media (like blogs, tweeting, etc.) [285] over conventional media (like Newspapers, TV, Radio, etc.) because the earlier one provides them the opportunity to read news as well as people's opinion about news while later one provides only news. Reinstein [285] beautifully describes the process of transfer of interest from conventional media to social media in pictorial form in figure 3.2.
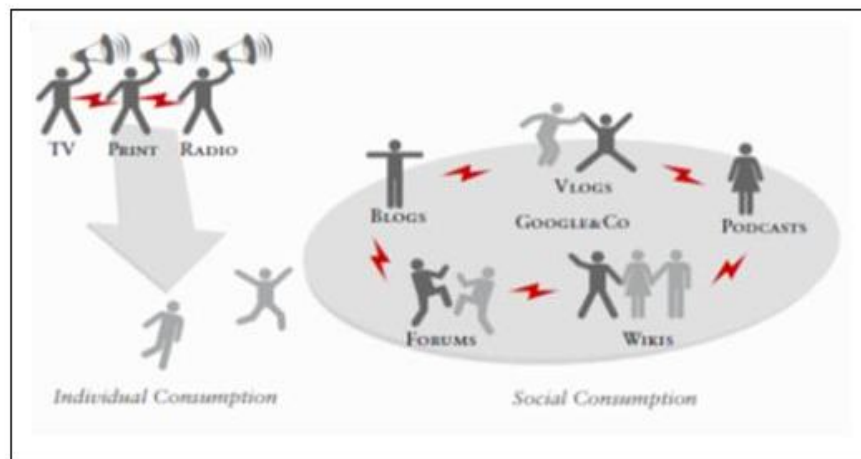
Figure 3.2: *Today: old media loses its audience to social media [285]*

Even the work of Reinstein is not that old but his revelations have already started to appear in industry as real commercial products. For example, the Google TV Project[3] from Google Corporation is a mélange of TV and Web as Google says itself:

TV meets web. Web meets TV

It can be considered as an interpretation of figure 3.3 where old media (TV in this case) will become part of web (social media). Interestingly, this trend has also been observed in the statistical results of surveys conducted. For example, Gillmor [111] argues that blogging has converted the once *Read Only Media* to more a *Conversation or Seminar*. Similarly in August

---

[3]http://www.google.com/tv/
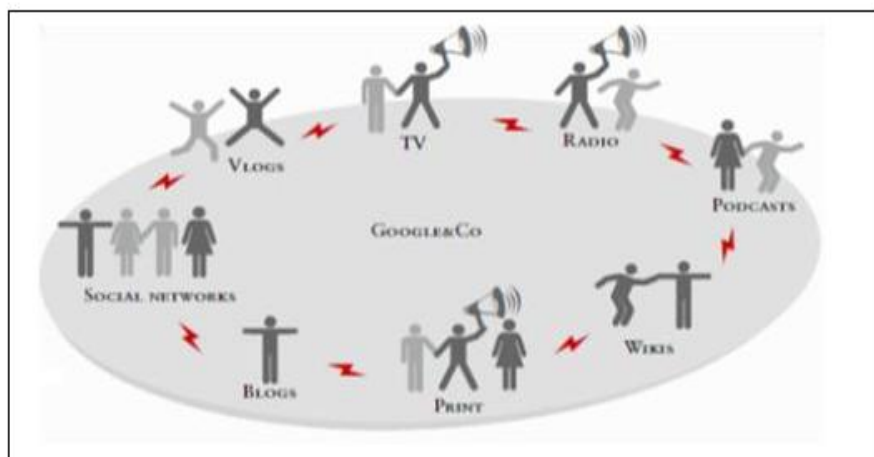
Figure 3.3: *Tomorrow: old media becomes part of the social media [285]*

2004 during the US presidential election, 28 million site visits were made to the ten most popular political blogs which is almost equivalent to the total audience for America's three online cable news network's [176]. In August alone the leading liberal blog, DailyKos, had seven million reader-visits, topping the 5.7 million audiences for Fox News.

With the increasing size of Web, the volume of opinionated information on the Web is also increasing. Given this large collection of Web Data, what if a user wants to have opinions of others on a topic he is interested in? Obviously, he has two choices:

1. To browse through the whole collection, search and extract the opinions by manually reading all the collection,

2. To use an Information Retrieval system (like Web Search Engines) to get a set of relevant opinions about the topic. This process is called *Opinion Mining*.

Obviously first choice seems intractable due to the sheer size of data collection and limited capability of the user. Therefore, the only choice left is the second one (i.e., to use an IR system to retrieve opinions about the product). Now the question arises, *Can existing classical search engines be used to satisfy this information need of the user?* In next section, we will try to find answer of this question with much more details about the process of "opinion mining".

3.3   Opinion Mining

Formally defining, *Opinion Mining is the process of extracting opinions from text documents* [202]. Opinion mining is also called "opinion finding" or "opinion extraction" or "opinion detection". Besides these, words like *sentiment* and *emotion* have also been used for opinions but more in the context of determining the polarities of the opinions (i.e., negative, positive or neutral). To better understand the definition of opinion mining, let us try to understand an opinion itself. Chaffee and Price [357] define opinions as: *Observable verbal responses to an issue or questions.* Another commonly found definition for opinion is: *An opinion is a statement that a person believes to be true but it cannot be measured against an objective standard.* Bethard et al. [37] defines opinion as: *A sentence, or a part of a sentence, that would answer the question, How does X feel about Y?*

It suggests that opinions are subjective to corresponding individual (i.e., if another individual is asked about the same issue then he/she might give a different opinion). For example, many people will agree with the following statement:

<center>*This colour is too bright to suit you*</center>

given by a person X but few may disagree too because there is no standard defined for *the best colour* for a specific person. On the other hand, fact can be described as: *Fact is a statement that can be proven true (or false) with some objective standard.* For example, the statement:

<center>*July 14 celebrates France National Day*</center>

This is a statement that can be validated however. These distinctions between factual and opinion information help researchers to extract opinions from documents.

Major tools used for searching information on the web are Search Engines like Google, Yahoo, etc. but they are more focused to retrieve factual information rather than opinion information [203] (it gives answer of the question raised at end of the previous section). Pang et al. [264] differentiate the treatment of opinionated text from classic text mining and fact-based analysis. According to them, traditionally text classification seeks to classify documents by topic. While dealing with topics, we can have as few

as two classes (like Relevant and Non-Relevant) or as many as thousands of classes (i.e., when classifying w.r.t. a taxonomy) for text classification. But in case of classifying opinions, generally we have few classes (like positive, negative or neutral, etc.). In addition, while dealing with topic-based categorization, different classes can be unrelated to each other but as far as opinion-based categorization is concerned (in light of its past work), the classes for categorization are always related somehow (i.e., whether they are opposite or they have some ordinal relation between them). Similarly, Tang et al. [337] and Ku et al. [179] argue that opinion retrieval is different from conventional topic-based retrieval. Adding further to arguments in this regard, Huifeng et al. [337] emphasize that opinion detection requires more robust techniques to be used than topic-based retrieval especially to associate opinions to corresponding topic while Lun-Wei and Hsin-Hsi [179] argue that topic-based retrieval (or conventional information retrieval) only focuses on retrieving relevant documents for the topic and do not report any positive or negative sentiments associated with the topic.

Researchers from *Cognitive & Psychological* and *Information Retrieval* domains have been working for the problem of opinion detection. While working on the cognitive level, we have come across different *Affect and Emotion Theories* like the discrete approach of Ekman [97], the dimensional approach of Russel [299] and the appraisal approach of Martin et al. [222]. On the other hand, the basic idea behind the IR approaches for opinion detection is to recognize the subjective information within these documents and the process of analyzing the subjective information is called *Subjectivity Analysis*.

Subjectivity Analysis is a problem closely associated with the problem of Opinion Detection [337]. It involves the identification of private states being expressed and also identification of their attributes [382]. Attributes of private states include who is expressing the private state, the type of attitude being expressed, about whom or what the private state is being expressed, the intensity of the private state, etc. For example given the sentence,

The Pakistani Cricket Team Coach Waqar Younis praised the players for their performance in World T-2O 2010 tournament

In above sentence, it is the coach *Waqar Younis* who is expressing a private state and the private state being expressed is indicated by the expression

praised the players. The type of attitude in the sentence seems positive and private state is being expressed for players of the team.

Just to give an example of how people are used to express their opinions on the Web, we give an example here of a comment posted on a blogpost of a famous blog[4] on issue of *Gulf of Mexico Oil Spill*:

I really can't understand why most industry officials and politicians talk about the oil spill as if there is only ONE problem; (1) capping the well. There are TWO problems: (1) capping the well and (2) dealing with the oil that has already gushed from the well. I completely agree that BP and/or the oil industry is the only entity with the technology that is capable of handling the first problem. However, the government is more than qualified to handle the second problem.

Figure 3.4: *An example opinion of a blog reader*

Similarly a digital camera review example taken from a Digital Camera review site[5] is given below:

3.4 Applications of Opinion Mining

In this section, we discuss the applications of opinion detection.

1. **Products Review Mining** Billions of people are expressing their opinions about different products on blogs or product specific review sites[6]. On the other hand, countless number of people are also consulting internet to search for people opinions about the products they are interested in buying. An ideal product mining tool will provide polarity-based categorized lists of opinions (i.e., list of negative opinions, list of positive or list of mixed opinions) for each feature of a given product. Researchers are trying their best to develop close-to-ideal (if not ideal) match of such mining tool. Feature-based product mining is one of the best applications of opinion mining. Generally it involves the extraction of a given product features and then retrieving and classifying (on behalf of its polarity, i.e., negative or positive) one or more review sentences for each product feature [83, 139]. Going one step further, few works (like [155, 204]) have also provided the features based comparison between different products. For example,

---

[4]http://www.huffingtonpost.com
[5]http://www.bhphotovideo.com
[6]like http://www.epinion.com/, http://www.wize.com/

Posted on: 5/18/2010 by XXXXX
Pros: Good Image Stabilization, Great Zoom, Nice Body Color, Short Lag Time
Best Uses: Family Photos, Landscape/Scenery, Sports/Action, Travel
Describe Yourself: Photo Enthusiast
Bottom Line: Yes, I would recommend this to a friend

My friends always come to me for advice when buying a camera due to my experience.
Most of the time they will ask me to buy and test it for them.

Of course, I cannot and will not compare a point and shoot camera to a BRAND-MODEL
but I will rate it fairly with my judgment on the price and performance.

I am really impressed and just noted few stuff.
1. AF is really fast and accurate.
2. Zoom range from wide to long.
3. WB is fairly accurate.
4. OIS really works.
5. HD video is smooth.
6. Nice quick menu controls.
7. Nice LCD resulution.

Quick note - I posted few sample shots on Facebook for the owner to see and it received
comments saying that it is impressive and asking what kind of lens is mounted. I say it is
a point and shoot.

Cons: I think I have to pop the flash manually which I prefer but may not be good for
others.
I gave it to the new owner so maybe there is a setting to auto pop it.

Figure 3.5: *Example of a review of Digital Camera taken from a product review site*

Liu et al. [204] propose a prototype system called *Opinion Observer*, which provides a visual comparison of two given products (see figure 3.6). For both products, it extracts the common product features and then shows their score on positive-negative scale.
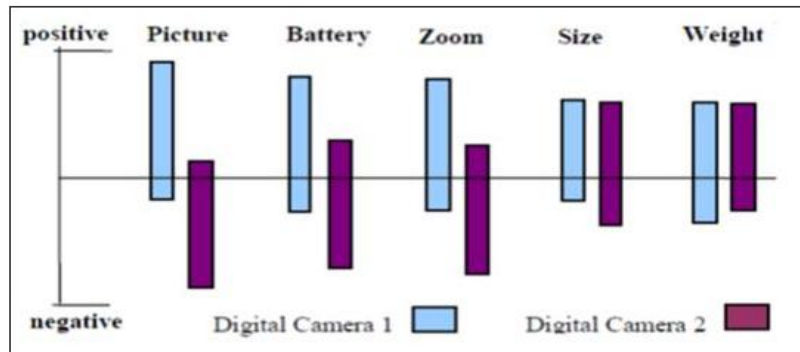
Figure 3.6: *Comparisons of features of two different digital cameras*

2. **Opinion Summarization** The task of Opinion Summarization is different from traditional text summarization. Opinion summarization is more about producing a sentiment summary from subjective sentences of an opinionated document [407]. While most of the work related to opinion Summarization is limited to Products Reviews, there are few works however for other domains [74, 407]. As far as the work for product review summaries are concerned, it is exactly the same as we discussed above in sub-section Product Review Mining except with the addition of last step where the system is supposed to produce a summary using the information it has collected [139, 140]. This short product summary can be really helpful for the reader especially when given product reviews are too long to read and make a decision of buying it. Li et al. [407] generate a summary for movie reviews in the form of positive and negative sentences for each feature class mined for that film. Similarly Conrad et al. [74] summarize the sentiments in legal blogosphere.

3. **Intelligent Market Reviews**
While on the one hand product consumers are getting benefit of opinion mining, on the other hand the product manufacturers are also taking its advantage. With the use of a blog opinion mining tool, product manufacturers can not only know about current acceptability rate of their products among public but can also can know the sta-

tus of their competitors. Opinion mining in blogs reveals the future trends forming among the public which helps product manufacturers to keep public interests in mind while creating new version of their products. These intelligent market surveys conducted with the help of opinion mining tools have produced a new wave of competition among manufacturing organizations.

4. **Trend Analysis** People use weblogs (or blogs in short) to express their thoughts, opinion or ideas that make blogs an ideal source to track trends over time. Blog sites like BlogPulse and Google Trends provides such trend analysis services like BlogPulse use the percentage of all posts concerned with a topic to show the trends in blogs [160].
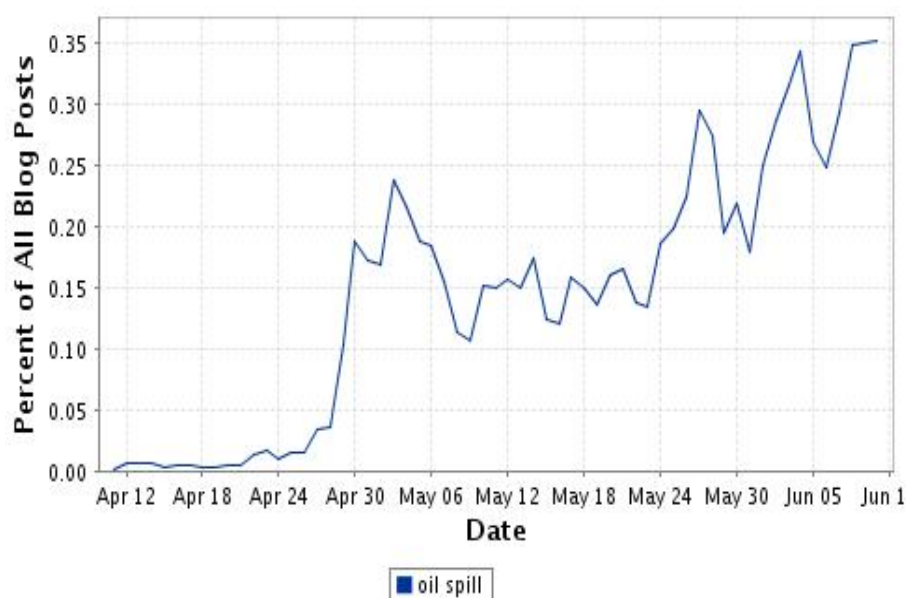
Figure 3.7: *Graph showing Trend Analysis over time*

Figure 3.7 taken from *www.BlogPulse.com* shows the trend analysis for term oil spill during two months April-2010 to June-2010. This figure shows a sudden increase in the number of relevant blogposts after 20th April because of the Gulf of Mexico oil Spill caused by an explosion on 20th April, 2010. A lot of literature work exists on trend analysis in blogosphere [25, 63]. For example, BlogScope [25] is a system used for spatio-temporal analysis of blogs, flexible navigation of the blogosphere and keyword correlation, etc.

3.5 Blogs as Data Source

In our thesis, we decided to choose Blogs as our data collection for the task of opinion detection because blogs present a tempting source of qualitative data. A *Blog (or Web Log)* is a collection of entries added frequently by an individual or a group of individuals and appear in reverse chronological order [3]. Similarly Winer [383] provides a more technical definition of blogs as: *Weblog is a hierarchy of text, images, media objects and data, arranged chronologically, that can be viewed in an HTML browser.* The individuals writing blogs are called *Bloggers* and each entry in a blog is called a *Blogpost*. Figure 3.8 shows the structure of a blog.

Blogging is the phenomenon of writing blogs and *Blogosphere* a social network of bloggers (or blogs). The opinionated and evolving nature of the blogosphere stands it unique among other online social networks. William Quick [280] defines the Blogosphere as: *It is the intellectual cyberspace that bloggers (i.e., those who write blogs) occupy* and many have [3, 159] referred to Blogosphere as a *universe of all Blogs*. A survey report by Technorati [339] demonstrates the increasing trend in volume of blogosphere and number of blogposts per day (figures 3.9 and 3.10).

The hyperlink connections between blogs are what form the base for structuring of the blogosphere. When one blog links to another one, the readers of the former blog are more likely to read the latter by following that link than they would have been otherwise [102]. *Blogroll* is a section of the blog where bloggers provide links to different blogs they read frequently or which talks of their common interests. In other words, blogroll represents the interests and preferences of bloggers. Tadanobu et al. [105] define four kinds of relations that can exist between two blogs:

○ **Citation:** Blog A and blog B are said to have a citation relation from A to B if an entry of blog A includes a hyperlink to blog B and vice versa.

○ **Blogroll:** If a blog A lists a blog B in its blogroll then there exists a blogroll relation between blogs A and B from A to B.

○ **Comment:** If the blogger of blog A comments on blog B then there develops a comment relation between two blogs from blog A to blog B.

○ **Trackback:** A trackback relation from A to B exists if an entry of blog B contains a back-reference by the trackback function to blog A

According to Daniel et al. [102], individual blogs can be considered as nodes while links between them as edges of the networked structure of blogosphere. The number of links to a particular blog is degree of that blog. Sometimes this link structure of blogosphere results in creation of small communities of blogs. Juan et al. [227] define a community as a set of blogs that have stronger relationships among them than rest of the sites of the same class. The discovery of online communities within blogosphere and analysis of information propagation is an interesting problem people have been working on for years [65, 200].
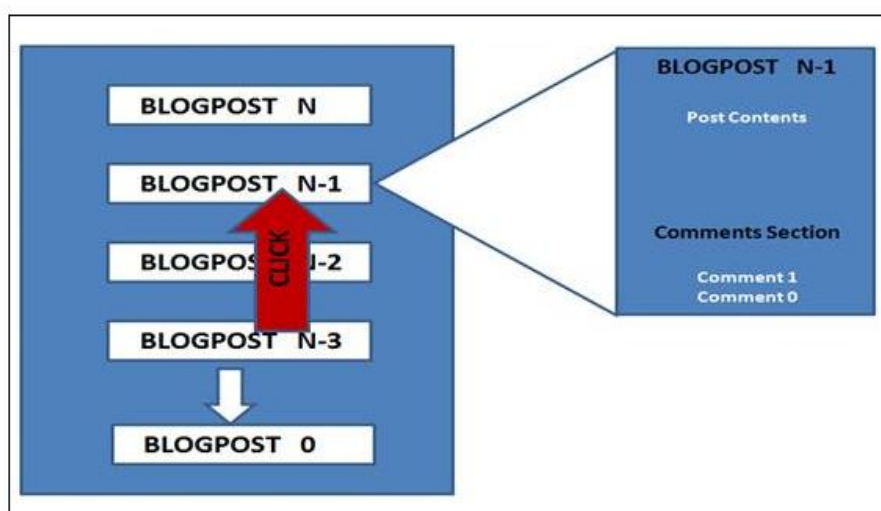
Figure 3.8: *The basic structure of a blog*

In [364], Warid and Cahill argue that a number of bloggers and people reading blogs are increasing. They say that easy availability of Internet and blog creation has provided a complete publishing platform to Internet users where they not only can produce content but can also market it. Kumar et al. [181] also observed a dramatic increase in connectedness and in local-scale community structure of Blogspace.

With the growing popularity of blogging among public, many opinion journals (like *The New Republic, The American Prospect, etc.*) and newspapers (like FOX News, ABC News, etc.) have developed their own blogs [102] which not only attract a great number of audience but also collect their opinions posted in the form of comments on blogposts. Each time a new
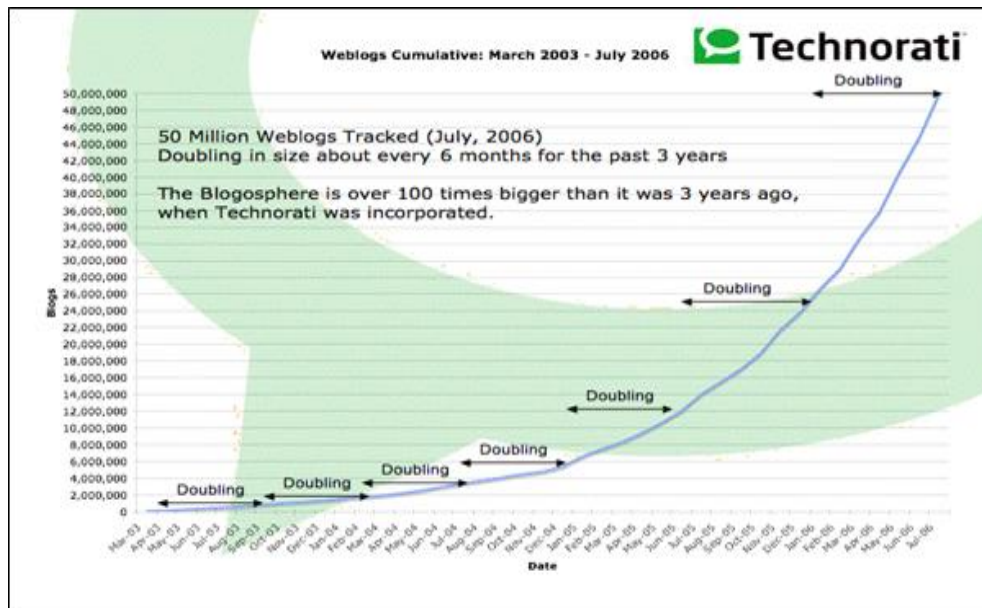
Figure 3.9: *Increase in volume of blogosphere [339]*
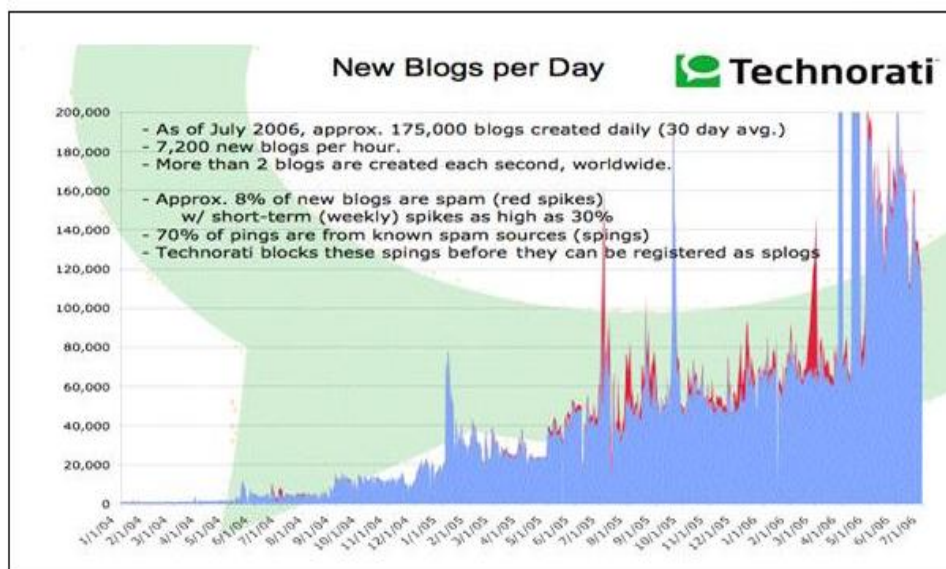
Figure 3.10: *Increase in number of blogposts [339]*

news story is published, bloggers emerge to link to it and a public discussion starts on the subject of the story. The public never had this level of freedom of expression that blogging has provided. One can find all kinds of opinions (i.e., positive, negative or neutral) on a variety of subjects being discussed in blogs. The topic range of blogs is very diverse [310] covering topics like politics, sports, education, drugs, health, literature, research, computer gaming and many other social issues.

Besides their increasing popularity and volume, blogs provide many benefits when considered as a source of data analysis [159].

- One of the major advantage of using blogs as a data collection is that they avoid the typical hard process of data collection (i.e., interviews, focus groups or surveys, etc.). Blogs, which are readily available in a electronic format with a separate document for each blogpost, offer an immediate availability of rich, codified data in an efficient package pre-prepared for analysis; hence are more convenient and less time and resource demanding for their collection. Another source of convenience can come from the content of the blog which is often categorized by the blogger according to their assignment of topic. This enables faster and easier access to information of greatest value.

- Blogs represent rich and deep personal interests of the blogger. This richness and depth is a result of the freedom in writers topic selection. Since bloggers choose their own topics, it is natural their choices reflect their areas of interest. Their writing of these issues is opinionated and often unbiased, as they are free to express their own views, expecting no tangible consequences. In addition, the motivation to express themselves forces bloggers to create their genuine profiles that provide valuable insight into the issues present on bloggers minds.

- Blog data is by nature primary data which is not subject to the influence or interference of the researcher. It is therefore away from the weaknesses and biases of many other forms of data collection, whether face-to-face or remotely collected. Typical of these influences is that of the Hawthorne Effect [124]. The Hawthorne Effect in particular causes respondents to provide or accentuate data they think will please the interviewer. Blog contents, however, are unaffected by the researcher, assuming that the blogger is not aware of the research when writing.

Another major reason behind choosing the blogs as a source of data collection is nonexistence of an effective IR system for opinion search in blogs,

considered one of the richest sources of opinions. There are many IR sys-
tems available for general web search like Google, Yahoo, Altavista, etc.
but Mishne et al. [233] are convinced that blog searches are different from
general web search with blog search more oriented towards theme and
entity-based search. They also identify the technology, entertainment and
politics as interest areas of blog searchers. All this motivates the research
industry to propose approaches for opinion search in blogs. Nowadays,
there exist many special blog search engines (like BlogPulse, BlogLines,
Technorati, etc.) and also major search engines like *Google, Yahoo, etc.*
have started to provide blog search services. But the blog search offered
(for example by Google) is more like web news search [133]. Technorati
does well with blog search by providing blogposts in topical categories but
remains focussed on relevancy only and do not takes into account the opin-
ionated nature of the blogs. Therefore there is a need for opinion search
systems in blogs. The work related to opinion detection dates back to late
1980s and early 1990s but with different titles and objectives. This work
was more limited to measuring other behavioral patterns like Aggression,
Politeness, hostility, etc. in text [47, 147, 328]. This trend was shifted to
the analysis of Subjectivity in late 1990s [49, 126, 175]. The actual work
on opinion detection for extracting opinions appeared as an emerging field
in early 2000s [83, 338, 343, 347, 372]. The earlier work was more lim-
ited to news corpuses or email messages while later on movie reviews and
discussion board's messages replaced the earlier corpuses with the spread
of forums and review sites. It was year 2003 when researchers recognised
the importance of blogs and an era of blog research began [181, 255, 311].
TREC and NTCIR took initiatives by starting special tracks for opinion
search in blogs.

The third major reason behind our decision of choosing blogs as a source
for the task of opinion detection is the presence of standard data collections
released by TREC and NTCIR (however our work limits itself to TREC
Collections). In next section of evaluation of opinion detection approaches,
we discuss the TREC and NTCIR opinion finding tracks in detail.

3.6 Evaluation

Most of the approaches for opinion finding have used *Precision (P) and Mean Average Precision (MAP)* measures for evaluation of their approaches that have already been defined in chapter 2 in evaluation section. However, in all equations related to precision, the word "relevant" should be considered as word "opinion" when talking about opinion finding task i.e. relevant documents are in fact opinionated documents.

In remainder of this chapter, we will discuss two prestigious evaluation campaigns for opinion finding organised by TREC and NTCIR.

3.6.1 TREC Blog Track

Considering the widespread authority of Weblogs in the World Wide Web (WWW), the prestigious TREC (Text Retrieval Conference) initiated a Blog track in year 2006 with the aim of providing a standard evaluation platform for approaches of opinion finding in blogs. Till now, TREC has released two data collections which are real snapshots of the blogosphere. Both collections are accompanied with query relevance judgments (qrels). The details of both collections, tasks defined for the track and other details are discussed in this section.

TREC Blog Data Collections

TREC Blog 2006 Data Collection   The test data collection TREC Blog 2006 released by TREC in year 2006 for Blog track was created in three steps [212]:

1. The selection of suitable blogs to crawl

2. Fetching the appropriate content from the web

3. Organizing the collection into a useable form

The selection of blogs was done using a list of top blogs from a famous blog specific search engine and by a manual selection of blogs related to different domains like sports, politics, health etc (US and UK only). Spam blogs (called splogs) were also made part of the collection to give a real world scenario to the participants. While most of the collection contains

English Language Blogs but a small portion of non-English blogs have been
included too. Further details about the TREC Blog06 data collection are
given in table 3.2.

Characteristic	Value
Number of Unique Blogs	100, 649
RSS	62%
Atom	38%
First Feed Crawl	06/12/2005
Last Feed Crawl	21/02/2006
Number of feed Fetches	753,681
Number of Permalinks	3,215,171
Number of Homepages	324, 880
Total Compressed size	25 GB
Total Uncompressed size	148 GB
Feeds (Uncompressed)	38.6 GB
Permalinks (Uncompressed)	88.8 GB
Homepages (Uncompressed)	20.8 GB

Table 3.2: *TREC Blog 2006 Collection Details [212]*

Figure 3.11 shows the excerpt of a RSS feed and figure 3.12 shows how a
permalink[7] document looks like in the collection.

```
<DOC>
<DOCNO></DOCNO>
<FEEDNO>BLOG08-feed-003540</FEEDNO>
<FEEDURL>http://alonsoquij.blogspot.com/feeds/posts/default</FEEDURL>
<BLOGHPURL>http://alonsoquij.blogspot.com/</BLOGHPURL>
<PERMALINKS>
http://alonsoquij.blogspot.com/2005/07/iii-ceremonia.html BLOG08-20080115-000-
0000000000
http://alonsoquij.blogspot.com/2005/07/ii-humildad.html BLOG08-20080115-000-
0000014292
</PERMALINKS>
<BLOG>http://alonsoquij.blogspot.com/</BLOG>
<DOCTIME>1200376164</DOCTIME>
<TIMESTATS>838620 1121043120</TIMESTATS>
<TIMES>1121043120 1121029200 1120738320 1120487100 1120393320 1120204500</TIMES>
<LINKCOUNT>15</LINKCOUNT>
<DOCHDR>
http://alonsoquij.blogspot.com/feeds/posts/default 0.0.0.0 2008115154924 16543
Cache-Control: max-age=0, must-revalidate, private
Connection: close
Date: Tue, 15 Jan 2008 05:49:24 GMT
```

Figure 3.11: *An example of a RSS Feed from Blog08 Data Collection*

---

[7]Permalink generally refers to URL of a specific blog entry (called "blogpost")

```
<DOC>
<DOCNO>BLOG08-20080115-004-0000000000</DOCNO>
<DATE_XML>1174496219</DATE_XML>
<FEEDNO>BLOG08-feed-003843</FEEDNO>
<FEEDURL>http://alekzmalchikspain.spaces.live.com/feed.rss</FEEDURL>
<BLOGHPURL>http://alekzmalchikspain.spaces.live.com/</BLOGHPURL>
<PERMALINK>http://alekzmalchikspain.spaces.live.com/Blog/cns!8D367438A6BBD4E0!
1635.entry</PERMALINK>
<DOCHDR>
http://alekzmalchikspain.spaces.live.com/Blog/cns!8D367438A6BBD4E0!1635.entry
0.0.0.0 200892323137 55301
Cache-Control: private
Date: Tue, 23 Sep 2008 00:01:35 GMT
Via: 1.0 bressay.dcs.gla.ac.uk:3128 (squid/2.6.STABLE17)
Server: Microsoft-IIS/6.0
Content-Length: 53591
Content-Type: text/html; charset=utf-8
Expires: Tue, 23 Sep 2008 00:01:35 GMT
Last-Modified: Fri, 25 Jan 2008 18:17:56 GMT
Client-Date: Tue, 23 Sep 2008 00:01:36 GMT
Client-Peer: 130.209.240.138:3128
Client-Response-Num: 1
```

Figure 3.12: *An example of a permalink from Blog08 Data Collection*

TREC Blog 2008 Data Collection   The TREC Blog tracks from year
2006 to year 2009 used the test collection Blog06. However a new bigger
data collection named BLOG08 was created by University of Glasgow for
TREC Blog track 2009 [210]. This blog collection covers a longer time
span period (i.e., from 14th Jan, 2008 to 10th, Feb 2009) which makes
over 1 year of time span. The details about data collection BLOG08 are
given below in table 3.3:

Characteristic	Value
Number of Unique Blogs	1,303,520
Number of Permalinks	28,488,766
First Feed Crawl	14/01/2008
Last Feed Crawl	10/02/2009
Total Compressed size	453 GB
Total Uncompressed size	2309 GB
Feeds (Uncompressed)	808 GB
Permalinks (Uncompressed)	1445 GB
Homepages (Uncompressed)	56 GB

Table 3.3: *TREC Blog 2008 Collection Details [53]*

TREC Blog Tasks

We will discuss the tasks defined for TREC Blog track in two steps. In
first step, we will describe the tasks defined for TREC Blog track of year

span 2006-2008 and in second step, we will discuss the tasks defined for TREC Blog track span 2009-2010.

Tasks 2006-2008   The tasks defined for TREC BLOG track 2008 covers all the tasks of previous years. Therefore, we will discuss all the tasks defined for year 2008 [53]

1. **Baseline Ad-hoc (Blog Post) Retrieval Task:** This is an ad-hoc information retrieval task where the objective is to *Find me blogposts about X*. This task was included in year 2008 in order to encourage the participants to evaluate their opinion finding approaches across different topical relevance baselines.

2. **Opinion Finding (Blog Post) Retrieval Task:** Opinion Finding task is the major task defined for TREC Blog track. It was defined for TREC Blog track 2006 [39] and continued to year 2008. The objective of this task is to reply the question *What do people think about X?* where X is a subject defined in the given topic. The basic idea behind this task is to retrieve all such blogposts that are not only relevant to X but also contains opinions about X.

3. **Polarity Opinion Finding (Blog Post) Retrieval Task:** Polarity task was added in TREC blog track 2007 [52] as an extension to opinion finding task. The basic objective was to *Find me positive or negative opinionated posts about X* (i.e., to identify the blogposts having positive or negative opinions about the topic).

4. **Blog (Feed) Distillation Task:** This task was defined for TREC Blog track 2007 and continued to 2008. The objective was to *Find me a blog with a principal, recurring interest in X* (i.e., to find the blogs frequently talking about a certain subject as expressed in topic so that this blog can be recommended to readers interested in X).

Note: The details of all the tasks defined for TREC Blog track can be found on TREC Blog WIKI[8].

Tasks 2009-2010   With the release of new collection in 2009, TREC changed the major tasks for TREC Blog track.

1. **Faceted Blog Distillation**
   Faceted Blog Distillation is a more refined version of the blog distillation task which takes into account the topic facets for retrieval. In

---

[8]http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

other words, a reader may not be interested to read all blogs having a recurring interest and principal interest in X but just a subset of such blogs that satisfy the conditions set by the topic facets. If re-defined formally then this task's objectives are to "Find me a good blog with a principal, recurring interest in X", where the sought quality of the blogs is characterized through the set facet inclinations [210].

The facets chosen for TREC Blog 2009 and 2010 are:

- **Opinionated Facet:** Opinionated facet restrictions allow retrieving only those blogs having a major interest in a given topic and also contain opinionated information for that topic. For this facet, the other restriction of interest is to retrieve only factual blogs.

- **Personal Facet:** The blogs have got too much for everyone (i.e., from individuals to organizations). A company may be interested to find some information from the other company's perspectives like the information about current market trend or the suitable financial institutions, etc. Similarly an individual might not be interested at all in knowing about marketing standards, etc. This facet will restrict the retrieval of blogs with major interest in the given topic and then sub-setting it according to company's and individual tastes. For this facet, the values of interest are Personal and Official blogs.

- **In-Depth Facet:** Another aspect from which readers might be interested to read blogs is the in-depth analysis of the issues and there may be others who just want to get shallow read of the issue. For this facet, the values of interest are In-depth vs. Shallow blogs (in terms of their treatment of the subject).

2. **Top stories identification** This task was introduced in TREC 2009 track to analyze the day to day news related dynamics of blogosphere. This task has two aspects:

(a) Identifying top headlines for a given unit of time and category

(b) Identifying relevant blog posts for a given headline which covers different/diverse aspects or opinions.

This task will help to analyze the temporal relationship between two media (i.e., news and blogosphere).

TREC Blog Topics and Assessments

From 2006 to 2009, TREC has been providing 50 new topics each year for their participating groups. A topic is actually a query (a set of terms) like the one we type to search for its relevant documents in some search engine. However in TREC queries are represented as shown in figure below. Generally each topic is represented by 4 tags (i.e., $< num >$ representing the topic number provided by TREC, $< title >$ the set of terms to be treated as a query, $< desc >$ gives a brief description of the title of the topic while $< narr >$ describes the relevance criterion for the documents for a topic). A typical representation of a TREC topic (for opinion task) is given in figure 3.13.

```
< top >
< num > Number: 851 < /num >
< title > March of the Penguins < /title >
< desc > Description:
Provide opinion of the film documentary "March of the Penguins".
< /desc >
< narr > Narrative:
Relevant documents should include opinions concerning the film documentary "March of
the Penguins". Articles or comments about penguins outside the context of this film
documentary are not relevant.
< /narr >
< /top >
```

Figure 3.13: *Standard TREC Blog Topic Format*

Besides this, TREC also provided the relevance assessments done by human annotators for each topic for the evaluation of the systems. The figure below shows a sample of assessments for topic 851 of TREC Blog track 2006.

The last column in the assessments represents the document labels that has been assigned by human assessors and are explained in table 3.4 given below:

3.6.2 NTCIR

The National Institute of Informatics (NII) runs annual meetings code-named NTCIR (NII Test Collection for Information Retrieval Systems).

```
851 0 BLOG06-20051207-068-0021005350 0
851 0 BLOG06-20051207-068-0023689277 0
851 0 BLOG06-20051207-068-0023708536 0
851 0 BLOG06-20051207-077-0002050920 0
851 0 BLOG06-20051207-079-0007582745 2
851 0 BLOG06-20051207-081-0015311921 0
851 0 BLOG06-20051207-086-0012716584 0
851 0 BLOG06-20051207-086-0012932787 0
851 0 BLOG06-20051207-088-0006532081 0
851 0 BLOG06-20051207-091-0008256494 0
851 0 BLOG06-20051207-092-0010317675 0
851 0 BLOG06-20051207-095-0019213818 0
851 0 BLOG06-20051207-097-0020977329 0
```

Figure 3.14: *An example of query relevance results(qrels) for topic-851*

Label	Caption	Description
−1	Not Judged	A label of −1 means that this document was not examined at all due to offensive URL or Header
0	Not Relevant	The post and its comments are not at all relevant to the topic
1	Relevant	The post or its comments contain some information about the topic but no opinion found about the topic concerned
2	Relevant, Negative Opinions	The post is relevant and contain a negative sentiment for the topic
3	Relevant, Mixed Positive and Negative Opinions	The post is relevant and contain both positive and negative opinions about the topic
4	Relevant, Positive Opinions	The post is relevant and explicitly positive about the topic

Table 3.4: *TREC Blog Relevance Judgements Labels*

Opinion analysis was featured at an NTCIR-5 workshop, and served as a pilot task at NTCIR-6 [84] and a full-blown task at NTCIR-7 [395] and NTCIR-8 [394].

NTCIR Opinion Analysis Data Collections

The data collection for NTCIR-6 opinion analysis pilot task was created using thirty queries over data from NTCIR Cross-Lingual Information Retrieval corpus covering documents from 1998 to 2001. The details of this corpus are given below in table 3.5.

Language	Topics	Documents	Sentences
Chinese	32	843	8,546
English	28	439	8,528
Japanese	30	490	12,525

Table 3.5: *NTCIR-6 Data Collection Details for Opinion Analysis Task*

Each set of documents for a topic is accompanied with its relevance assessments. An example of the topics in the NTCIR-6 opinion analysis corpus is shown in the figure 3.15 representing the topic 010 with title, "History Textbook Controversies, World War II".

```
< topic >
< num > 010 < /num >
< title > History Textbook Controversies, World War II < /title >
< desc > Find reports on the controversial history textbook about the Second World War
approved by the Japanese Ministry of Education.
< /desc >
< narr >< back > The Japanese Ministry of Education approved a controversial high
school history textbook that allegedly glosses over Japans atrocities during World War
Two such as the Nanjing Massacre, the use of millions of Asia women as "comfort women"
and the history of the annexations and colonization before the war. It was condemned by
other Asian nations and Japan was asked to revise this textbook.< /back >< rel >
Reports on the fact that the Japanese Ministry of Education approved the history
textbook or its content are relevant. Reports on reflections or reactions to this issue
around the world are partially relevant. Content on victims, "comfort women", or Nanjing
Massacre or other wars and colonization are irrelevant. Reports on the reflections and
reactions of the Japanese government and people are also irrelevant.< /rel >
< /narr >
< /topic >
```

Figure 3.15: *Standard NTCIR Opinion Analysis Topic Format*

More recent data collections were used for NTCIR-7 (for details, consult

[395]) and NTCIR-8 (for details, consult [394]) whereas Chinese Language portions were further divided into Traditional Chinese and Simplified Chinese.

NTCIR Opinion Analysis Tasks

NTCIR opinion analysis task started with total four subtasks in NTCIR-6. However, one additional subtask was introduced for NTCIR-7 and NTCIR-8 making total number of subtasks equal to five. The details of these subtasks are given in table 3.6.

Subtask	Values	Description
Opinionated Sentences	Yes, No	To determine whether a given sentence is opinionated or not
Relevant Sentences	Yes, No	To determine whether a given sentence is relevant or not
Opinionated Polarities	POS, NEG, NEU	To determine whether a given opinion expression is positive, negative or neutral
Opinion Holders	String, Multiple	To determine the entity who is expressing an opinion about something
Opinion Targets	String, Multiple	To determine the entity about which an opinion is being expressed

Table 3.6: *Subtasks defined for NTCIR Opinion Analysis Task from NTCIR-6 to NTCIR-8*

3.7 Chapter Summary

In this chapter we described the origin and sources of opinions on the web and the process of Opinion Detection on the web. We discuss in detail the applications of Opinion Mining. The section about Blogosphere explained the structure of the blogs and nature of the Blogosphere. Later on we described the TREC Blog track in detail with its tasks, topics and query relevance assessments (QRELS). At the end, we also discuss the TREC Blog Evaluation framework and then we end the chapter with challenges of Opinion Detection in blogs. In next chapter we discuss the related work of opinion detection from various aspects describing pros and cons of the approaches where possible.

Opinion Detection: From Word to Document Level

No and Yes are words quickly said, but they need
a great amount of thought before you utter them.
Baltasar Gracian

4.1   Introduction

This chapter reviews the literature concerning the process of computational treatment of opinions, sentiment, and subjectivity in text. In the literature this process is known by expressions like "opinion mining", "sentiment analysis", and/or "subjectivity analysis" [264]. Other commonly used terms for this process are "opinion detection", "sentiment detection", "polarity detection" and "opinion finding". In addition to this, many other terms have been used for opinion related work (like "affective computing" [362], "review mining" [407], "appraisal extraction" [41], etc.) but in this manuscript we will limit ourselves to the use of most common terms mentioned above. By definitions it appears that in broader sense subjectivity analysis, sentiment analysis, and opinion mining denote the same field of study. The term "opinion mining" was first coined by Dave et al. [83]. The basic aim of opinion mining is to determine human opinion from text written in natural language and recently has attracted lot of attention from researchers of this domain. Similarly, the popularity of sentiment analysis recites the same story as for opinion mining. The term *sentiment* started appearing in re-

search articles like [82, 342] published in 2001, [265, 345] published in 2002. A number of papers used the term of *sentiment analysis* (like [250, 389]) that explains its popularity in research community. Many of the articles that used the term of sentiment analysis focused on the task of classifying given text into positive or negative classes. However, nowadays this term is used in a broader sense and is meant for computational treatment of opinion, sentiment, and subjectivity in the text [264]. Wiebe [374] defines subjectivity as a function of private states (i.e., the states that are not open to objective observation or verification). Opinions, evaluations, emotions, and speculations all fall into this category [264]. The process of analyzing these opinions and emotions is called *Subjectivity Analysis* whose objective is to recognize the opinion-oriented language to distinguish it from objective language.

Year 2001 was the beginning of widespread awareness of the research problems related to opinion mining which caused hundreds of papers published on this subject (see figure 4.1)[1]. Pang et al. [264] describe the factors behind this sudden increase in interest in the field of opinion mining and sentiment analysis:

○ the popularity of machine learning methods in natural language processing and information retrieval,

○ the availability of datasets for machine learning algorithms to be trained on, due to the blossoming of the World Wide Web and, specifically, the development of review-aggregation web-sites; and,

○ realization of the fascinating intellectual challenges and commercial and intelligence applications that the area offers.

This chapter is organized as follows: In section 4.2, we will provide a brief overview of already existing works that have given different classifications of opinion-related work. Section 4.3 defines the opinion detection in few steps and gives a novel classification of opinion related work categorized with respect to each step of this process. Basically, each step of opinion detection process defines a different granularity-level of text processing unit (from words to documents). In section 4.4, we discuss several challenges that field of opinion mining is facing with brief discussions of their related work.

---

[1]Taken from slides of talk by Andrea Esuli on the topic of opinion mining in Istituto di Scienza e Tecnologie dell'Informazione Consiglio Nazionale delle Ricerche, Pisa, Italy, 14th June 2006
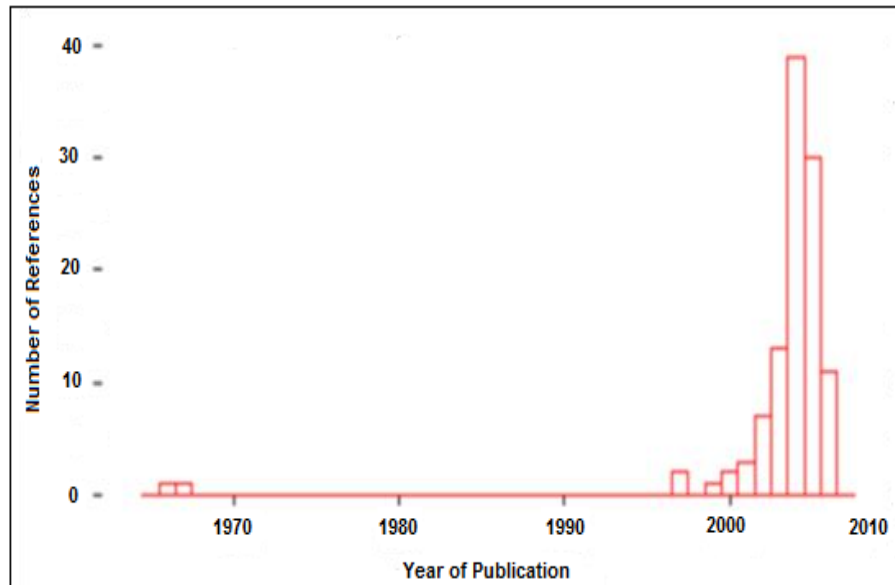
Figure 4.1: *Emerging trend in number of articles for opinion mining research*

4.2   Major Opinion Mining References

In this section, we will present prominent existing works that have summarized the work related to opinion mining and have classified the related work with respects to different aspects.

4.2.1   Work by Tang et al.

Tang et al. [337] present a detailed survey of work for sentiment detection in product reviews. They identify three kinds of major approaches in the literature for sentiment detection in real-world applications:

○ **Machine Learning Approaches** In this type of approaches, generally a machine learning classifier is trained on already annotated data to create a model of the trained data and then this model is used to estimate the classes of documents in the test data.

○ **Semantic Analysis Approaches** Lexical resources play a very important role in this type of approaches. Semantic relations of concepts, extracted from some lexical resource, are used to provide some evidences about the subjectivity. Use of synonyms and antonyms has

been very common in this regard.

- ○ **Natural Language Processing Approaches** Approaches exploiting the Parts-of-Speech (POS) information, complex syntactical structural information, etc. are part of this type of approaches.

Besides this, Tang et al. also highlight the related work in context of major tasks like "subjectivity classification", "sentiment classification", etc.

4.2.2 Work by Esuli et al.

Similarly, Esuli et al. [100] have categorized the related work in three classes according to the nature of tasks associated with sentiment detection. These three classes are:

- ○ **Determining Text SO-Polarity:** The type of approaches belonging to this class focus on the task of deciding whether a given text is factual or contains opinions on a topic (i.e., a binary text categorization with classes *Subjective and Objective*).

- ○ **Determining Text PN-Polarity:** The task this type of approaches focus on is to evaluate the polarity of a subjective text (i.e., whether given subjective text contain positive or negative opinion about the target).

- ○ **Determining the strength of Text PN-Polarity:** Once it has been decided whether a given text is positive or negative then the task of determining the degree of its positivity or negativity becomes active. The approaches in this category of classes calculate this degree of positivity or negativity.

4.2.3 Work by Pang et al.

While Esuli et al. only describes three tasks related to problem of opinion mining, Pang et al. [264] identify a set of relatively larger number of opinion related tasks in the literature. Few major tasks are listed below:

- ○ **Sentiment Polarity Classification:** It is a binary classification task in which the polarity of a given opinionated document is estimated to be positive or negative.

- ○ **Likely vs Unlikely:** Another related task identified by Pang et al. [264] is classifying predictive opinions in election forums into *likely to win* and *unlikely to win* classes.

○ **Good vs Bad News:** Classifying a news article as a *good news* or *bad news* has also been identified as a sentiment classification task.

○ **Reviewer's Evaluation:** Another task is to determine reviewer's evaluation with respect to a multi-point scale (e.g., one to five stars for a review). This problem can be seen as a multi-class categorization problem.

○ **Agreement Detection:** Given a pair of texts, deciding whether they should receive the same or different sentiment-related labels based on the relationship between elements of the pair.

○ **Opinion Strength:** Another task identified was to determine the clause-level opinion strength (e.g., *How mad are you?*).

○ **Viewpoint Classification:** Classifying the viewpoints and perspectives into classes like *liberal, conservative, libertarian, etc.* is another task identified.

○ **Genre Classification:** This task focuses on determining the genre of a given piece of text i.e. whether the given text is an editorial, advertisement or announcement, etc.

○ **Source Classification:** Classifying the documents according to their source or source style. Authorship identification is a very good example of such task or similarly classifying the documents according to their publisher (e.g., *The Washington Post or The Daily News*).

Prominant Classification Features

A fundamental technology in many current opinion mining applications is *Classification* [264]. Different approaches have experimented with different sets of features proposed to distinguish opinionated documents from non-opinionated documents. In addition to defining opinion related tasks, Pang et al. [264] also present major features specifically proposed for the task of sentiment analysis in related work. Below, we present a summary of opinion features discussed by Pang et al.

○ **Term Frequency:** Term frequencies have traditionally been important in classical IR, however story seems to be different in opinion mining. For instance, Pang et al. [265] obtained better performance with presence of the terms rather than their frequency. In their work, the use of binary-valued feature (i.e., with values 0 or 1) representing the presence (1) or absence (0) of a term performed better than

the results obtained by using term frequencies for the task of polarity classification. This finding is an indication of differences between natures of topic-based text categorization and sentiment classification (i.e., while a topic is more likely to be emphasized by frequent occurrences of certain keywords, overall sentiment usually remains behind the scenes through repeated use of the same terms).

○ **Parts of Speech**: Parts-of-speech (POS) information has been frequently used as an evidence in sentiment analysis and opinion mining. Using POS information has given very interesting findings though. For example according to few works (like [49, 127]), adjectives have been reported to be more subjective than other parts-of-speeches. The fact that adjectives are good predictors of subjectivity of a sentence does not, however, imply that other parts of speech do not contribute to expressions of opinion or sentiment. In fact, in a study by Pang et al. [265] on movie-review polarity classification, using only adjectives as features did not perform as well as the same number of most frequent unigrams. The researchers point out that nouns (e.g., *gem*) and verbs (e.g., *love*) can be strong indicators for sentiment. Riloff et al. [298] specifically studied extraction of subjective nouns (e.g., *concern, hope*) via bootstrapping. There have been several targeted comparisons of the effectiveness of adjectives, verbs, and adverbs, where further subcategorization often plays a role [31, 250, 376].

○ **Syntax**: The use of syntactical structure of textual units has been one of the most important features for the task of sentiment classification. For instance, Kudo and Matsumoto [180] report that for two sentence-level classification tasks (i.e., sentiment polarity classification and modality identification (*opinion, assertion or description*)), a subtree-based boosting algorithm using dependency-tree-based features outperformed the bag-of-words baseline (although there were no significant differences with respect to using n-gram-based features). Nonetheless, the use of higher-order n-grams and dependency or constituent-based features has also been considered for document-level classification; Dave et al. [83] on the one hand and Gamon [106], Matsumoto et al. [223], and Ng et al. [251] on the other hand come to opposite conclusions regarding the effectiveness of dependency information. Parsing the text can also serve as a basis for modeling valence shifters such as negation, intensifiers, and diminishers [170]. Collocations and more

complex syntactic patterns have also been found to be useful for subjectivity detection [289, 376].

○ **Negation:** Handling negation can be an important concern in opinion and sentiment-related analysis. For example, compare these two sentences:

I like this laptop

and

I dont like this laptop

While using similarity measures, these two sentences might be very similar to each other but they fall in opposite classes when analyzed for their sentiments. The negation *not* inverses the polarity of the word *like*, hence inverting the polarity of the overall sentiment of the sentence. There are few works that take into account this evidence of inverse in polarity in their approaches. For example, Das and Chen [82] propose attaching *NOT* to words occurring close to negation terms (such as *no* or *dont*) so that in the sentence *I dont like deadlines*, the token *like* is converted into the new token *like-NOT*.

However, not all appearances of explicit negation terms follow the same tradition of reversing the polarity of the sentence. For instance, for the sentence *No wonder this is considered one of the best*, the approach of Das and Chen [82] will not work. However, Na et al. [246] attempt to target this problem of negation. They look for specific part-of-speech tag patterns (where these patterns differ for different negation words), and tag the complete phrase as a negation phrase. For their dataset of electronics reviews, they observe about 3% improvement in accuracy resulting from their modeling of negations. Further improvement probably needs deeper (syntactic) analysis of the sentence [170]. Similarly, Wilson et al. [381] discuss other complex negation effects.

○ **Topic-Oriented Features:** Opinion-Topic association is very important in opinion mining. A document may contain opinions about many topics. Therefore, to extract only those opinions (and sentiments) which are related to given topic, effective techniques are required. Mullen and Collier [244] examine the effectiveness of various features based on topic (e.g., they take into account whether a phrase follows a reference to the topic under discussion) under the experimental condition that topic references are manually tagged. Thus, for example, in a review of a particular work of art or music, references

to the item receive a *THIS WORK* tag. Topic-sentiment interaction has also been modeled through parse tree features, especially in opinion extraction tasks. Relationships between candidate opinion phrases and the given subject in a dependency tree can be useful in such settings [275].

4.3   Granularity-based State-of-the-Art

While works by Tang et al. [337], Esuli et al. [100], and Pang et al. [264] have organized the related work based on the nature of tasks and type of approaches adopted, we provide a novel granularity-level (word, sentence/-passage, and document) classification of the related work for opinion mining. We describe the opinion mining process in few steps and then work related to each step is discussed in separate sections. This organization of work is very useful for researchers working on the task of opinion detection at any granularity level. In addition to this, we also discuss related work for major challenges of field of opinion detection which gives an overview of opinion mining work from another perspective.

4.3.1   Opinion Detection Process

The process of "opinion detection" can be described in following major steps:

1. Retrieve the relevant set of documents for a given topic (Topic Relevance Retrieval) if needed,

2. Compute the word-level polarity orientations (determining whether a word is positive or negative) and polarity strengths (determining the strength of the positivity or negativity of a word),

3. Combine the word-level subjectivity scores, polarity orientations or strength to calculate the polarity orientations and strengths on sentence-level (or passage-level),

4. Combine the sentence-level subjectivity scores, polarity orientations or strengths to compute the polarity orientations and strengths of the given document.

Each step of the above process sacks lot of research work which demonstrates the importance of opinion mining process in IR field. Therefore,

we classify the related work in light of this step-wise process of opinion detection by discussing approaches for each step.

In the rest of this chapter, we will discuss the related work for opinion detection on word, sentence and document levels. The approaches falling in each category are further classified (if needed) according to the nature of lexical resources, data collections and other used techniques to give an overview of related work from various perspectives.

4.3.2 Word Level Processing

The work on word level generally corresponds to prediction of sentimental orientation of words in a document and calculating their sentimental strength. Predicting sentimental orientations of words is necessary for estimating the sentimental orientation of a sentence or a document [127]. The sentiment of a word indicates the direction the word deviates from the norm for its semantic group or lexical field [187]. It also restricts the word's usage in the language [100]. Positive sentimental orientation indicates praise (e.g., *honest, intrepid*) and negative sentimental orientation indicates criticism (e.g., *disturbing, superfluous*). The expressions like "sentiment (or semantic) tagging", "semantic orientation", and "polarity orientation" are also used for sentimental orientation sometimes.

The work related to computation of sentimental orientation of words is comprised of several approaches that involve the use of synonyms, antonyms, language constructs (like conjunctions) and lexical resources (like Word-Net [230]). WordNet groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. We can identify three word-level sentiment analysis tasks in the literature [100] as described below:

○ To determine subjectivity of words in a document (i.e., whether the word is subjective or objective)

○ To determine orientation or polarity of words (i.e., whether the word is positively subjective or negatively subjective)

○ To determining strength of orientation (i.e., how much positive or negative a word is)

Most of the approaches found in literature do not differentiate between these tasks. Therefore, discussion of the related work for word-level sen-

timent analysis will not be done in context of these tasks but type of techniques used for these tasks.

A general overview of the related work unveils an interesting observation about the use of adjectives for the task of computing semantic orientation. It was reported by few early works [49, 375] that adjectives are more vulnerable to be subjective than other parts-of-speech. Therefore most of the later work for word level semantic orientations mostly focused on the use of adjectives.

Generally two kinds of approaches have been proposed for determining the sentiment orientation of words [12]: first, *Corpus-Based approaches* and second, *Dictionary-Based Approaches*. However, few approaches combine both to propose a mixed approach. Before going into details of these type of approaches, we would like to summarize the contents of this section in list form for better understanding:

1. Corpus-Based Approaches
   - Using Language Constructs
   - Using Co-occurrence Evidence
2. Dictionary-Based Approaches
   - Use of Semantic Relations
   - Use of Gloss Definitions
   - Using WordNet Affect
3. Mixed Approaches

Corpus-Based Approaches

Corpus-based approaches generally exploit the inter-word relationships (syntactic or co-occurrence relationships) in large corpora to perform any of the three tasks defined above [117, 126, 174, 346, 396]. We discuss the major works for such kind of approaches by classifying them according to the nature of evidences used.

Using Language Constructs   This kind of approaches generally take support of language constructs (conjunctions, prepositions, grammar rules, etc.) to perform their tasks. For example, Hatzivassiloglou et al. [126] proposed a method for automatically tagging the adjectives with a sentimental tag (positive or negative) with the help of conjunctions (and,

or, but, either-or, or neither-nor) joining them. The basic principle behind their approach was that adjective combined with the conjunction *and* (like *beautiful and calm*) are supposed to have same orientation while those joined by conjunction *but* (like *justified but brutal*) generally differs in their sentimental orientations. The experiments were conducted in a large corpus of 21 million words of Wall Street Journal articles (which is a subset of TIPSTER[2] document collection). A classification precision of over 90% was observed for adjectives that occur with modest number of conjunctions in the corpus. Other studies [127, 370] showed that restricting features, used for classification, to those adjectives that come through as strongly dynamic, gradable, or oriented improved performance in the genre-classification task.

Using Co-occurrence Evidence   Baroni et al. [27] used a list of subjective adjectives as a seed set to rank a list of adjectives that are to be ranked in descending order by their subjectivity. The motivating factor behind this work was the intuition that subjective adjectives are most likely to co-occur with other subjective adjectives. They calculated the subjectivity score of target adjectives by computing their mutual information with the adjectives of seed set and *Pointwise Mutual Information (PMI)* [67] technique was used for this purpose. The idea of calculating the mutual association using Pointwise Mutual Information (PMI) between words was taken from work of Turney et al. [348]. There are few studies [67, 117, 332, 348] that have already demonstrated the effectiveness of PMI in comparison with other sophisticated association measures such as *log-likelihood ratio and cosine similarity*. PMI can be defined as [67]:

$$PMI(t, t_i) = \log_2 \left(\frac{p(t \& t_i)}{p(t) \cdot p(t_i)} \right) \tag{4.1}$$

Where $p(t \& t_i)$ is the probability that terms t and t_i occur together. In other words, above equation represents the measure of the degree of statistical dependence between t and t_i. For measuring the co-occurrence of adjectives, NEAR operator of the AltaVista[3] search engine was used where the NEAR operator produces a match for a document when its operands appear in the document at a maximum distance of ten terms, in either order.

---

[2] http://www.ldc.upenn.edu/
[3] http://www.altavista.com/

Turney and Littman [346, 348] proposed an approach to determine the sentimental orientation of terms. They prepared two sets of seed terms (i.e., one for negative terms and other for positive terms) as given below:

$$S_p = \{good, nice, excellent, positive, fortunate, correct, superior\} \quad (4.2)$$

$$S_n = \{bad, nasty, poor, negative, unfortunate, wrong, inferior\} \quad (4.3)$$

The basic idea behind is to infer semantic orientation (sentimental) from semantic association. The semantic orientation of a given word is calculated from the strength of its association with a set of positive words, minus the strength of its association with a set of negative words. For example, the orientation value of a given term t. $O(t)$ is computed as:

$$O(t) = \sum_{t_i \in S_p} PMI(t, t_i) - \sum_{t_i \in S_n} PMI(t, t_i) \quad (4.4)$$

Where $PMI(t, t_i)$ is the Pointwise Mutual Information [67] score for term t with each seed term t_i as a measure of their semantic association.

The results show that this approach required a large data collection for good performance. Even this is understandable because the reliability of the co-occurrence data increases with the number of documents for which co-occurrence is computed but still it is a limitation of this approach. Another drawback with this approach is that it did not deal with ambiguous terms (having both positive and negative senses at a time like the word *mind, unpredictable, etc.*) because the ambiguous terms were deleted from the set of testing words.

Dictionary-Based Approaches

The second type of approaches for word-level sentiment analysis benefit from the flexibility provided by various lexicons (like WordNet [230]) through its structure and lexical relations. The definitions like terms' glosses [99] and semantic relations (like synonyms and antonyms) [162] provide enough level of liberties to the researchers to be exploited for finding semantic orientations of words.

Use of Semantic Relations   Use of semantic relations has always been part of classical IR and it has got equal importance in the field of opinion mining and sentiment analysis. There exists a number of publications exploiting lexical semantic relations between concepts to estimate their subjectivity which eventually assits to estimate the subjectivity of a document. For example, Kamp et al. [162] developed a distance based WordNet measure to determine the semantic orientations of adjectives. This measure is based on the distance of a word from two selected reference words, "good" and "bad" and is given below:

$$SO(t) = \frac{d(t, bad) - d(t, good)}{d(good, bad)} \tag{4.5}$$

Where $d(t_1, t_2)$ is the shortest path between connecting any two terms t_1 and t_2. The adjective t is considered as positive if $SO(t) > 0$, and the absolute value of $SO(t)$ determines the strength of this orientation (the constant denominator $d(good, bad)$ is a normalization factor that constrains all values of SO to lie in the interval $[0, 1]$). Good results were reported after evaluation against manually constructed lists of General Inquirer [99, 404]. Williams et al. [377] use lexical relations of WordNet to assign polarity strength to adjectives. They use a small set of reference positive and negative terms to build an adjective graph, by using the lexical relations defined in WordNet. To compute the polarity strength of adjectives, they used various combinations of lexical relations. The best results were achieved when using the lexical relations of related words and similar words in addition to the standard synonym relation commonly used.

Use of Gloss Definitions   WordNet is a large lexical database containing about 150,000 words organized in over 115,000 synset entries for a total of 203,000 word-sense pair [266]. Each word comes along with a short description for all of its senses which is called its gloss definition. The glosses are usually one or two sentences long. For example, gloss definitions for the word *Car* are:

○ a motor vehicle with four wheels; usually propelled by an internal combustion engine

○ a wheeled vehicle adapted to the rails of railroad

○ the compartment that is suspended from an airship and that carries personnel and the cargo and the power plan

○ where passengers ride up and down

○ a conveyance for passengers or freight on a cable railway

There are few approaches [99, 100, 312] that make use of the quantitative analysis of the gloss definitions of terms found in online dictionaries to determine their semantic orientations. The motivation behind the work of Esuli et al. [99] is the assumption that if a word is semantically oriented in one direction, then the words in its gloss tends to be oriented in the same direction. For instance, the glosses of terms *good* and *excellent* will both contain appreciative expressions; while the glosses of *bad* and *awful* will both contain derogative expressions.

Like Turney et al. [348], Esuli et al. [99] started with seed set of positive and negative terms. This seed set was further expanded using lexical relations of WordNet. The gloss definitions for each term in new expanded set are obtained (and collated if more than one). This creates the training data for a binary text classifier. The experimentation was performed using data collections of [126, 162, 348] while using the seed set of previous works [162, 348] for a fair comparison with their results. The learning algorithms used for this work are *Multinomial Naive Bayes Model*, *Support Vector Machines (SVM)* using linear kernels, and the *PrTFIDF probabilistic version of the Rocchio learner* [157]. This work outperformed the results of all previous work including the best published results of that time [348].

Sebastiani [312] extends the work presented in [99] by including an additional task of determining term subjectivity. However a decrease in performance of state of the art approach [99] was noted once it is modified for the task of determining term subjectivity. The results suggest that deciding term subjectivity (including term orientation) is a substantially harder task that deciding term orientation alone.

Further extension to these works led to the creation of an automatic subjectivity lexicon SentiWordNet (SWN) [100]. SWN assigns three numerical scores $(Obj(s), Pos(s), Neg(s))$ to each synset of the WordNet describing how objective, positive or negative the terms within a synset are. The range of three scores lies in interval [0, 1] and sum of all the scores equals to 1. This process of assigning scores makes the task of determining semantic orientation and semantic strength more precise than the one in which terms are labeled just with tags subjective or objective(for semantic orientation task) or Strong or Weak (for polarity strength task). All of three scores are obtained by combining the results of eight ternary classi-

fiers, all characterized by similar accuracy levels but different classification behavior. A template of SWN is shown in figure 4.2.

```
# POS offset      PosScore    NegScore    SynsetTerms
a    1000003      0.0    0.125 form-only#a#1
a    1000159      0.25   0.0    dress#a#1 full-dress#a#1
a    1000307      0.0    0.0    titular#a#5 nominal#a#6
a    1000440      0.0    0.0    prescribed#a#4 positive#a#5
a    1000554      0.0    0.25   perfunctory#a#2 pro_forma#a#1
a    1000681      0.0    0.5    semiformal#a#1 black-tie#a#1 semi-formal#a#1
a    10007 0.0    0.625 abstentious#a#1 abstinent#a#1
a    1000859      0.0    0.0    starchy#a#2 buckram#a#1 stiff#a#4
```

Figure 4.2: *Template of SentiWordNet with first column: Parts of Speech (POS) of the Synset, 2nd column: Offset of the Synset in WordNet, 3rd Column: Positive Score of the Synset, 4th Column: Negative Score of the Synset, 5th Column: Entries of a Synset*

Quantitative analysis of the glosses of the synsets is performed to obtain three scores as mentioned above. The basic intuition behind the creation of SWN was that different senses of a term might have different semantic orientations. For example, the term "estimable" is objective (i.e. $Obj(estimable) = 1.0$ with its $Pos = Neg = 0.0$) corresponding to its sense *may be computed or estimated* and SWN scores for the same term become as $Obj(estimable) = 0.25$, $Neg(estimable) = 0$ and $Pos(estimable) = 0.75$ when its sense *deserving of respect or high regard* is taken. SWN has been used in many opinion related approaches [5, 58, 92] and has performed well.

However, there are few other works [11, 174, 334] too who have treated the task of determining semantic orientation same as [100] (i.e., instead of viewing the properties of positivity and negativity as categories, graded versions of these properties have been proposed.)

Using WordNet Affect   Valitutti [349] developed a lexicon called *WordNet Affect* for representation of affective knowledge by selecting and tagging a subset of WordNet synsets with the affective concepts like emotion, trait and feeling etc. For building this lexicon, a support was taken from another lexicon *WordNet Domains* [213]. WordNet Domains is a multilingual extension of the WordNet and provides at least one domain label (like Sports, Politics, and Medicine, etc.) for each of its synset. It has a hierarchy of almost two hundred domain labels. WordNet-Affect is an additional hierarchy of the affective domain labels, independent from the

domain hierarchy, wherewith the synsets that represent affective concepts
are annotated. Bobicev et al. [42] has used WordNet-Affect to develop
another multilingual (Russian and Romanian) WordNet-Affect lexical re-
source.

Generally, it has been observed that corpus-based approaches for word-
level subjectivity classification perform better than dictionary-based ap-
proaches. However, the performance of corpus-based approaches is badly
affected across different domains. On the other hand, most of the dictionary-
based approaches generally take support of domain-independent lexical re-
sources (e.g., SentiWordNet, WordNet); hence avoiding the drawback of
corpus-based approaches. However, performance of dictionary-based ap-
proaches might vary with the nature and scope of the lexicon being used.

Mixed Approaches

There are very few works though where both of above approaches (i.e.,
dictionary-based and corpus-based approaches) were combined to improve
the results. One of the examples of such work is the work by Kim et
al. [404] whereby they prepared a long list of opinion words to identify
opinion bearing sentences. Three resources were used to prepare the list
of opinion words. First, they prepare a list of few opinion and non-opinion
words (verbs and adjectives) manually. This list was expanded with their
synonyms and antonyms in WordNet (see figure 4.3) assuming that syn-
onyms and antonyms of opinion words are opinionated words too. For each
target word (i.e., synonym or antonym), its WordNet distance to the two
sets of manually selected seed words plus their current expansion words
was measured. They assigned the new word to the closer category. The
following equation represents this approach:

$$argmaxP(c|w) \cong argmaxP(c|syn_1, syn_2, syn_3, \ldots, syn_n) \qquad (4.6)$$

Where c is a category (opinion bearing or non-opinion bearing), w is the
target word, and syn_i is the synonyms or antonyms of the given word by
WordNet.

Second, another list of opinion bearing words was prepared using *Wall
Street Journal (WSJ)* data by assuming that words that appear more of-
ten in newspaper editorials and letters to the editor than in non-editorial
news articles could be potential opinion bearing words. The collection was
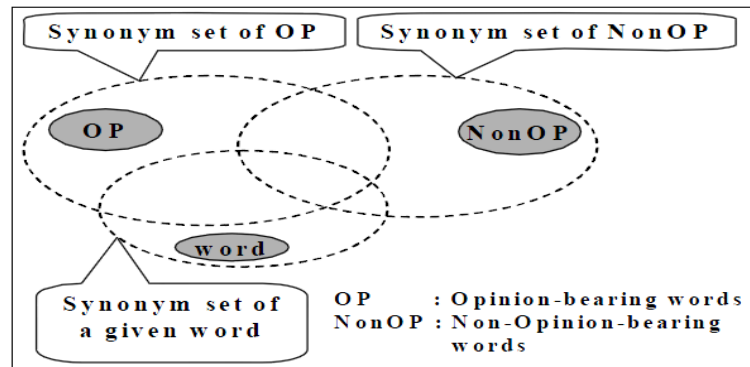
Figure 4.3: *Automatic word expansion using WordNet Synonyms*

classified into Editorial and Non-Editorial sets. They separated out opinion words from non-opinion words by considering their relative frequency in the two sub-collections. The list of opinion bearing words was prepared and filtered to a have final list.

The third list of opinion words was provided by *Columbia University*. Finally, all three lists were merged to prepare a final list of opinionated words. The score of words were averaged and normalized and then top 2,000 opinion bearing words and top 2,000 non-opinion bearing words for the final word list. This final list of opinionated words was later used for automatic selection opinion bearing sentences for three different data sets (MPQA data collection [371], an internal data collection and TREC 2003 Novelty track data [319]).

4.3.3   Sentence Level Processing

Most of the work related to opinion mining on sentence level focuses on following two tasks:

- To determine whether a sentence is subjective or objective,
- To determine whether a sentence is positive or negative.

In this section, we will discuss few major contributions for both tasks.

Sentence Subjectivity Identification

In this section, we will discuss approaches that have used different types of evidences to determine whether a given sentence is subjective or objective.

Using Presence of Subjective Words   Most of the approaches rely on the evidence of presence of subjective words in a sentence to analyze the subjectivity of that sentence. For example, Kim et al. [404] proposed two models for identifying opinion bearing sentences:

First Model (Model-1): This model depends on the total opinion score of all words in a sentence.

Second Model (Model-2): The basic idea behind this model is that if a sentence contains even a single strong opinionated word then it is an opinion sentence.

They prepared a list of opinion words from different sources like WordNet (by measuring the distance of a given word from a set of seed set of positive and negative opinionated terms (see figure 4.3)), WSJ data collection (by using the relative frequencies of opinion terms) and the word list of Yu and Hatzivassiloglou [396]). The list of words prepared from all three sources were merged to have a final list of opinionated terms.

The results of the experimentation performed on TREC 2003 Novelty track data with different system cut-off values show that model-2 performs better than model-1 (see table 4.1).

Table 4.1: *System performance with different models and cutoff values on TREC 2003 data*

Model	System Parameter λ	F-Score
Model-1	0.2	0.398
	0.3	0.425
Model-2	0.2	0.514
	0.3	0.464

However, an interesting relation between presence of adjectives in a sentence and it's subjectivity has been explored by many works. For example, Bruce et al. [49] proved that adjectives are statistically, significantly and positively correlated with subjective sentences in the corpus on the basis of the log-likelihood ratio. The probability that a sentence is subjective, simply given that there is at least one adjective in the sentence, is 55.8%, even though there are more objective than subjective sentences in the corpus.

The work of Bruce et al. motivated further research to look for relation between presence of adjectives in a sentence and sentence subjectivity. Hatzivassiloglou et al. [127] experimented with several lexical features of adjectives to determine their ability to affect the subjectivity of a sentence. The objective was to observe the effects that an adjective's semantic ori-

entation and gradability has on its probability of occurring in a subjective sentence (i.e., to check who the best predictor of sentence subjectivity is). The set of adjectives S for this study was a union of few sets of adjectives (dynamic adjectives, gradable adjectives, adjectives with semantic orientation labels; both manually and automatically collected) from several previous studies [49, 126]. A very simple method was adopted to predict the subjectivity of a sentence:

A sentence is classified as subjective if at least one member of a set of adjectives S occurs in the sentence and objective otherwise.

Experiments were performed by varying the set S (i.e., with all adjective, with only gradable adjective, with only positive or negative adjectives, etc.) to assess the impact of each subset of set S on sentence subjectivity. The major findings of this work are:

○ All subsets involving dynamic adjectives, positive or negative adjectives or gradable adjectives are better predictors of sentence subjectivity than the class of adjectives as a whole

○ In most cases, automatically classified adjectives are comparable or better predictors of sentence subjectivity than the manually classified adjective

○ The probability of predicting the subjectivity of sentences correctly improves or remains same as additional lexical features are added

A further investigation was done by Wiebe [370] in this regard to verify the role of further high quality adjective features. In this work, a baseline was created using the performance of the simple adjective features and higher quality adjectives features were identified using the results of a method for clustering words according to distributional similarity [197], seeded by a small amount of detailed manual annotation. In addition, lexical semantic features of adjectives (i.e., polarity and gradability) [126] also form part of the feature space. A 10-fold cross validation experimentation shows that new features performed much better than the baseline. The work presented in article [373] is another useful contribution by Wiebe et al. in this regard.

We have seen that most of the earlier work depends on presence of adjectives within a sentence for subjective classification of a sentence but Riloff et al. [290] proposed an approach which exploits the subjectivity of nouns for identification of subjective sentences. They develop a method to learn

the sets of subjective nouns using bootstrapping algorithms. Then a Naive Bayes classifier was trained using the subjective nouns, discourse features, and subjectivity clues (from previous research work) to check the impact of these features on subjective classification of sentences. The results show a precision of 81% on subjective classification of sentences.

Use of Sentence Similarities   The results of approaches mentioned above have shown that the presence of adjectives in a sentence can be a good clue for a sentence to be subjective. However, other evidences have also been used for estimating the subjectivity of a sentence which involves comparison of the given sentences with subjective sentences, use of Parts of Speech (POS) information of the words present in a sentence and also the sequence of polar words in a sentence [396]. Hong et al. [396] proposed three different approaches for classifying a sentence as opinionated or factual sentence. The first one is the Similarity approach with the hypothesis that, within a given topic, opinion sentences will be more similar to other opinion sentences than to factual. A state-of-the-art sentence similarity algorithm SIMFINDER [125] was used. SIMFINDER calculates the similarity between two sentences with the help of shared words, phrases, and WordNet synsets. In their second method, Naive Bayes classifier was trained with features like words, bigrams, trigrams, parts of speech (POS) and number of positive and negative words in each sentence [127]. In addition, the counts of the polarities of sequences of semantically oriented words (e.g., ++ for two consecutive positively oriented words) and the counts of POS combined with polarity information (e.g., $JJ+$ for positive adjectives).

Using Product Features Information   Hu et al. [139] provide a summary of a product review by selecting a set of opinion sentences for each feature of a product. A combination of data mining and natural language processing techniques is used to mine the product features [140]. A very simple method of selecting the opinion sentences from the customer reviews was adopted. If a sentence contains one or more product features and one or more opinion words, then the sentence is called an opinion sentence.

Sentence Polarity Tagging

It is to be noted that the performance of an approach developed for predicting the polarity orientation of a sentence is dependent on the performance of the approach proposed to estimate the polarity estimation of words within that sentence. Therefore, it is only an effective combination of techniques on both levels that can eventually give good performance for predicting the sentimental orientation of sentences.

Using Word-Level Polarity Scores   The approach proposed by Hong et al. [396] tags the opinion sentences with polarity tags (i.e., positive or negative). They used a co-occurrence measure including a seed-set of semantically oriented words to estimate the polarity orientations of words in a sentence. This has been discussed in previous section in detail. For evaluation purposes, they aggregated the word-level polarity scores to estimate sentence level polarity orientations with different combinations of parts-of-speeches (i.e., adjectives, adverbs, nouns, verbs). However, maximum accuracy was obtained (90% over baseline of 48%) when they combined word-level evidences for adjectives, adverbs and verbs.

Using Number of Subjective Words   Hu et al. [139] proposed an approach for summarizing the customer reviews for product features. A combination of data mining and natural language processing techniques is used to mine the product features [140]. Selection of positive and negative opinion sentences is done for each product feature and then presented to user as a summary. A very simple method for detection of sentence polarity orientation was adopted by Hu et al. (i.e., if a sentence contains more number of positive words than negative words, it is considered as a positive sentence otherwise negative). In the case where there are equal numbers of positive and negative opinion words in the sentence, they predict the orientation using the average orientation of effective opinions or the orientation of the previous opinion sentence. In this research, Hu et al. used the adjective synonym set and antonym set in WordNet to predict the semantic orientations of given words whose orientations need to be determined. For each feature in the sentence the nearby adjective is referred to as its effective opinion. For example, *horrible* is the effective opinion of *strap* of the camera in sentence given below:

The strap is horrible and gets in the way of parts of the camera you need access to.

Their system performed well by giving an average accuracy of 84% in predicting the sentence sentimental orientation.

We have seen that most of the sentence-level work depends on the semantic orientations of the words (discussed already above) contained within a sentence to calculate its semantic orientation. But it should be noted that polarity of a word is likely to change when it is surrounded by other words in a sentence. In other words, polarity of an individual word *(prior polarity)* and polarity of a word in a sentence *(contextual polarity)* are most likely to be different. For example, take the following sentence:

John's house is not beautiful at all

We know that word *beautiful* has a positive prior polarity but in above sentence the contextual polarity of the word *beautiful* is negative because of the presence of negation *not* just before the word *beautiful* in the sentence. In rest of the discussion for sentence polarity tagging, we will present some works that have proposed sentence polarity approaches by focusing on the problem of contextual polarity of words.

Using Word-Level Context-Aware Polarity Approaches   Contextual polarity of a term is the polarity which is generated after modification of the prior polarity of the term. This modification of the prior polarity occurs because of change in the context. Here we define few major contexts responsible for polarity shift of the terms:

○ This type of contextual polarity is defined by the presence of negations (like *not, neither, nor or never,* etc.) in surroundings of a given word. For example, *Good* is a positive word but if preceded by a negation like *not or never*, its contextual polarity is changed from positive to negative. It can also be said that prior and contextual polarities of words remain same as long as they are not modified by some negative words in their surroundings.

○ The second type of contextual polarities are caused by the senses of a word as found in a everyday dictionary (like WordNet). A word can have many senses. This is called *Polysemy.* For example, *bank* can be used as a *financial institute* or a *river shore.* Similarly, the polarities

of words can be different for different senses of a word. For example, while the word *strong* is considered a positive adjective (with positive score of 0.75 and negative score 0.0) when used as sense strong#a#7, it is more likely to highlight its negative aspect (with negative score of 0.5 and positive score of 0.0) when used as sense strong#a#8 in subjective lexicon SentiWordNet(SWN) [100].

o Third type of contextual polarity is defined by the type of the topic (or query) we are searching for so we call it *Topic-Dependent contextual polarity*. For example, the word *unpredictable* in an opinion document containing opinion about a film as *unpredictable film plot* will be taken as a positive. In contrary, if the same word is used in another document containing opinion about a digital camera as *unpredictable functional response* then this time it will be considered as a negative word. Hence, a change in term's semantic orientation is observed with the change in topic-class i.e. from movie class to product class.

However, there are few works [118, 139, 174, 178] that have dealt with the problem of local contextual polarities by focusing on negations like *no, not, never,* etc. However, works like [250, 380, 389] also focus on other type of contextual polarities.

Kim et al. [174] propose three models for classifying the sentences as positive or negative using the sentiment orientation of the words present in the sentence. A window based approach was used for calculating the sentiment of a sentence. Four kinds of windows are defined around the holder of the opinion (already identified).

Table 4.2: *The four types of windows defined [174]*

Window 1	Full Sentence
Window 2	Words between holder and topic
Window 3	Window 2 $\pm$ 2 words
Window 4	Window 2 to the end of the sentence

Model 0: This model only considers the polarities of the words in a sentence to decide the semantic orientation (positive or negative) of the sentence. It takes into account the negations like *not and never* that reverse the prior polarity of the word following them.

Model 1: It takes the harmonic mean of the sentiment strengths in the region

Model 2: It takes the geometric mean of the sentiment strengths in the

region

Ku et al. [178] consider the contextual polarity of words while deciding the opinion tendency of sentences and propose the following algorithm for deciding the polarity of the sentence:

○ For every sentence

 * For every sentiment word in this sentence

 * If a negation operator appears before, then reverse the sentiment tendency.

○ Decide the opinionated tendency of this sentence by the function of sentiment words and the opinion holder as follows:

$$S_p = S_{OpinionHolder} \times \sum_{j=1}^{n} S_{w_j} \qquad (4.7)$$

Where S_p, $S_{opinionHolder}$ and S_{wj} are sentiment score of sentence p, weight of opinion holder and sentiment score of word w_j, respectively and n is the total number of sentiment words in p. For experimental purposes, documents related to the issue of animal cloning were selected from NTCIR collection (17 documents) and blogosphere (20 documents). The table given below (table 4.3) shows the results.

Table 4.3: *Opinion extraction at sentence level [178]*

Source	NTCIR	BLOGS
Precision	34.07%	11.41%
Recall	68.13%	56.60%
F-Measure	45.42%	18.99%

We can see in the table that this algorithm shows poor performance for precision values. The obvious reason behind this id that the algorithm proposed only considers opinionated relations but not relevant relations. Many sentences, which were non-relevant to the topic animal cloning, were included for opinion judgment. The non-relevant rate reported is 50% and 53% for NTCIR news articles and web blog articles, respectively. Wilson et al. [380] propose some features to automatically identify the contextual polarities of sentimental expressions. They annotated the subjective expressions in Multi-perspective Question Answering (MPQA) opinion corpus with contextual polarity judgments. Annotators were instructed to

tag the polarity of subjective expressions as positive if the expression is positive, as negative if the expression is negative, as both if expression contains both sentiments in it or as neutral if the expression does not contain any opinion . For example, expressions like *I AM HAPPY* will be annotated with Positive tag; expression like *I AM SAD* will be annotated with Negative tag. An example of an annotated sentence is given below:

Besides, politicians refer to good and evil (both) only for purposes of intimidation and exaggeration.

In total, 15,991 subjective expressions from 425 documents (8,984 sentences) were annotated. The kappa value of agreement measurement for this annotation is 0.72 i.e. 82% of agreement was found between annotators. For experimentation, a two-step process was used with machine learning and a variety of features. In the first step, each phrase was classified as neutral or polar while in the second step, all phrases marked as polar in first step were taken and disambiguation of their contextual polarities (positive, negative, both or neutral) is performed with the help of features defined. For both steps, they developed classifiers using the BoosTexter AdaBoost. HM [309] machine learning algorithm with 5,000 rounds of boosting. The classifiers are evaluated in 10-fold cross-validation experiments with 28 features for first step and 10 features for second step. Features defined belong to five different categories i.e. word based features (5), modification features (8), sentence based features (11), structure features (3) and document level features (only 1). The system performed well for both steps against the baselines. For the first step, an increase in accuracy of 1.3% was marked against the baseline while results for second step show that the system identified the contextual polarities with an accuracy of 65.7% beating the baseline by 4.3%.

Further work from Wilson and her colleagues exists [378, 379] among which the sentence level subjectivity detection tool (i.e., Opinion Finder [378] is a very effective tool and is being used in opinion finding research [130, 308]). The work of Wilson et al. [378] closely resembles to work of Nasukawa, Yi, and colleagues [250, 389]. They also classify the contextual polarity of sentiment expressions. They classify expressions that are about specific items, and use manually developed patterns to classify polarity. These patterns are high-quality, yielding quite high precision, but very low recall. Their system classifies a much smaller proportion of the sentiment expressions in

a corpus than ours does.

Xiaowen et al. [96] focuses particularly on feature based contextual polarity by proposing a holistic lexicon-based approach. Ding and Liu [106] explores the idea of intra-sentential and inter-sentential sentiment consistency with the help of conjunctions like but, and, however, etc. They proposed to consider both opinion words and object features together, and use the pair (object_feature, opinion_word) as the opinion context. Empirical evaluation of the approach reveals better results.

4.3.4   Document-Level Processing

Different works have focused on different granularity levels while working on the problem of sentiment detection. The details of work on word and sentence level approaches have already been given and now we will discuss document level approaches in this section. Most of the earlier work on document-level sentiment detection is limited to the use of data collections like news articles and product reviews. However, with the popularity of online social networks, various types of data collections have emerged (like collection of blogs and tweets) that have given boost to the research work in this field. For example, a significant increase in interest for research in opinion mining field has been noticed after start of TREC Blog track in year 2006 (see figure 4.1).

In this section, we will discuss the approaches focusing on identifying opinionated documents and classifying them according to their polarities (i.e., positive, negative or neutral). A two-step approach is generally followed by most of the works for the task of opinion detection with very few exceptions (like [15]) that adapt a single step method to identify opinionated documents. In the first step, called *Topical Relevance Retrieval*, a set of relevant documents is retrieved for a given topic. In the second step, called *Opinion Finding step*, the set of relevant documents retrieved during first step are processed and re-ranked according to their opinionatedness. The details of the related work for both approaches are given below in corresponding sections.

Topical Relevance Retrieval   In this step (i.e., Topical Relevance Retrieval), the objective is to retrieve top relevant documents for a given topic. It has been observed that good performances on the opinion find-

ing task are strongly dominated by good performances on the underlying topic-relevance task [52, 53, 260]. Therefore, the conclusion made is that a stronger topical relevance baseline is more likely to improve the results of opinion finding task than a weaker baseline if an effective opinion finding approach is applied.

Various methods have been practiced for topic relevance retrieval in opinion finding approaches. However the Language Modeling (LM) [18, 196, 201, 214], TF*IDF [38, 198] and BM25 [2, 173, 315, 351] have been among the favourites. Language model has been combined with some popular smoothing models like Jelinek-Mercer, Dirichlet, Baysian and Absolute Discounting, etc. For example, the work by Liao et al. [196] estimates the multinomial language model with the help of *Maximum Likelihood Estimator (MLE)* using *Dirichlet smoothing*. Similarly Hoang et al. [201] performed topic-relevance experimentation with *Kullback-Leibler (KL) divergence*. KL is a statistical language model which scores and ranks documents by the KL-divergence (i.e., relative entropy) between the query language model and the document language model [184]. Additionally, they applied Bayesian smoothing method using Dirichlet priors with default prior parameter set to 1,000. Overall, Language Models have performed well for the retrieval task [18, 231, 365, 388]. For example, top ranked KLE group for TREC 2008 topics [388] deployed a passage-based retrieval language modeling approach for topic relevance retrieval.

Many of the researchers take support of query expansion for improvement of the results of topic relevance. Sometimes this support was provided by Pseudo Relevance Feedback [120, 279], sometime by Wikipedia [68], sometimes by the web [143] and few used a combination of all of them to expand the query [367, 387].

There are various search engine toolkits that have been used for topic relevance retrieval by many and Lucene [279], Lemur [198], Terrier [2] and Indri [18] are among the most popular. Some have used proprietary tools developed at their labs [173, 351, 385]. Indri is reported to give better retrieval performance than other systems [365] because of the ease it provides for using query models.

It has been seen that many approaches have used various kinds of topic-relevance methods to obtain a set of relevant documents. Knowing that performance of an opinion finding approach depends on performance of topic-relevance baseline, it becomes meaningless to compare two opinion

finding approaches using two different topic-relevance baseline. This is the reason TREC Blog track provided five standard topic relevance baseline runs (chosen from the baselines submitted by participants for topic relevance task) to its participants of TREC 2008 to evaluate the performance of different approaches on common baselines which can give better idea of effectiveness of an approach. The details of these baselines are given below in tables 4.4 and 4.5. It is to be noted here that an *Automatic Run* involves no human intervention at any stage while in a Manual Run, queries could be extended or modified manually. Similarly, a *Title Only Run* is a run in which only title of the topic is used while in a *Title-desc Run*, information from two parts of the topic i.e. title and description is used to generate the run.

Table 4.4: *TREC provided baselines' details [53]*

Baseline	Run Type	Topics
baseline1	Automatic	Title Only
baseline2	Automatic	Title-desc
baseline3	Automatic	Title-desc
baseline4	Automatic	Title-desc
baseline5	Manual	Title Only

Table 4.5: *TREC provided baselines' Relevance and Opinion MAP (over all 150 topics from year 2006 to 2008) [53]*

Baseline	Relevance		Opinion Finding	
	MAP	**P10**	**MAP**	**P10**
baseline1	0.3701	0.7307	0.2639	0.4753
baseline2	0.3382	0.7000	0.2657	0.5287
baseline3	0.4244	0.7220	0.3201	0.5387
baseline4	0.4776	0.7867	0.3543	0.558
baseline5	0.4424	0.7793	0.3147	0.5307

We have summarized the work of TREC Blog participants (from year 2006 to 2008) in the form of a table (see appendix A) in context of several characteristics (like subjectivity lexicon, external data collections, relevance feedback method, etc.) of experimentation process. This table also summarizes the topic-relevance approaches used by TREC Blog participants.

Opinion Finding   In this section, we will discuss work related to two document-level opinion related tasks (i.e., opinion detection and opinion

polarity detection). Various approaches have been proposed for the task of opinion finding. The degree of diversity that is found in these approaches can be estimated by the classification of the work of opinion finding as done by Ounis et al. [260]:

- **Lexicon Based Approaches:** In this type of approaches [129, 231, 367, 385], the support of sentiment lexicons is taken to decide about the subjective nature of the document. These lexicons are either manually or automatically built (using external or internal corpus) or assistance from already available lexicons (like SentiWordNet [100], General Inquirer [404], etc.) is leveraged.

- **Shallow Linguistic Approaches:** There are few approaches [2, 131] that exploit the inter-word syntactic relationships or other features related to Parts of Speech (POS) to estimate the subjectivity of a document.

It has been observed that opinion detection approaches that create their own sentiment lexicons [129, 131, 172, 368] using some opinion data collection perform better than those using ready-made available lexicons [231, 315, 385]. Below we discuss the related work for document-level opinion finding from different perspectives. Globally, we discuss the related work with respect to the lexical resources and machine learning techniques used. However, we also discuss the major data collections used for opinion finding task and the role relevance feedback has performed for this task. In addition to this, we acknowledge the importance of TREC Blogs opinion finding task by summarizing its key findings over years.

Using Corpus-Based Dictionaries

In this section, we discuss the approaches that use an opinion lexicon for identifying opinionated documents. These lexicons may be explicitly prepared by using the given test corpus (or some external corpus) or one can use ready-made lexicons [100, 230] especially available for such kind of tasks.

Using Internal Corpus-Based Dictionaries   There are few works [109, 129, 132] that have used the target collection itself to build the opinion lexicons which were to be used for opinion finding task. For example, He

et al. [129] automatically created a lexicon of opinionated words with the help of Skewed Query Model [54] from the document collection (TREC Blog 2006 collection) they used for experimentation. Skewed Query Model was used to filter out too frequent or too rare terms in the collection. The terms are ranked in descending order by their collection frequencies using the skewed model. The terms, whose rankings are in the range (s·#terms, u·#terms), are selected to be part of the resulting dictionary. #terms are the number of unique terms in the collection. s and u are parameters of the skewed model (with values $s = 0.00007$ and $u = 0.001$). For weighting the terms, they adopted *Divergence from Randomness (DFR)* mechanism which measures the divergence of a term's distribution in pseudo-relevance set from its distribution in the whole collection. The weighting model used is the *Bo*1 *term weighting model* based on the Bose-Einstein statistics which measures how informative a term is in the set of relevant opinionated documents i.e. $D(opRel)$ against the set of relevant documents i.e. $D(Rel)$ [187]. According to this model, the weight of a term t in the opinionated document set $D(opRel)$ is calculated as:

$$w_{opn}(t) = tf_x . \log_2 \frac{1+\gamma}{\gamma} + \log_2(1+\gamma) \qquad (4.8)$$

Where tf_x is the frequency of the term t in opinionated documents and γ is the mean of the assumed Poisson distribution of the term t in the relevant documents and is calculated as:

$$\gamma = tf_{rel}/N_{rel} \qquad (4.9)$$

Where tf_{rel} is the of the term t in relevant documents and N_{rel} is the number of relevant documents. Top X terms are selected to make them part of the final query. Finally opinion scores of the documents are retrieved using BM25 or PL2 DFR model. The final ranking of the documents is done with combination of opinion and relevance score (obtained with original unexpanded query) of the documents. This approach managed to improve the TREC strongest baseline of that time [18] and further all improvements were statistically significant according to the Wilcoxon test at 0.01 level. Similarly, Shima et al. [109] chose not to rely on external lexicons of opinionated terms, but investigate to what extent the list of opinionated terms can be mined from the same corpus of relevance/opinion assessments that are used to train the retrieval system. They calculate the opinion score

of a term t by taking ratio (Weighted Log-Likelihood Ratio [252, 254]) of relative frequency of the term t in set of opinionated documents (set O) to the set of relevant documents (set R and $O \subset R$). The weight of the term t is calculated using two ways i.e. *Likelihood Ratio (LR)* and *Weighted Log-Likelihood Ratio (WLLR)* as given below:

$$Opinion_{LR}(t) = \frac{p(t|O)}{p(t|R)} \tag{4.10}$$

$$Opinion_{WLLR}(t) = p(t|O) \cdot log\frac{p(t|O)}{p(t|R)} \tag{4.11}$$

where $p(t|O)$ is

$$p(t|O) = \frac{\sum_{d \in O} c(t,d)}{\sum_{d \in O} |d|} \tag{4.12}$$

and similarly $p(t|R)$ is given below:

$$p(t|R) = \frac{\sum_{d \in R} c(t,d)}{\sum_{d \in R} |d|} \tag{4.13}$$

where $c(t,d)$ represents the the number of occurrences of term t in document d and $|d|$ is the total number of words in the document.

The above equations quantify the dissimilarity between two sets of documents (i.e., O and R just like *Kullback-Leibler divergence* [218]). In order to calculate an opinion score for an entire document, average opinion score over all the words in the document is calculated as:

$$Opinion_{avg}(d) = \sum_{t \in d} Opinion(t) \cdot p(t|d) \tag{4.14}$$

where $p(t|d) = c(t,d)/|d|$ is the relative frequency of term t in document d.

While most of the manually/automatically built subjectivity lexicons provide just a list of subjective words without any subjectivity scores associated with them (like in [169]), others lexicons like SentiWordNet (SWN) provides the positive and negative scores for each synset of the WordNet or some provide the gradability (i.e., strong or weak) of the subjective words [379].

Using External Corpus-Based Dictionaries   There are many who took the support of external opinionated data collections for building their own lexicons. There is always a trade-off between domain independency and performance in building a lexicon from external data collections (i.e., a lexicon built using external data collection tend to be more generalized but a bit poor in performance relative to a lexicon built from the given test data collection). Yang et al. [144] creates the simplest form of dictionary created through web. This dictionary created, was composed of of positive and negative verbs and adjectives was downloaded from the web. Finally manual selection was used to shorten the list so that it is short enough to not to lengthen the retrieval time too much. This short dictionary is shown in the table 4.6.

Table 4.6: *Opinion Word Dictionary*

Positive Verb	Negative Verb	Positive Adjective	Negative Adjective
love, like	hate, dislike	Good, best, better, happy, extraordinary, successful, glad, desirable, worthy, remarkable, funny, lovely, entertaining, decent, beautiful, fascinating, brilliant, gorgeous, perfect, nice, fantastic, impressive, fabulous, amazing, desirable, excellent, great, awesome, splendid, distinctive	bad, awful, suck, worse, worst, poor, annoying, stupid

Similar to Yang et al. [144], Seki et al. [169] adopt a very simple approach to build a lexicon of opinion terms from reviews of `www.amazon.com`. They explored to use 27,544 positive/negative customer reviews harvested from `www.amazon.com` in order to find good sentiment terms as features. Another work that make use of external sources for building an opinion lexicon is [131]. They prepared a lexicon of 12,000 English words derived from various linguistic sources which gave an improvement of 15.8% over its baseline.

Using Ready-Made Dictionaries

Use of domain-independent ready-made dictionaries is very common in the field of opinion mining. Dictionaries like General Inquirer, SentiWord-Net, and WordNet Affect, etc., are available to researchers for this task. Many [170, 231, 385] have used the lexicon General Inquirer (GI) for their work related to opinion finding. General Inquirer is a large-scale, manually-constructed lexicon. It assigns a wide range of categories[4] to more than 10,000 English words. The categories assigned are Osgood's semantic dimensions and emotional categories. The following word categories are used as indicators of the existence of an opinion in the text: the two valence categories i.e. Positive and Negative; the emotional categories (i.e., Pleasure, Pain, Feel, Arousal, Emot, Virtue, and Vice); the pronoun categories (i.e., Self, Our, and You); the adjective categories (i.e., *IPadj* (relational adjectives) and *IndAdj* (independent adjectives)); and the Respect category. For example, Positive and Negative categories of GI contain 1,915/2,291 terms having a positive/negative orientation. Examples of terms in positive category are advantage, fidelity and worthy, etc., while examples of negative terms are badly, cancer, stagnant.

Gilad Mishni [231] used the words present in different categories of General Inquirer (GI) to predict the subjectivity of a blogpost. A subset of topic relevant sentences is selected from each document to check for the occurrences of opinion words (from GI) within them. An almost similar use of the General Inquirer is noted by Liao et al. [385]. Liao et al. [385] used the lexicon General Inquirer for performing the task of document polarity detection. Their system was trained on TREC 2006 data collection using DragPush machine learning classifier. They compared the results of lexicon

---

[4]A complete list of the General Inquirer categories is given at `http://www.wjh.harvard.edu/inquirer/homecat.htm`

based run with another run based on multidimensional representation of the blogpost using a bag of words approach. The results show that lexicon based approach performed way better than bag of words approach. Moffat et al. [170] simply use the positive and negative categories of GI for the task of document sentiment classification. They combined GI's positive and negative categories with words from various other sources [128]. Experiments with SVM classifier show that results were improved by addition of external list of subjective words (accuracy from 0.803 to 0.820).

The use of sentiment lexicons is very helpful for the tasks related to opinion detection but there is a need for more sophisticated lexicons and techniques that can get benefit of the information these lexicons are providing. Simply counting the occurrences of the opinion words in a document to calculate the document's subjectivity is not an optimal solution and is subjected to many drawbacks. Given two subjective words, one might be stronger in its subjectivity than the other one. Intuitively, a document containing stronger subjective words should be ranked higher than a document with equal number of subjective words but with lesser subjectivity. Therefore such a lexicon is needed that not only categorize the words as positive or negative but also assigns subjectivity scores to the words to avoid the problem mentioned above.

SentiWordNet (SWN) [100] solves the problem mentioned above by providing objective and subjective (i.e., positive and negative subjectivity scores) scores for each synset of the WordNet. The range of scores lie in interval $[0, 1]$ and sum of all three scores equals 1. Few approaches [2, 315, 400] showed their interest in using SWN as a lexical resource. All of these approaches sum the opinion scores of the words in a document to calculate the opinion score for that document. Zhang et al. [400] fixed a threshold value of 0.5 for an adjective to be considered as subjective. Zhao et al. [315] follow a similar approach but on document level (i.e., if

$$
P(d) \begin{cases} \geq 0.4 \text{ then document d is positive} \\ \leq 0.2 \text{ then document d is negative} \\ 0.2 < P(d) < 0.4 \text{ then document d is neutral} \end{cases} \tag{4.15}
$$

Where $P(d)$ is the document's subjectivity score.

A question which creates space into our mind is that which sense of the word to be considered when using these subjectivity scores from SWN. A

word may have more than one senses (like the word good have 27 senses in WordNet: 4 as a Noun, 21 as an Adjective and 2 as Adverb) then how to determine which sense of the word is being used in a particular context because each sense might have different subjective and objective scores. It is very unfortunate that most of the opinion finding approaches did not work too much on this problem of sense disambiguation but have tried to deal with it using very simplified statistical approaches. For example, Bermingham et al. [2] considered the positive SentiWordNet score for a word w to be the mean of the positive scores for all the word senses of that word i.e.,

$$S_{pos}(w) = \frac{1}{n} \sum_{i=0}^{n} \left(\frac{1}{m} \sum_{k=0}^{m} (PosSWN_{i,k}) \right) \tag{4.16}$$

where n is the number of synsets the word appears in, m is the number of word senses in the synset for that word and $PosSWN_{i,k}$ is the positivity score for word sense k in synset i for word w. The positive score for a document is the mean $S_{pos}(w)$ for all words in the document and is given by:

$$Score_{pos}(d) = \frac{1}{p} \sum_{i=0}^{p} S_{pos}(w_i) \tag{4.17}$$

for a document d with p words. The negative score of the document is calculated similarly.

Text Classification Approaches

Text classification approaches [16, 30, 43, 68, 143, 151, 169, 265, 305, 400] generally make use of some machine learning classifier trained on already annotated opinionated data and then is tested on test data. Most of the commonly used classifiers for opinion detection in blogs are Support Vector Machines (SVM) [16, 120, 134, 151, 153, 169, 170, 283, 316, 326, 368, 402], Logistic Regression Classifier [68, 400] and Maximum Entropy classifier [143].

SVM has been the most preferred machine learning classifier because SVMs are reported to perform better as compared to other machine learning algorithms. Most of the approaches have proposed very simple features for the opinion related tasks. The major ones used:

○ The number of subjective words in a document d,

○ The number of positive and negative words in a document d,

○ The number of subjective sentences in a document d,

○ The number of positive and negative sentences in a document d,

○ The proximity approach (i.e., a fixed number of sentimental words around the topic words in a document or the fixed number of words around adjectives, verbs or adverbs),

○ The use of punctuations like smiley faces : or sad faces 9, etc.,

○ The sum of the classification scores of the sentences in d that are classified to be positive relevant,

○ The sum of the classification scores of the sentences in a document d that are classified to be negative relevant,

○ Average score of classified positive relevant sentences in d,

○ Average score of classified negative relevant sentences in d,

○ The ratio of the number of the classified positive relevant sentences in d, to the number of the classified negative relevant sentences in d,

○ The ratio of the sum of the scores of the classified positive relevant sentences in d to the sum of the scores of the classified negative relevant sentences in d,

Role of External Data Collections

Many opinion finding approaches seek help of some external data collection whether for query expansion or for training the classifier for opinion detection task. An external data collection means the data collection other than the one used for evaluation of an approach. The most common and popular data collections used for training the machine learning classifiers are movie review data provided by Pang and Lee [263, 265] and customer Review Data provided by Hu and Liu [139]. The movie review data includes 5,000 subjective sentences and 5,000 objective sentences. The subjective sentences are sentences expressing opinions about a movie. The objective sentences are descriptions or the storytelling of a movie. The customer review data contains 4,258 sentences in total with 2,041 positive examples and 2,217 negative examples. The customer reviews are from

Table 4.7: *Document-Level Summarization of Work in Context of Collections and ML-Classifiers used*

Title of the Paper	ML-Classifier	Data Collection
Customizing sentiment classifiers to new domains: A case study [16]	Naive-Bayes and Support Vector Machines	Pang and Lee (2004) movie review data set (2000 reviews), book review data of 100 positive and 100 negative reviews, Product Support Services web survey data with 2564 examples of positive feedback and 2371 examples of negative feedback, Knowledge Base web survey data They consist of 6035 examples of bad feedback and 6285 examples of good feedback.
The Sentimental Factor: Improving review classification via human-provided information [30]	Naive-Bayes	Pang and Lee (2002) movie review data set (1400 reviews)
Sentiment Classification of Movie Reviews Using Contextual Valence Shifters [170]	Support Vector Machines	Pang and Lee (2004) movie review data set (2000 reviews)
Which side are you on?: identifying perspectives at the document and sentence levels [199]	Naive-Bayes	http://www.bitterlemons.org 591 articles
Sentiment Classification Using Word Sub-sequences and Dependency Sub-trees [223]	Support Vector Machines	Pang and Lee (2002) movie review data set (1400 reviews) and Pang and Lee (2004) movie review data set (2000 reviews)
Sentiment analysis using support vector machines with diverse information sources [244]	Support Vector Machines	Pang and Lee (2002) movie review data set (1400 reviews)
A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts [262]	Naive-Bayes, SVM	5000 movie review snippets (e.g., bold, imaginative, and impossible to resist?) from www.rottentomatoes.com 5000 sentences from plot summaries available from the Internet Movie Database (www.imdb.com).
Thumbs up? Sentiment classification using machine learning techniques [265]	(Naive Bayes, maximum entropy classification, and support vector machines	Pang and Lee (2002) movie review data set (1400 reviews)
Using emoticons to reduce dependency in machine learning techniques for sentiment classification [284]	Naive-Bayes, SVM	Pang and Lee (2002) movie review data set (1400 reviews), Internet Movie Review Database archive of movie reviews, Emoticon corpus
Automatic Opinion Polarity Classification of Movie Reviews [305]	Naive Bayes and Markov Model	Pang and Lee (2002) movie review data set (1400 reviews)
Using Appraisal Taxonomies for Sentiment Analysis [369]	SVM	Pang and Lee (2004) movie review data set (2000 reviews)
Sentiment extraction from unstructured text using tabu search-enhanced Markov blanket [384]	Markov Blanket Classifier, SVM, Naive-Bayes, Max. Entropy, voted Perceptron	Pang and Lee (2002) movie review data set (1400 reviews)
Automatic extraction of opinion propositions and their holders [37]	Naive-Bayes	FrameNet is a corpus of over 100,000 sentences, PropBank is a million word corpus consisting of the Wall Street Journal portion of the Penn TreeBank that was then annotated for predicates and their arguments.
Mining the peanut gallery: Opinion extraction and semantic classification of product reviews [83]	Naive-Bayes	C-Net and Amazon customer reviews

`www.amazon.com` about 5 electronic products including digital cameras, DVD players and jukeboxes.

Yang et al. [144] used a passage based retrieval approach and retrieved 1,000 passages for each query. Logistic regression was used to predict the subjectivity of each sentence in a passage. The logistic regression binary classifier predicted labels Y for test set sentence S, $Y = 1$ when S is an opinion and $Y = -1$ when S is an objective sentence. Logistic regression model was trained using the Pang et Lee movie review data [263, 265] and Hu and Lie [139] customer review data. Similarly Zhang et al. [279] calculate the subjectivity score of each sentence using a CME classifier trained on movie review data [263, 265] using unigram, bigram features of a sentence. SVM classifier then predicts the opinion score of each blogpost on behalf of the subjective sentences contained in a blogpost. Almost similar kind of approach was used by Robin et al. [295] for opinion finding task using movie review data of Pang et al. with other data sources with Naïve Bayes Classifier. Besides being used as training data for classifiers, these external sources have been used for expanding the queries or for generating a list of opinionated words (individual terms or phrases). For example, Yang et al. [173] used the Pang et Lee movie review data for building lexicon for their IU module. The phrases or collocations with pronouns I and You were extracted from the movie review data to be used in their IU module. Similarly, Li et al. [38] benefited from Pang et Lee's movie review data, Hu et Lie's customer review data and data of 256 hotel reviews for creating a sentiment lexicon for the task of opinion finding. They used a list of seed words (positive and negative) to compute their co-occurrences statistics with other adjectives. Adjectives that co-occur with any seed word over 10 times were considered as sentiment terms and are made part of the sentiment lexicon. The list of positive and negative seed words is given in table 4.8 while table 4.9 lists a sample of new sentimental adjectives added to the list of seed words.

Table 4.8: *seed words used in [38]*

Positive	good, excellent, wonderful
Negative	bad, poor, terrible

In addition to these data collection, there are others too which have been providing support for several approaches for opinion detection. Few of them are listed below:

Table 4.9: *New added adjectives in seed word's list [38]*

Positive	good, excellent, wonderful, relaxing, glorious, delicious, priceless, decorated, helpful, superb,
Negative	bad, poor, terrible, worse, absent, stupid, problematic, boring, threatening,

- ○ Yahoo Movie Review Data (used in [400])

- ○ Epinion Digital Camera Review data (used in [400])

- ○ Reuters Newswire Data (used in [400]])

- ○ Reviews from www.Rateitall.com (used in [367, 368]

- ○ Reviews from www.amazon.com (used in [169])

- ○ AQUAINT-2 news corpus (used in [18, 131]

- ○ Internet Movie Database plot summaries (used in [295, 387])

- ○ Reviews from Rotten Tomatoes (used in [295])

It is hard to conclude that which external data source has performed well because no data collection has as such given distinctive results consistently. Therefore, we believe that it is not the type of data collection which improves the system's performance but more the way that data collection is being used. After an analysis of the top performing opinion finding approaches, it can be concluded that systems using data collections as a way to expanding the given query or creating an opinion lexicons have performed well. Looking at table 4.7 and table in appendix A could be interesting to have an idea about various data collections.

Role of Relevance Feedback

A general overview of opinion finding approaches reveals an interesting observation about the use of relevance feedback. If we look at the top most effective opinion finding approaches then it can be noted that most of the top performing approaches [18, 231, 365, 367, 368, 388] have benefited

from the use of Pseudo Relevance Feedback on topical retrieval step to have improved topic relevance MAP. Knowing already that the performance of the opinion finding task is dominated by the performance of topic relevance task, it can be suggested that use of Pseudo Relevance Feedback at retrieval step can influence the performance of opinion finding phase.

Major TREC Findings (from year 2006 to year 2008)

After year 2006, most of the opinion finding approaches conducted their experiments with the standard TREC Blog data collections and performed evaluations under TREC evaluation framework. Therefore, in this section we will discuss major findings of the TREC Blog track for years 2006, 2007 and 2008.

TREC 2006

- TREC Blog 2006 overview paper [260] reports that systems retrieve more spam documents at later ranks than earlier ranks. In particular, the average number of spam documents retrieved by all systems in the top 10 documents was 1.3 which also shows the effectiveness of participant's approaches for removing spam blog documents.

- There is no strong relation between the opinion finding MAP performance of systems and their likeliness to retrieve spam. However, as the correlation is not negative, it is not the case that low performing systems were more likely to retrieve spam. This suggests that presence of spam blogs do not markedly affect the retrieval performance of the systems.

- Variable performance was reported on behalf of lexicon-based approaches with some groups observing slight degradation of results compared to their base retrieval scores, and others observing some improvement.

- The success of machine learning approach was limited, possibly because of the difference between training data and the actual opinionated content in blog posts.

- Like Machine Learning approaches, shallow linguistics-based approached could not perform very well too.

○ It was found that the performance of the opinion retrieval is strongly dominated by the performance on the underlying topic relevance task, emphasizing the importance of a strong retrieval baseline.

TREC 2007

○ The retrieval performances of the participating groups in TREC 2007 are noticeably higher than those reported in TREC 2006 on the same task [52]. However it is unclear that whether this is due to the TREC 2007 topics being easier than those used in TREC 2006 or due to the use of more effective retrieval approaches by the participants.

○ A strong correlation is observed between opinion finding MAP and polarity classification of documents (i.e., the systems which are more successful at retrieving opinionated documents ahead of relevant ones, they will then have more documents for which they can make a correct classification). Systems which perform poorly at retrieving opinionated documents are by definition not going to have the chance to classify as many documents correctly, so the strong correlation is expected.

TREC 2008

○ The results show that topics of year 2008 seem easiest [53].

○ The more an opinion finding technique consistently improves the opinion finding retrieval performance of the 5 provided baselines, the more likely that it is effective.

○ The TREC 2007 topic set appeared to be the easiest for the retrieval of positive opinionated documents, while the three topic sets (TREC 2006, TREC 2007 and TREC 2008) showed the same level of difficulty when searching for negative opinionated documents.

4.4   Challenges for Opinion Mining

Most of the opinion detection approaches model the presence of subjective words in a given document. They use several methods to identify subjective words and process this information to identify and retrieve opinionated sentences or documents (as discussed above). However, proposing approaches

that can process subjective information effectively requires overcoming a number of challenges. In this section, we discuss the major problems that researchers working in this domain are facing.

4.4.1   Identifying Comparative Sentences

Although most of the opinion detection approaches exploit the presence of subjective words in a document, these are not as simple as counting the number of subjective words in a document. The syntactic and semantic relations between words in a sentence play very important role in this regard. For example, the comparative sentences: *Mobile phone A is better than B* and *Mobile phone B is better than A* convey total opposite opinions. To well understand the meanings of these comparative phrases, an effective modeling of sequential information and discourse structure is required. The use of comparative sentences is very common in product reviews. Product reviews contain opinions of experts about the products, hence are subjective but, on the other hand, comparisons can be subjective or objective. Jindal et al. [156] explains this by giving the following examples of an opinion sentence, a subjective comparison sentence and an objective comparison sentence as shown in table 4.10.

Table 4.10: *A comparison of opinion, subjective comparative and objective comparative sentences*

Car X is very ugly	Opinion Sentence
Car X is much better than Car Y	Subjective Comparison
Car X is 2 feet longer than Car Y	Objective Comparison

We can see that in general comparative sentences use quite different language constructs from typical opinion sentences. Identification of comparison sentences is challenging because although there are few indicators that can help to identify such sentences (i.e. comparative adverbs and comparative adjectives like *better, longer,more* etc.) but such indicators are also present in sentences that are not comparative, e.g., *I do not love you any more*. Similarly, many sentences that do not contain such indicators are comparative sentences, e.g.,*Cellphone X has Bluetooth, but cellphone Y does not have* [156].

Jindal and Liu [156] take a data mining approach to identify the comparison sentences. They use class sequential rule (CSR) mining with supervised

learning approach to identify comparative sentences in customer reviews, forum discussions, and news articles. They prepare a list of words using WordNet [230]. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

The list of words prepared by Jindal and Liu are used to express comparisons (like *prefer, superior, outperform, beat,* etc.). Their approach successfully identifies almost all of the comparative sentences with precision of 79% and recall of 81%.

Hou and Li [136] apply another data mining technique, *Conditional Random Fields (CRF)* to a manually annotated corpus of Chinese comparative sentences. They identify six semantic parts of comparative opinion: Holder, Entity 1, Comparative predicates, Entity 2, Attributes, and Sentiment, and extract them using Semantic Role Labeling (SRL), a statistical machine learning technique [110]. They achieved maximum precision of 93% for recognizing and labeling of comparative predicates. The results show the effectiveness of SRL method for mining Chinese comparative sentences.

4.4.2   Leveraging Domain-Dependency

The performance of effective opinion mining approaches [40, 98, 144, 261, 284] differ from domain to domain [202]. For example, the opinion finding approach of Seki et al. [314] performed exceptionally well for "products" related topics while it fails to give good results for topics of type "politics" and "organization". The one major and obvious reason is the difference in vocabularies across different domains. Developing domain-based approaches (or topic-based approaches) might give an edge as far as their performance is concerned but this performance is achieved on cost of its generalization. On the other hand, a domain-independent approach (or topic-independent approaches) is more generalized but might suffer from low performance. Therefore, developing an opinion finding approach that maintains its generalization and gives better performance is a big challenge for researchers working in this domain. An approach which can combine both types of approaches to benefit from positive points of both will be an ideal solution.

There exists a lot of work in the literature for both kind of approaches. Owsley et al. [261] show the importance of building a domain-specific classifier. Read [284] reports that standard machine learning techniques for opinion analysis are domain-dependent (with domains ranging from movie reviews to newswire articles). Na et al. [247] proved that building a query-specific subjectivity lexicons helps improving the results for opinion finding task. They prepare a domain-dependent subjectivity lexicon after updating a domain-independent lexicon by observing top-retrieved documents according to how a given word frequently occurs in documents with high degrees of subjectivity. Their work got significant improvements over baseline.

Similarly, there exist few approaches that exploit domain-independent features for the task of opinion mining. Yang et al. [144] take the following simple approach to domain transfer: they find features that are good subjectivity indicators in both of two different domains (in their case, movie reviews versus product reviews), and consider these features to be good domain-independent features. Blitzer et al. [40] explicitly address the domain transfer problem for sentiment polarity classification by extending the structural correspondence learning algorithm (SCL) [10], achieving an average of 46% improvement over a supervised baseline for sentiment polarity classification of 5 different types of product reviews mined from www.amazon.com. Topic-independent approach of Zhang et al. [400] use external data collections (Yahoo! Movie Reviews[5], Epinions[6] Digital Camera reviews and Reuters newswire) for training their logistic regression classifier. Bigrams and trigrams extracted from these data collections form the feature space. Wang et al. [363] used a set of domain-independent features to train a neural network for the task of opinion detection which includes:

○ document length,

○ number of positive words,

○ number of negative words,

○ number of objective words,

○ ratio of positive words to total number of words in document (after stop word removal, same below),

○ ratio of negative words to total number of words in document, and

---

    ○ ratio of objective words to total number of words in document.

Liao et al. [196] and Mishni et al. [231] use positive and negative categories of the lexicon General Inquirer [331] for the task of opinion finding while Seki et al. [169] use positive/negative customer reviews of www.amazon.com for the same task.

In chapter 8, we present our approach for opinion detection which combines both topic-dependent and topic-independent approaches and outperforms the results of the previous best published results for topics of TREC Blog 2007.

4.4.3   Opinion-Topic Association

A document can contain information about many topics and might have opinions on many of them too. In this situation, determining the opinion on a given topic requires a very effective approach which should not only separate opinionated information from factual information but also look for opinion-topic associations in the documents. Processing the documents on sentence and passage level might be a good idea to help solve this problem of finding opinion-topic associations. Various techniques have been used in the past to find this association between the given topic and the corresponding opinion; here we will discuss some prominent work done in this regard.

Natural Language Processing (NLP) techniques (like *POS Tagging* and *Syntactic Parsing*) have been used to identify opinion expressions and analyze their semantic relationships with the topic [249]. POS tagging can be helpful to disambiguate polysemous expressions (such as the word *like*) which assists in identifying the correct sense of an ambiguous word to relate an opinion expressions with the topical terms. Similarly, syntactic parsing is used to identify relationships between sentiment expressions and the subject term. In their approach, Jeonghee et al. [389] extract ternary expressions (T-expressions) and binary expressions (B-expressions) from text, in order to find opinion-topic association. For each opinion expression detected, its target and final polarity (positive or negative) can be determined by sentiment pattern database which contains sentiment extraction patterns for sentence predicates. If no corresponding sentiment pattern is available, the B-expressions can be created for making the sentiment assignment. From a T-expression, sentiment of the verb (for sentiment

verbs) or source (for trans verb), and from a B-expression, sentiment of the adjective, is assigned to the target.

Besides NLP techniques, there exist approaches (like [15, 151, 307]) that have been using proximity-based techniques for finding the opinion-topic associations in textual documents. For example, Santos et al. [307] hypothesized that the proximity of the query terms to the subjective sentences in the document helps to find that level of opinion-topic association necessary for opinion finding task. In the first step, they propose two approaches to select a set of subjective sentences. One approach is based on NLP-based subjectivity classification and second one is a dictionary-based approach. In their first approach of NLP-based classification, they used *OpinionFinder*, a subjectivity analysis system which provides information about opinions expressed in text and also who expresses them. OpinionFinder employs a Naive Bayes classifier to distinguish between objective and subjective sentences. In their second approach for selection of subjective sentences, a dictionary of subjective terms is automatically derived from the target. This list of terms is ranked with terms' within-collection frequencies and then it is filtered for too common and too rare terms. Using a training set of queries, the remaining terms from the list are weighted based on the divergence of their distribution in the set of opinionated documents retrieved for these queries against that in the set of relevant documents retrieved for the same set of queries.

During retrieval time, an aggregated subjectivity score sw is calculated for each sentence s of a retrieved document d as:

$$sw(d, s) = \frac{1}{|s|} \times \sum_{t \in s} w(t) \qquad (4.18)$$

where $t \in s$ corresponds to the set of all terms t in sentence s, $|s|$ is the number of terms in s, and $w(t)$ corresponds to the weight of term t according to the generated dictionary of subjective terms. Sentence subjectivity weight is normalized by the total number of sentences in a document. Finally, sentences with a weight greater than a predefined threshold are considered as subjective sentences. Given a retrieved document d and a set of subjective sentences S_d, the score of document d with respect to a query Q is boosted according to the following linear combination:

$$score(d, Q) = \lambda_1 \times score(d, Q) + \lambda_2 \times \sum_{t \in Q} \sum_{s \in S_d} prox(t, s) \qquad (4.19)$$

where $score(d, Q)$ is the score of the document d retrieved against a query Q, $t \in Q$ corresponds to the set of all query terms, $s \in S_d$ is the set of all subjective sentences in document d, $prox(t, s)$ is the proximity score assigned to the query term t and the subjective sentence s in document d, and λ_1 and λ_2 are free parameters of the linear combination.
$prox(t, s)$ is calculated as shown in equation 4.20:

$$
\begin{aligned}
prox(t, s) = w\,(q) \times sw\,(d, s) \times \frac{1}{pf + 1} \times [&-\log_2(wc + 1) \\
&+ \log_2(pf + 1) \\
&+ \log_2(wc - pf + 1) \\
&- pf \times \log_2 \frac{1}{wc} \\
-(wc \times pf) \times \log_2 &\left(1 - \frac{1}{wc}\right) \qquad (4.20)
\end{aligned}
$$

where $w\,(q)$ and $sw\,(d, s)$ are the weights of the query term t and the sentence s respectively, $wc > 0$ is the number of windows of size ws sentences in document d where ws is a free parameter and pf is the frequency of the pair (t, s) within windows of size ws sentences in the document.

Santos et al. used TREC Blog 2007 and 2008 data collections with TREC provided 5 baselines (i.e., baseline1, baseline2, baseline3, baseline4, baseline5) for evaluation purposes. Experiments are conducted by combining subjective sentence selection approaches (i.e., OpinionFinder (OF) and dictionary based approach (Dict)) with the proposed proximity approach, hence represented as OFProx and DictProx in results. A comparison of experimental results (i.e., OF vs. OFProx and Dict vs. DictProx) is also reported. Results for TREC Blog 2007 collection show that OpinionFinder (OFProx) significantly improved over its counterpart (OF) in 4 out of the 5 baselines in terms of both topic-relevance and opinion MAP (Mean Average Precision), while our dictionary-based approach (DictProx) was not significantly different from its base approach (Dict) across the considered baselines. Similarly if we compare the results of proposed approach with TREC baselines then it was observed that this approach significantly improve over the baselines in 8 out of 10 cases in terms of topic-relevance

MAP, and in 7 out of 10 cases in terms of opinion MAP.

However the story is different for TREC Blog 2008 data collection. For this data collection, DictProx significantly improved over Dict for 3 baselines in terms of topic-relevance MAP, and for 2 baselines in terms of opinion MAP. More surprisingly, our approach using OpinionFinder (OF) could only significantly improve over its counterpart for baseline5. In addition to this, if we compare the results of this approach with TREC baselines for TREC blog 2008 data collection then it was noted that it significantly outperform the standard baselines in 6 out of 10 possible cases for both topic-relevance and opinion MAP.

Relative to approach proposed by Santos et al. [307], simpler proximity approaches were adopted by Java et al. [151] and Attardi and Simi [15] where they just check for occurences of opinionated terms around the query terms. However, comparison between all these approaches is not possible because all of these approaches use different data collections and baselines. Similarly, a comparison of results for approaches of Java et al. [151] and Attardi and Simi [15] cannot be justified because both approaches use different topic-relevance baselines.

The approaches adopted by Yang et al. [144] and Lee et al. [388] use passages for finding opinion-topic associations within documents. Yang et al. [144] adopt a passage-based procedure for topic-relevance retrieval and sentences for the task of opinion mining while Lee et al. [388] use a language modeling approach for topical-relevance retrieval of documents. For opinion finding task, they prepare a query-specific lexicon using the best passage extracted using the complete-arbitrary passage approach [248] from the top N relevant documents.

We propose our sentence and passage-level approaches for finding opinion-topic associations in documents that are discussed in chapter 6 and chapter 7 respectively.

4.4.4 Feature-based Opinion Mining

A document might be overall positive about a certain topic while it may also contain some negative opinions about few aspects of the topic. For example in review of a digital camera, a reviewer might be overall satisfied with the camera but it is possible that he is not happy with one or two of its features like he might not be satisifed with the size of the screen or is

not happy with quality of its optical zoom. Feature-based opinion mining is considered a big challenge for opinion mining and it involves two tasks, *Feature Extraction* and *Feature-Sentiment Association*. To explain these tasks, let us take the example of the following sentence:

I love picture quality of this camera

In above sentence, *picture quality* is a product feature and *love* is the sentiment associated with it. If a feature appears in subjective text, it is called *explicit feature*. If a feature appears in text other than subjective and is implied then it is called *implicit feature*. For instance in sentence given above, the feature *picture quality* is an explicit feature while *size* is an implicit feature in the sentence given below as it does not appear in the sentence but it is implied [202]:

This camera is too large

Mining implicit feature is harder than mining explicit feature because the feature word is not explicitly mentioned in the text. Li et al. [407] proposed an approach which seeks for feature-opinion pairs to mine explicit and implicit features in a movie review data collection. They use a dependency grammar graph to mine some relations between feature words and the corresponding opinion words in training data. The mined relations are then used to identify valid explicit feature-opinion pairs in test data. For mining implicit feature-opinion pairs, Li et al. dealt simply with very simple case of implicit features with the help of opinion words or phrases appearing in the text. They defined classes of movie domain related features with a set of opinion words allocated to each class (just two shown in figure 4.4). Therefore, when such opinion word is found in a sentence, corresponding feature class can be decided even when a feature word is not mentioned in the sentence.

Opinion words only for feature class OA:
    entertaining, garbage, masterpiece, must-see, worth watching
Opinion words only for movie-related people
    clever, masterful, talented, well-acted, well-directed

Figure 4.4: *Some opinion words frequently used for only feature class OA (overall) or movie-related people*

Yi et al. [389] proposed two feature term selection algorithms based on

a mixture language model and likelihood ratio. Likelihood Test method performed better than language model in their experimentation. Following is principle of the Likelihood Test method: Let D_+ be a collection of relevant documents for a topic T, D_- be a set of non-relevant documents, and bnp a candidate feature term extracted from D_+. Then, the likelihood ratio $-2log\lambda$ is defined as follows:

$$-2\log\lambda = -2\log\frac{max_{p_1 \leq p_2}L(p_1,p_2)}{max_{p_1=p_2}L(p_1,p_2)}$$
$$p_1 = p(d \in D_+|bnp \in d)$$
$$p_2 = p(d \in D_+|\bar{bnp} \in d) \tag{4.21}$$

where $L(p_1,p_2)$ is the likelihood of seeing bnp in both D_+ and D_-. The higher the value of $-2\log\lambda$, the more likely the bnp is relevant to the topic T. For each bnp, compute the likelihood score, $-2\log\lambda$, as defined in formula 4.21. Then, sort bnp in decreasing order by their likelihood score. Feature terms are all $bnps$ whose likelihood ratio satisfying a pre-defined confidence level. Alternatively simply only the top N $bnps$ can be selected.

For instance, Liu et al. [204] proposed a method to extract product features from product reviews (pros and cons) based on association rules. Their method starts with parts-of-speech (POS) tagging of features and replacing them with the word $feature. For example, the sentence

Camera has large Screen.

is converted into $\langle Camera, NNhas, VBlarge, JJ\$feature, NN \rangle$. Duplicates are distinguished by giving them numbers and then word stemming is performed. They then use association mining system *CBA (Classification based on Association)* [137] to extract the rest of the features. Once the features are found, they are grouped using WordNet synsets. For example, words *photo, picture, and image* all refer to the same feature in the digital camera. So, if they are found to be synonymous, they become known synonyms of the same feature.

However, association rule mining is not suitable for this task because association rule mining is unable to consider the sequence of words, which is very important in natural language texts. Thus, many complex ad hoc post-processing methods are used in order to find patterns to extract features. Hu and Liu [138] propose a more principled mining method based

on sequential pattern mining. In particular, they mine a special kind of sequential patterns called *Class Sequential Rules (CSR)*. As its name suggests, the sequence of words is considered automatically in the mining process. Unlike standard sequential pattern mining, which is unsupervised, they mine sequential rules with some fixed targets or classes. Thus, the new method is supervised. If we compare the results of work by Hu and Liu [138] with work of Liu et al. [204], we observe that the technique proposed by Hu et Liu [138] generates comparable results as the association rules of Liu et al. [204]. However, feature extraction using association rules needs a lot of extra post-processing and manual involvement as association rule mining is unable to consider the sequence of words, which is very important for natural language texts. However the sequential pattern based feature-extraction approach proposed by Hu and Liu [138] is thus a more principled technique.

The work of Liu et al. [204] was further improved by Popescu et al. [275] by removing those noun phrases that may not be product features. They proposed an algorithm which evaluates each noun phrase by computing a *Pointwise Mutual Information (PMI)* score between the noun pharse and meronymy (part-of something or member-of something relation) discriminators associated with the product class. PMI has already been defined in equation 4.1.

Given a set of relations of interest, their system calculates PMI between each feature and automatically generated discriminator phrases. For example, *scanner* class would be compared with phrases like *of scanner, scanner has, scanner comes with*, etc. which are used to find components or parts of scanners by searching the Web. The PMI scores are then converted to binary features for a Naive Bayes Classifier, which outputs a probability associated with each feature [101]. In the end, a rich system of features is developed, a part of which is shown in table 4.11.

Table 4.11: *Feature Information*

Explicit Features	Examples
Properties	Scanner Size
Parts	Scanner Cover
Features of Parts	Battery Life
Related Concepts	Scanner Image
Related Concept's Features	Scanner Image Size

But the approaches above discover only explicit features. Implicit features can be extracted using the context of already known features. The rule mining technique described by Liu et al. [204] can be extended to implicit features by tagging each feature-specific template with its respective feature. Once all features have been extracted, the techniques discussed above in opinion-topic association(see section 4.4.3) can be used for associating sentiments with extracted features.

4.4.5 Contextual Polarity of Words

An accurate identification of polarity of words requires a deep analysis of their contexts. The prior polarity of a word is always subjected to changes under the context defined by its surrounding words. The new polarity of the word defined by its context is called its *contextual polarity*. Let us take an example to understand contextual polarity:

*Information Secretary of National Environment **Trust**, Robin Hood, said that Ricky is not a **good** guy.*

Although the word *trust* has many senses that express a positive sentiment, but in above sentence, the word *trust* is not being used to express a sentiment at all and is part of the name of the organization *National Environment Trust*. In other words, the contextual polarity of the word *trust* is neutral in this case relative to its prior polarity which is generally positive. Similarly because of the presence of negation word *not* just before the word *good* which is positive in its prior polarity, the contextual polarity of word *good* is negative.

The context can be defined by negations (like *not good, never right,* etc), by word senses (like the word *plant* can be used as *nuclear plant* or *biological plant*), by the syntactic role of words around the given word (like *killers* vs *they are killers*), by intensifiers (like *very beautiful*), by diminishers (like *little problem*), or even by the domain/topic (like *unpredictable movie plot* is positive while *unpredictable camera functions* is negative) [382]. Polanyi and Zaenen [273] give a detailed discussion of many of the above types of polarity influencers.

There exist few works that have proposed approaches to identify the contextual polarities in opinion expressions [275, 336, 389]. Yi et al. [389] use a lexicon and manually developed high quality patterns to classify contextual polarity. Their approach shows good results with high precision over

the set of expressions that they evaluate. Popescu and Etzioni [275] use an unsupervised classification technique called *relaxation labelling* [145] to recognize the contextual polarity of words. They adopt a three-stage iterative approach to assign final polarities to words. They use features that represent conjunctions and dependency relations between polarity words. Suzuki et al. [336] use a bootstrapping approach to classify the polarity of tuples of adjectives and their target nouns in Japanese blogs. Negations (only *not*) were taken into account when identifying contextual polarities.

The problem with above approaches is their limitation to specific items of interest, such as products and product features, or to tuples of adjectives and nouns. In contrast, the approach proposed by Wilson et al. [380] seek to classify the contextual polarity of all instances of the words in a large lexicon of subjectivity clues that appear in the corpus. Included in the lexicon are not only adjectives, but nouns, verbs, adverbs, and even modals. They dealt with negations on both local and long-distance levels. Besides this they also include clues from surrounding sentences. It was first work to evaluate the effects of neutral instances on the performance of features for discriminating between positive and negative contextual polarity.

4.4.6   Use of Social Features for Opinion Detection

With the spread of opinionated content in online social networks, later has become an important source of opinions. It does not only provide researchers with an opportunity to have a huge amount of real-world data but also a chance to exploit the social and networked structure of these networks for the task of opinion detection. However, identifying potential social evidences in online social networks (like blogosphere) and implementing them for the task of opinion detection remains a big challenge for researchers working in this domain.

Most of related work for opinion mining in blogs have been using content-based evidences [52, 53, 260]. However, there exist few works who have exploited the network structure of blogosphere to identify the most influential and opinionated blogs within it [142, 161, 325]. Song et al. [325] propose an algorithm *InfluenceRank* to identify the most influential opinion leaders within blogosphere. *InfluenceRank* is based on characteristics of the opinion leaders as identified by them. This algorithm rank blogs

according to how important they are in the network and how novel the information they provide. The top blogs ranked by *InfluenceRank* tend to be more influential and informative in the network, and thus are more likely to be opinion leaders. Song et al. present an example to describe the basic principle behind their algorithm. In their example, they show a blog network of seven nodes A, B, C, D, E, F, and G. Blogs A, B, C, and D discuss the same topic (let's say topic is *how to use Riya to find similar faces and objects in images across the web*). Later on blogger of blog E publishes a post about a rumor of Google acquiring Riya, and links to blogs A and C that introduce how to use Riyas visual search. Following blog E, blogs F and G start to discuss this acquisition rumor. In this simple example, blog A and blog E are opinion leaders because they introduce innovative opinions and influence the opinions of other blogs. These opinion leaders capture the most representative opinions in the blog network, and thus are important for understanding and capturing the opinions in this Riya network.
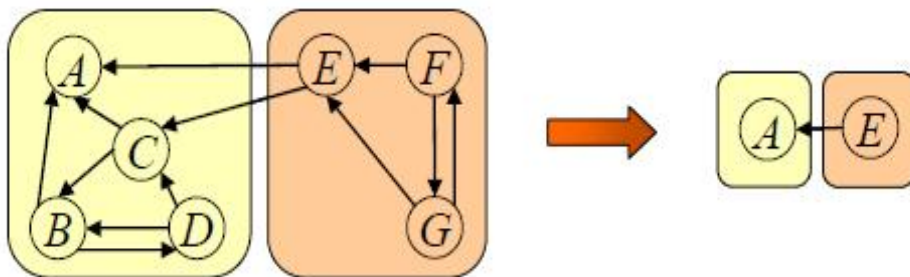
Figure 4.5: *A motivating example: (left) a blog network (right) opinion leaders*

Hui et al. [142] propose a novel method for quantifying sentiment and influence with respect to a hierarchy of topics, with the specific aim of facilitating the computation of a per-topic, influence-weighted sentiment measure. They set a criterion for a blog to be an influential blog as given below. An influential blog:

- has a non-trivial number of followers,
- generates a non-trivial amount of user feedback, in the form of comments on posts, and
- has a large proportion of posts on the topic being analyzed.

The work by Kale et al. [161] presents an approach to model trust and influence in blogosphere using link polarity. Their approaches uses the link

structure of a blog graph to associate sentiments with the links connecting two blogs. Sentiments associated with the links are named as *link polarities*. The sign and magnitude of link polarities are computed by analyzing the text present around the corresponding link from one post of a blog to a post of another blog. Then they take support from trust propagation models to spread this sentiment from a subset of connected blogs to other blogs to generate the fully connected polar blog graph. The approach of Kale et al. somehow resembles to our work but differs on very major points which are given below:

- Our proposed framework includes trust and quality scores associated with a blogger to perform its required tasks while work of Kale et al. considers only trust measure.

- Approach proposed by Kale et al. to compute the link polarity is based on the text present around the corresponding link while our work adopts a twofold approach for the same purpose. We propose to use content-based opinion finding evidences to calculate the sentiment of a particular post for a particular topic. In case of absence of enough content-based evidences (like in case a blog has not any relevant post for a particular topic), our approach proposes to exploit history of related posts (not relevant) or use social evidences to predict the sentiment of a blogger for that specific topic.

We have proposed a framework for opinion finding in blogs which is based on both content and social evidences of blogosphere. It is discussed in detail in chapter 9 on page 229.

4.5   Chapter Summary

In this chapter, we discussed the related work for opinion mining in detail. We classified the work on word, sentence and document level in accordance with the opinion finding process. We also highlighted the TREC Blog track, its tasks and topics.

We have seen that most of the literature for opinion detection is prevailed by the lexicon-based approaches. These subjectivity lexicons whether are already available to the researchers (e.g., General Inquirer, SentiWordNet) or they are readily prepared from the target data collection or some exter-

nal data collection is used for this task. Lexicon based on external data collections have played a very important role to improve the performance of opinion finding approaches but this advantage is traded with loss of generalization of the approach because most of the external data collection used in several approaches are domain-specific. However, a very good choice is available now in the form of TREC Blog data collections that contain data from various domains ranging from sports to politicis. Researchers are taking ful advantage of these data collections by focusing on different challenges of the field of opinion mining. At the end of the chapter, we discussed the related work in context of these major challenges.

Chapter 5

Entity Ranking

The foolish and the dead never change their opinions.
James Russell Lowell

5.1   Introduction

Classical Web Information Retrieval aims at satisfying the user's information need by selecting and providing him/her a list of documents relevant to his/her supplied query (verbalized information need). This process of classical IR is shown in the figure 2.1 in chapter 2 on page 41.

In classical IR, the value of user's information need cannot be ignored. Different users can give different written forms (query) to the same information need but actual intent behind it might remain same. However, different information needs can have different user's intents. Andrei Broder [45] classifies the user's information needs into following three classes according to the intent behind them:

1. **Navigational** The intent behind this kind of queries is to reach a particular web page. A query to find the home page of a site is an example of navigational query.

2. **Informational** Informational queries aims at finding some information about a topic expressed in the form of a query.

3. **Transactional** Transactional queries are more focused on performing a particular task on a certain web page. Examples of such queries include finding gaming servers, online shopping, etc.

However, with the evolution in the nature of the Web, the nature of information needs of users is also changing. Their desire of having relevant documents for satisfying their information need is becoming more and more specific. One of the examples of this change is user's desire of having only the relevant sections of documents in the search results instead of a long list of relevant documents. Similarly, finding relevant named entities instead of/in addition to relevant documents is another major user requirement that is on demand. The classical IR systems cannot satisfy this demand [28, 59, 268]. For example, if a user types a query *Cricket Players of England* in search box of a search engine to find a list of all English Cricket players, it will return a list of documents relevant to English Cricket and then user himself has to extract the name of the players from these relevant documents. Therefore, it seems that a system is needed to find the relevant entities and rank them just like classical IR systems rank documents. The process of finding relevant entities is called *Entity Retrieval*. Kaptein et al. [287] describe the characteristics specifically associated with the process of retrieving entities and hence, making it different from traditional document retrieval. According to them, in entity retrieval:

- ○ returned documents have to represent an entity,
- ○ the returned entity should belong to a specified entity type, and
- ○ an entity should be returned ony once to create a diverse result list

Entity retrieval systems may initially retrieve documents (pertaining to a given topic or entity) but they must then extract and process these documents in order to return a ranked list of entities [270, 318]. This ranking is done with respect to their relevancy with the given topic or given entity and the process of ranking entities on behalf of their relevance is called *Entity Ranking (ER)*. The process of entity ranking includes the process of Entity Retrieval. However sometimes both terms are used interchangeably [353]. In the Information Retrieval (IR) context, entities are more commonly known as *Named Entities*. Different people have defined Named Entities differently like Desislava et al. [269] define it as:

> *A named entity is a semantic category, a pointer to a real world entity*
> *such as a city, an organization, a movie, a book, or a historical event*

While they are more generalized in their definition of an entity, TREC[1]

---

[1]http://ilps.science.uva.nl/trec-entity/guidelines/

gives a more task-oriented definition for an entity which is: *an entity is a person, product, or organization with a homepage*, where an entitys homepage is considered the representative of that entity on the web. Entities have been very important in IR related research and therefore are associated with many IR related tasks like *Entity Extraction from text, Entity Disambiguation, Question-Answering,* etc. Keeping ourselves limited to Entity Retrieval, we define major entity related tasks in the context of entity retrieval.

5.2   Related tasks of Entity Retrieval

In this section, we discuss some major tasks that involve the process of entity retrieval.

5.2.1   Entity Ranking

Given a topic, the objective of Entity Ranking task is to find a list of relevant entities for that topic. For example, to retrieve a list of entities related to the topic of *Information Retrieval.* Related entities are retrieved and are ranked in order of their relevancy scores. Each retrieved entity can be supported with one or more documents with details about that entity.

A variation of this task (known as "Entity List Completion (ELC)" task) [88, 91, 177, 361] can be designed by imposing some constraints on the type of entities to be returned. The returned entity type (generally called "target entity") can be explicitly mentioned or an example entity (known as "source entity") can be provided along with the query. For this new task, above example can be modified as follows: "To return a list of *researchers* working in the field of Information Retrieval". Another example for this task can be to retrieve the *organizations* that are related to Hollywood film star *Tom Cruise.*

Google Sets[2] is an excellent example of this task which can be used to explain both scenarios. It allows us to automatically create sets of items from a few examples. Google Sets identifies groups of related items on the web and uses that information to predict relationships between items. To explain its working, we give examples of two scenarios. In the first scenario,

---

[2]http://labs.google.com/sets
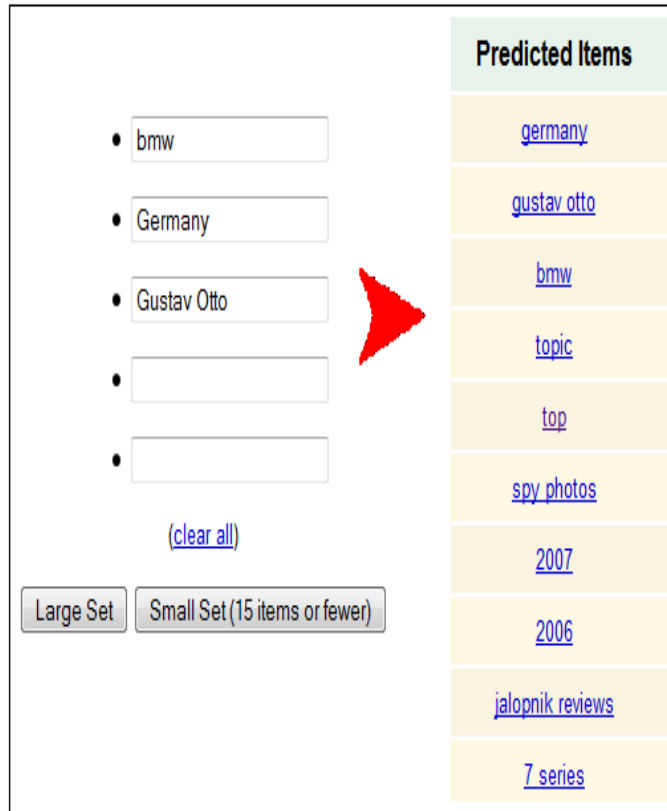
Figure 5.1: *Google Sets: Scenario 1*

we provide three entities *BMW*, Germany and *Gustav Otto* as an input to Google Sets (see figure 5.1). In the second scenario (see figure 5.2), the three input entities are of the same type (i.e., car brands) and therefore, Google Sets returns a set of related car brands because it automatically determines the type of entities to be returned by analyzing the relationship between input entities. On the contrary, Google Sets returns diverse type of entities (among which few are not even correct entities) because even the input entities were related but they were of different types.

This task becomes more complex by adding a restriction on relation between source entity and target entities. For example, "to retrieve a list of *products* that are developed by the *Microsoft Corporation*". This example imposes two restrictions on the returned entity types: First, all returned entities should be of type "Product" and second restriction is to retrieve only those product entities that are developed by the input entity (i.e., Microsoft Corporation).
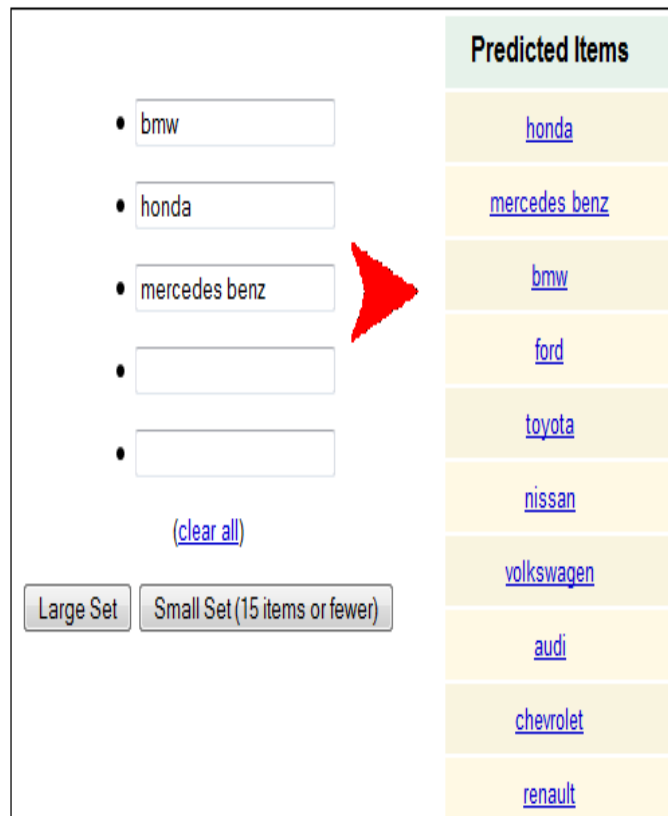
Figure 5.2: *Google Sets: Scenario 2*

5.2.2   Expert Finding

This task aims at finding the answer of the question: "Who are experts on topic X?", and can be considered a more specific form of entity ranking task in which type of the entity to be returned is fixed as "person". This is needed when someone needs the expertise of a person for a certain project or for some other related problem. This search becomes more specific when certain restrictions are applied on type (e.g., only those experts who work as CEO), location (e.g., experts of Information Retrieval in London city), or their history (e.g. experts who have ever worked in Microsoft), etc. There are many commercial expert finders available online (like Askme[3], Linkedin[4], etc.) which shows the popularity of expert finding task in research industry. In figure 5.3, we show the results of *Linkedin* listing the

---

[3]http://www.askme.com/
[4]http://www.linkedin.com/

experts in the field of Information Retrieval. In this figure, the criteria to refine this search can also been seen on its left side.

Figure 5.3: *Linked Results for Experts in IR field*

5.2.3 Entity-level Opinion Detection

Entity-level opinion detection [107] is getting popular in research community where researchers are proposing approaches to know people's sentiments about particular entities. Generally in document-level opinion detection, a document is analyzed to know author's opinion about a particular topic. However it is a fact that a document talks about many entities and it may contain opinion about all of those being discussed in it. Therefore, finding entities in a document which are relevant to a given topic and

extracting the correct opinion associated to those entities is actual task of entity-level opinion detection. A related subtask of this task is known as *Attribute Identification* (also called *Feature Extraction*). It aims at returning a list of key attributes of an entity given as input. For example, if the input entity is a sports car, then the list of attributes to be returned should include its "manufacturer", "top speed, acceleration", "number of seats", etc. Once this list is determined, opinions about these particular attributes of an entity can be extracted using some opinion finding technique.

Besides the tasks discussed above, there are many other tasks where entities play a vital role. For example, Raghavan et al. [282] aimed to gather some information about entities using language models of each entity. Entity models are built and then several methods are applied to these entity models to understand how these models can be applied to to extract information about these entities. Similarly, Meij et al. [226] proposed an approach to suggest query completions using entity and entity type information.

In just few years, a lot of interest by IR community has been shown in the problem of Entity Ranking. Respecting this interest, INEX and TREC both moved forward to provide the interested researchers a common platform for Entity based IR research. In the next two sections, we describe both tracks (i.e., *INEX Entity Ranking (INEX-XER) Track* and *TREC Entity Track*) in detail.

5.3 INEX-XER Track

The INEX-XER track started in year 2007 and continued till year 2009. INEX took initiative to provide a common platform to researchers working in this domain to experiment and evaluate their entity retrieval approaches. The details of three INEX-XER tracks can be consulted in articles [88, 91, 361].

5.3.1 Data Collection

Data collection used for INEX-XER track is Wikipedia XML corpus based on an XML-ified version of the English Wikipedia in early 2006 [93]. The original Wiki syntax has been converted into XML, using general tags of the layout structure (like article, section, paragraph, title, list, and item), typographical

tags (like bold, emphasized), and frequently occurring link-tags. INEX-XER 2007 and 2008 tracks used the same collection as described above while INEX-XER 2009 track uses the new Wikipedia 2009 XML data based on a dump of the Wikipedia taken on 8 October, 2008 and annotated with the 2008-w40-2 version of YAGO [335]. The Wikipedia pages and links are annotated with concepts from the WordNet thesaurus.

Characteristic	Value
Number of Documents	659,338
Size of the Collection	6 GB
Average number of XML nodes per document	161
Average depth of a node in XML tree of the document	6.72

Table 5.1: *Data Collection Details*

5.3.2   INEX-XER Tasks

All INEX-XER tracks from year 2007 to 2009 focused on two main tasks which are: Entity Ranking (ER) task and Entity List completion (ELC) task and both of these are explained below.

Entity Ranking

The objective of entity ranking task is to return a list of entities of a specific type that are relevant to the topic described in natural language text [361]. The entity type to be returned is mentioned in the topic (see figure 5.4). The results consist of a list of Wikipedia pages corresponding to relevant entities. For example, given the topic text as *Cricket Teams* and *Country* as input category, then the results should include the names like Pakistan, Australia, England, India, Sri-Lanka, etc. Of course, the entity type is only loosely defined by its category and correct answers may belong to other categories close to this category in the Wikipedia category graph, or may not have been categorized at all by the Wikipedia contributors.

Entity List Completion

The objective of this task is to complete a given list of example entities. The example entities are mentioned in the topic (see figure 5.5). The results to

```
< inex_topic  topic_id = 9999 >
< title >Impressionist art in the Netherlands< /title >
< description >
I want a list of art galleries and museums in the Netherlands that have impressionist art.
< /description >
< narrative >Each answer should be the article about a specific art gallery or museum
that contain impressionist or post-impressionist art works.
< /narrative >
< categories >
< category >art museums and galleries< /category >
< /categories >
< /inex_topic >
```

Figure 5.4: *INEX-XER Topic format for Entity Ranking Task*

be returned are same as for entity ranking task (i.e., corresponding Wikipedia
pages). For example, when completing the list of Universities for topic *the
list of universities in France* with given examples of *Université de Toulouse,
Université de Lyon*, the system should add other French university names in
the given example list like *Université Européenne de Bretagne*, *Université de
Savoie* etc. While evaluation of this task, only added list of entity types are
taken into account.

Besides these two tasks, a pilot task of *Entity Relation Search (ERS)* [91] was
introduced in INEX-XER 2008. The purpose of this task was to find the relation
between retrieved relevant entities and other related entities. For example, in
above example topic of *Cricket teams* with return entity type of *Country*, it can
be asked to find the name of current captain of each country team retrieved
(i.e., Shahid Afridi for Pakistan, M. Dhoni for India, etc). But participants
did not take too much interest in this task; therefore it was not continued for
INEX-XER 2009. However, we can say that ERS task includes the task of
entity ranking and can be very helpful to explore the connections between IR
and Natural Language Processing (NLP), Question Answering (QA) and the
Semantic Web (SW).

5.3.3   INEX-XER Topics

Each year INEX XER released a set of topics for the participants to perform
experiments with. The details of topics per year are given in table 5.2 while the
format of topics for both tasks defined are shown in figures 5.4 and 5.5.

```
< inex_topic topic_id = 9999 >
< title >European countries where I can pay with Euros< /title >
< description >
I want a list of European countries where I can pay with Euros.
< /description >
< narrative >
Each answer should be the article about a specific European country that uses the Euro as
currency.
< /narrative >
< entities >
<entity id=5843419>France< /entity >
<entity id=11867>Germany< /entity >
<entity id=26667>Spain< /entity >
< /entities >
< /inex_topic >
```

Figure 5.5: *INEX-XER topic format for Entity List Completion Task*

Year	Number of Topics
2007	28
2008	25
2009	35

Table 5.2: *Number of Topics per Year*

5.3.4 Evaluation

The evaluation measure used for INEX-XER 2007 is *Mean Average Precision
(MAP)*. For a set of test topics, MAP is the mean of the average precisions for
all the test topics and is used to evaluate the overall retrieval performance of
an IR system. In 2008, a new evaluation measure *xInfAP* [392] was introduced.
The *xInfAP* is an estimation of *Average Precision (AP)* for the case where the
judgment pool has been built with a stratified sampling approach [85]. This
means that the complete collection of documents is divided into disjoint con-
tiguous subsets (strata) and then documents are randomly selected (sampling)
from each stratum for relevance judgment. In this case it is possible to give
more importance to documents retrieved higher by ER systems (e.g., by having
a complete assessment of top 20 retrieved results) still going down into the list
of retrieved entities (e.g., by having a partial assessment of results retrieved be-
tween rank 30 and 100). The metrics xInfAP is computed exploiting (similarly
to infAP [391]) the estimation of precision at each relevant documents in each

stratum.

5.4   TREC Entity Track

The TREC Entity track [177] started in year 2009 with the same objective as of INEX-XER track (i.e., to create a test collection for the evaluation of entity related research). In the TREC Entity Track framework, an entity is defined as:

A person, product, or organization with a homepage. The entity's homepage is considered the representative of that entity on the web.

5.4.1   Data Collection

The data collection used for TREC Entity track is *Category B* part of the ClueWeb09 collection [177]. The ClueWeb09 data collection (5TB compressed size) was crawled in Jan-Feb, 2009. It covers web data in 10 languages (i.e. Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Portuguese and Spanish). However *Category B* part of the collection only includes English Language web pages and its details are given in table below:

Characteristic	Value
Size	1.0 TB
Number of Pages	50 Million
Unique URLs	428,136,613
Total Outlinks	454,075,638

Table 5.3: *Details of Category-B part (English subset) of collection ClueWeb09*

5.4.2   TREC Entity Track Tasks

In TREC Entity track 2009, the only task proposed was *Related Entity Finding (REF)*. However for TREC 2010, a pilot task (i.e., *Entity List Completion (ELC)*) has been proposed in addition with REF task. In following two subsections, we provide details of topics (total number and formats of topics), input, outputs and evaluation measures for both of these tasks.

Related Entity Finding (REF)

The task of REF is defined as follows:

Given an input entity, by its name and homepage, the type of the target entity, as well as the nature of their relation, described in free text, find related entities that are of target type, standing in the required relation to the input entity [177].

In the TREC Entity track framework, entities are represented by their primary homepages. Therefore, it can be said that searching for entities thus corresponds to ranking these homepages.

Entity List Completion (ELC): Pilot Task

For year 2010, TREC introduce a pilot task, called Entity List Completion (ELC) task and is also known as *Semantic Search* or *Semantic Data Search*. The objective of this task is to find entities in the Semantic Web, or, in other words, to perform entity search in the *Linked Open Data (LOD)* cloud. The problem of the Entity list completion (ELC) task is defined as follows:

Given a list of input entities, represented by their URIs, complete the list with additional entities from a specific collection of Linked Open Data.

5.4.3   TREC Entity Track Topics

REF Task Topics

For TREC Entity track 2009, only 20 topics were created and assessed. However, 50 new REF topics were created and assessed for TREC 2010. For each topic or query, TREC provides following information is provided (see in figure 5.6):

⋄ Input entity, defined by its name and homepage,

⋄ Type of the target entity (person, organization, product, or location),

⋄ Narrative (describing the nature of the relation in free text).

TREC restricts the target entity types to four: people, organizations, products, and locations (entity type *location* was added for TREC 2010 and was missing in TREC 2009). However, it is to be noted that there is no obligation for an input entity to be limited to these four types.

```
< query >
< num >7< /num >
< entity_name >Boeing 747< /entity_name >
< entity_URL >clueweb09-en0005-75-02292< /entity_URL >
< target_entity >organization< /target_entity >
< narrative >Airlines that currently use Boeing 747 planes.< /narrative >
< /query >
```

Figure 5.6: *Information need* find organizations that currently use Boeing 747 planes *is represented in TREC Entity track format*

ELC Task Topics

As stated by TREC, it will use (most of) the 20 topics developed in the 2009 pilot run of the track. For each of these topics, the answer entities identified in the 2009 Entity Track will serve as the list of examples.

Topic definitions follow the same format as for the REF task, with the addition of tags $< examples > .. < /examples >$ that will contain the URIs of known relevant entities, referred to as examples.

5.4.4   Evaluation

Assessment Procedure for REF Task

The assessment procedure is completed in two phases. In the first phase, homepages are judged as primary or relevant. For the primary homepages, the entity name (returned along with the homepage) is judged whether it is correct or not. Then, in phase two, homepages belonging to the same entity are grouped together. The output of the assessments will therefore include a set of homepages and a set of names that all refer to one entity; one or more of these homepages identified as primary, a set of homepages identified as relevant, and one of names identified as correct.

```
< query >
< num >7< /num >
< entity_name >Boeing 747< /entity_name >
< entity_URL >clueweb09-en0005-75-02292< /entity_URL >
< target_entity >organization< /target_entity >
< narrative >Airlines that currently use Boeing 747 planes.< /narrative >
< examples >
< entity >
< URI > http : //dbpedia.org/resource/KLM < /URI >
< URI > http : //www.linkedin.com/companies/klm < /URI >
< URI > http : //www.reference.com/browse/KLM < /URI >
< /entity >
< entity >
< URI > http : //dbpedia.org/resource/Northwest_Airlines < /URI >
< /entity >
< /examples >
< /query >
```

Figure 5.7: *Example Topic for TREC ELC pilot task*

Assessment Procedure for ELC Pilot Task

Judging will be done by participants. Entity resolution (i.e. the same entity represented under different URIs) will be done during the assessment phase. The following measures are used for evaluation in TREC Entity track:

⋄ NDCG@R, where a primary homepage gets gain 3 and a relevant home-page gets gain 1 (note that we reward primary homepages more than last year)

⋄ P@R and MAP, computed for relevance level 1 (both relevant and primary accepted) and 2 (only primary accepted)

Official evaluation results will be based on the homepage field only; alternative rankings of systems will also take entity names into account i.e. accept an entity (homepage) as primary/relevant only if a correct name is also provided.

Data Set: To ease collection building and at the same time simplify partici-pation by the target community, the track will use the Billion Triple Challenge 2009 dataset (`http://vmlion25.deri.ie/`). The same data collection has been used for the Semantic Search challenge posed by the Semantic Search work-shop held at WWW 2010, so should be easy to process for those researchers we specifically organize the pilot task for.

In the rest of the chapter, we will discuss the work related to Entity Retrieval.

5.5   State of the Art

Entity Retrieval is not a very old field in IR domain but it has attracted a lot of attention in the time span of few years. The earlier proposed approaches [45, 59, 60] mainly focus on scaling efficiently on large datasets but not on the effectiveness of search. However, with the passage of time, different entity-related tasks have been defined and worked on by many research groups. In this section, we will highlight some major attempts and contributions for each of the related tasks defined in section 5.2.

5.5.1   Entity Ranking

Dealing with Entity Types

A generalized entity ranking approach aims at ranking all kinds of entities (e.g. persons, locations, places, organizations). However, a more robust mechanism is required for this task which can identify and classify different type of entities. Conrad et al. [75] proposed a framework which was capable to identify entities of types person and organization. In addition, their framework was also able to determine the relationships between entities of both types. Another approach proposed by Vallet et al. [350] not only identifies and classifies the entities in a the test data collection but also in the query itself. The entity types they dealt with include location, person and organization. Their approach extract entities from top ranked relevant passages retrieved against a given query and the entity type (i.e. location, person or organization) associated with most of the entities extracted from relevant passages is assigned to the query. Using their approach, the majority of queries were correctly classified by top entity types.

Role of Wikipedia

Zaragoza et al. [397] discuss the problem of ranking entities of different types. They use Wikipedia as a resource to identify a number of candidate entities. They took support of a statistical entity extractor to identify 5.5 million entities in Wikipedia and created a retrieval index containing both text and the identified entities. Different graph centrality measures are used to rank entities in

an entity containment graph. Also a web search based method is used to rank entities. Here, query-to-entity correlation measures are computed using page counts returned by search engines for the entity, query and their conjunction. Their approaches are evaluated on a self-constructed test collection. Both their approaches outperform methods based on passage retrieval. Kaptein et al. [287] report some interesting findings from their experiments for entity ranking. They found that:

◇ In principle, the problem of web entity ranking can be reduced to Wikipedia entity ranking. Majority of entity ranking topics can be answered using Wikipedia, and that with high precision relevant web entities corresponding to the Wikipedia entities can be found using Wikipedias external links.

◇ The structure of Wikipedia can be exploited to improve entity ranking effectiveness. Entity types are valuable retrieval cues in Wikipedia. Automatically assigned entity types are effective, and almost as good as manually assigned types.

◇ The web entity retrieval can be significantly improved by using Wikipedia as a pivot. Both Wikipedias external links and the enriched Wikipedia entities with additional links to homepages are significantly better at finding primary web homepages than anchor text retrieval, which in turn significantly improved over standard text retrieval.

Entity Related Commercial Products

There exist few search engines that can rank entities of different types using different approaches. The semantic search engine NAGA[5], for example, builds on a knowledge base that consists of millions of entities and relationships extracted from web-based corpora [166]. A graph-based query language enables the formulation of queries with additional semantic information such as entity types. Similarly, search engine ESTER combines information of Wikipedia with ontology search capabilities of YAGO [28]. Another interesting idea proposed and demonstrated in this regard is Yahoo Correlator[6]. Correlator provides new way of running a search query. It extracts and organizes information from text, and searches for related names, concepts, places, and events for a given query.

---

[5]http://www.mpi-inf.mpg.de/yago-naga/naga/
[6]http://correlator.sandbox.yahoo.net/

Similarly, Google also introduced Google Squared[7] which is an experimental search tool that collects facts from the web and presents them in an organized format.

INEX Entity Track Approaches

For INEX entity ranking task, returned relevant entities should be of the same type as mentioned in the topic. Several approaches tried to exploit the Wikipedia category information for this purpose and have been quite successful. Various techniques have been used to compute the similarity between categories of returned entities and target entities. The similarity scores are estimated based on the ratio of common categories between the set of categories associated with the target categories and the union of the categories associated with the candidate entities [356] or by using lexical similarity of category names [355].

Tsikrika et al. [344] use entity graph for the propagation of relevance to neighborhood nodes. The entity graph is actually a query-dependent link graph, consisting of all articles (entities) returned by the initial retrieval as vertices and the link-structure among them forming the edges. Links to other articles not returned in the initial ranking are not considered in the entity graph.

A language modeling based probabilistic framework was proposed by Balog et al. [23] to rank entities. Their model takes into account the probability of a category occurrence and allows for category-based feedback. Finally, in addition to exploiting Wikipedia structure (i.e., page links and categories) [85] applies natural language processing techniques to improve entity retrieval. Lexical expressions, key concepts, and named entities are extracted from the query, and terms are expanded by means of synonyms or related words to entities corresponding to spelling variants of their attributes. Table 5.4 briefly summarizes the INEX participant's approaches for different tasks.

5.5.2   TREC Entity Track Approaches

TREC 2009 participants have approached the entity ranking task in two main steps. First, candidate entity names are extracted, using entity repositories such as Wikipedia, or using named entity recognizers. Link information of the given entity can be used to make a first selection of documents. In a second step,

---

[7]http://www.google.com/squared

Group	Entity Ranking	List Completion	External Tool
L3S (2008) [77]	Combination of NLP and IE techniques	Wikipedia Categories	Two entities are related if they co-occur in a sentence sized window
CSIR at INEX 2008 [154]	Language Modeling	Language Modeling	1) Searching for a list of entities relevant to the topic 2) for each relevant entity e retrieved, finding a group of target entities that have the specified relation with given entity 3) re-rank entity pairs all together.
Uams (2008) [366]	Language Modeling	Language Modeling	X
Uams (2009) [20]	Language Modeling	Language Modeling	X
Waterloo at INEX (2009) [148]	Factoid question answering approach	Factoid question answering approach	X
Guindy [291]	1) Extracting required information from the query and the provided category information 2) Extracting the relevant documents and 3) Ranking the retrieved documents making use of the structure available in the Wikipedia Corpus. Their ranking mechanism combines various approaches that make use of category information, links, titles and WordNet information, initial description and the text of the document.	They used the categories of the given example entities as reference set . This set is compared against the set of categories the retrieved document belongs. The ratio of the match is used to find the similarity between the retrieved entity and the example entities.	X

Table 5.4: *Summarization of INEX Entity Track participant's approaches*

candidate entity names are ranked, and primary homepages retrieved for the top ranked entity names. The University of Glasgow method builds entity profiles for a large dictionary of entity names using DBPedia and common proper names derived from US Census data [281]. At query time, a voting model considers the co-occurrences of query terms and entities within a document as a vote for the relationship between these entities. Purdue University expands the query with acronyms or the full name of the source entity [386]. Candidate entities are selected from top retrieved documents, heuristic rules are applied to refine the ranking of entities. Most of the approaches of TREC 2009 entity track depend heavily on Wikipedia and use it as a a large repository of entity names and types. This is the reason TREC imposed the restriction of not accepting Wikipedia pages as entity homepages for TREC Entity Track 2010.

Table 5.5 and table 5.6 summarize the approaches for TREC 2009 and TREC 2010 entity track participants for REF task. Table 5.6 lacks two fields because the articles for TREC 2010 have not been made available yet by TREC.

5.5.3   Expert Finding

Early research in entity retrieval was more focused on specific kinds of entities, for example ranking "persons" for expert finding task [22]. There exists a lot of work focusing on the task of expert finding.

Yimam-Seid and Kobsa [393] provide an overview of early automatic expertise finding systems. Early expert finding systems were limited to the use of specific document genres like emails [57] or software documentations [242] for expert finding. However, deficiency of these approaches to deal with hetergenous sources of data was tackled by later approaches [78]. Research in Expert Finding made more progress when TREC decided to launch an expert finding task as part of the Enterprise track [256]. It provided a common platform for evaluation of approaches devised for Expert Finding task. TREC defines expert finding task as,

Given a set of documents, a list of candidate names, and a set of topics, the goal then is to find experts from the list of candidate names for each of these topics.

Further details on expert finding can be consulted in proceedings of TREC Enterprise track [19, 321].

Group	Basic Technique	Machine Learning	External Tool
Purdue [390]	Exploit the structure of tables and lists to identify the target entities,	Logistic Regression	Indri, Google search results WordNet
uogTr [281]	Voting Model for people search adapted for ER task. Considers the co-occurrences of query terms and entities in a document as a vote for the relationship between these entities	X	X
UAms [177]	Statistical Language Model built from Window of text in which entities co-occur	X	X
TUDelft [177]	Treats Wikipedia as a repository of entities. Select the top ranked article, or one of the most highly ranked articles (using external links of primary page) which was the most related to the primary homepage of the query entity	X	Lemur+ Dbpedia+yago
UAms (ISLA) [177]	Entity Co-occurrence + Language Model	X	X
BUPTPRIS [177]	2 stage Approach= Retrieve relevant documents + Extract entities of target type from relevant documents	X	Indri + Stanford NER
UAms (Amsterdam) [177]	Exploits Wikipedia information i.e. its links and category information	X	X
BIT [177]	Entity extraction from relevant snippets of text.	Maximum entropy classifier	Indri + Stanford NER
EceUdel [177]	Passage-based retrieval, extract entities from passages and rank using language modeling approach of publication "Probabilistic models for expert finding" by Fang et al.	X	Stanford NER

Table 5.5: *Summary of approaches for TREC-2009 REF Task*

Most of the Expert Finding approaches use two kinds of models, Candidate Model an Document Model [22]. Candidate model based approaches build a textual (usually term-based) representation of candidate experts, and rank them based on query/topic, using traditional ad-hoc retrieval models. The Document model-based approaches, on the other hand, find documents which are relevant to the topic, and then locate the associated experts. Thus, it seems like finding experts using a document retrieval system. Document model based approaches are also called *Query-Dependent approaches* [269]. Nearly all of the

Group	Basic Approach
CARD-UALR(2010) [21]	Used Entity-Entity co-occurrence graph. Given the query entity, relevant entities are extracted based on a novel centrality measure (Cumulative Structural Similarity-CSS) using the intuition that an important entity will share many common neighbors with adjacent entities. Additionally, PageRank, HITS and Ensemble- based approaches are submitted.
FDWIM (2010) [21]	Extract entity with NER tools, Wikipedia and text pattern recognition. Filter with stop list. Features like keywords from narrative, page rank, combined results of corpus-based association rules and search engine are considered.
HPI (2010) [21]	1) Enrich query, 2) Retrieve relevant documents, 3) Extract potential entities, 4) Homepage retrieval. Exploits advanced features of different web search engines for enriching query and homepage retrieval for entities. Genetic learning algorithm used to compute weight of each feature used
Uamsterdam (2010) [21]	Uses Wikipedia as a pivot to search for entities. Wikipedia topic categories are manually assigned to the query topics, To search web entities the external links in Wikipedia are used, and an anchor text index is searched
Waterloo (2010) [21]	Entities extracted from top documents retrieved for a query, refined this list of entities using statistical and linguistic methods. One of the key components of their method consists of finding hyponyms of the category name specified in the narrative, representing candidate entities and hyponyms as vectors of grammatical dependency triples, and calculating similarity between them.

Table 5.6: *Summary of approaches for TREC-2010 REF Task*

participants that took part in the Expert Finding task at TREC implemented (variations on) one of these two approaches. However, there are few approaches that cannot be categorized into any of these categories. For example, Macdonald and Ounis [209] propose to rank experts with respect to a topic based on data fusion techniques, without using collection-specific heuristics; they find that applying field-based weighting models improves the ranking of candidates. Macdonald, Hannah, and Ounis [211] integrate additional evidence by identifying home pages of candidate experts and clustering relevant documents. Rode, Serdyukov, Hiemstra, and Zaragoza [297] represent documents, candidates, and associations between them as an entity containment graph, and propose relevance propagation models on this graph for ranking experts. For other models and techniques, we refer the reader to numerous variations proposed during the TREC Enterprise track (see [19, 321].).

5.5.4   Entity-level Opinion Detection

We have already discussed related work for Entity-Level opinion detection while discussing challenges for opinion mining in chapter 4 in section 4.4.4.

5.6   Challenges for Entity Retrieval

Entity Retrieval is all about identifying entities, their types and relations and ranking them according to their relevancy with the input query.

5.6.1   Identifying Potential Entities

The most important step in the task of Entity Ranking is to identify valid set of entities and then filter them to obtain only those relevant to the given query. However, identifying different types of the entities (e.g., person, organization products) is also to be considered.

Generally, Named Entity Taggers are used to tag the entities in a given text. There has been a lot of work aimed at developing accurate named entity taggers [66, 104, 207, 352, 401, 405]. Most of them perform very well but their accuracies drop considerably when used in different domains because they are designed to perform well over a particular collection or type of document. The reason for this is that many NE taggers systems rely heavily on complex linguistic resources, which are typically hand coded, for example regular expressions, grammars, and gazetteers etc. The alternatives for extracting entities are present in the form of ontologies [276] and dictionaries [73]. However, these are also subjected to problem of domain-dependency and lack of richness in their contents.

Another challenge related to entity extraction is multi-lingual extraction of entities [333] which aims at extraction of entities from multi-lingual data collections. Once entities have been identified, filtration of these entities is done with respect to the types of entities (e.g., people, location, oranization) an approach is designed to deal with. Only a robust approach can deal with entities of multiple types and that is another challenge of this field.

5.6.2   Identifying Relations

Identifying relations is one of the primary challenges for entity ranking task especially for the approaches wherein it is the relevant relations that are identified first and later these relations are used to identify the relevant entities [399]. Generally, relation identification is needed when we need to determine the related entities for a given source entity or when we need to determine whether given two entities are related through a specific relation in question or not. Various kinds of approaches have been proposed to in the literature to deal with this problem. Using co-occurrence of query terms and entities is one of the most common approach to find the relations or related entities [77, 208, 281]. Another very effective evidence for finding relations between two entities is to consult Wikipedia categories [341, 356]. Some systems [17, 95, 166, 335] explicitly encode entities and their relations (and general knowledge) in RDF (Resource Description Framework), the W3C recommendation of data model for Semantic Web. They can thus leverage the rich expressiveness of query languages like SPARQL[8] for querying entities. According to Li et al. [191], their proposed entity-relationship query could also be used to identify entities and their relationships.

5.6.3   Ranking Entities

Ranking entities is one of the most interesting problems of the field of entity retrieval. Different approaches adopt different methods and techniques to rank entities according to their relevance with the given topic or an example input entity. Most of the approaches look for occurrences of the entities in documents or similarity of the entity page (a page relevant to the given entity) with the given topic [268] for entity ranking task. There are many groups who have also exploited the popular knowledge resource Wikipedia for this task [61, 86, 268]. Other approaches proposed in this regard include the use of topic knowledge [354], Vector Space Modeling [87] or co-occurence statistics [46]. Despite the presence of some very effective entity ranking techniques, it still remains a challenge for researchers working in this domain of research.

---

[8]http://www.w3.org/tr/rdf-sparql-query

5.7 Chapter Summary

In this chapter, we gave a general overview of the field of entity ranking. Besides its definition and motivation, we describe the tasks related to entity ranking. Later on, we provide the details of TREC and INEX entity ranking tracks. We also highlight few challenges that researchers are facing while working in this field. At the end, we discuss some major related work related to this field. With this chapter, we end our part of introduction and next chapter start second part of our contributions for this thesis.

Part 2

Contributions

Section I

Opinion-Topic Association

Chapter 6

Sentence-Level Opinion Detection in Blogs

Errors of opinion may be tolerated
where reason is left free to combat it.
Thomas Jefferson

6.1   Introduction

In this chapter, we propose our opinion finding approach which uses semantic relations of WordNet [230] to improve opinion finding results by focusing on sentence-level opinion-topic associations (OTA) within documents. We propose a novel method of two dimensional query expansion with the purpose of adding relevant and opinionated terms to the query. Experimentation results show a significant opinion finding (O.F.) MAP improvement of almost 29% over topic-relevance baseline.

This chapter is organized as follows: Section 6.2 gives few words about the motivation for this work. In section 6.3, we describe our approach in detail. Section 6.4.1 briefly analyzes different proposed features by comparing thier probability distributions between opinionated and non-opinionated documents. In section 6.4.2, we describe our experiments with analysis of their results. At the end, we report some limitations of our work (see section 6.5).

6.2   Motivation

The basic purpose of the opinion finding task is to retrieve documents that are not only relevant to a given topic but are also opinionated about that topic. Generally, relevancy of a document is computed by using an IR model (as discussed in chapter 3) where each document is given a relevance score according to its relevance to a given topic. However, the computation of opinion score of a document is required to deal with many challenges. One of these challenges is to find opinion-topic associations (OTA) within documents [337]. Any opinion finding approach lacking to deal with this problem is subjected to show poor performance [174]. The basic idea of finding OTA is to identify the textual segments that contain opinions about the given topic. Various approaches have been proposed to identify such textual segments (see chapter 4), however basic idea behind our approach for identification of such textual segments is to check the presence of relevant and opinionated terms in textual segments and score them with respect to associations found between both type of terms.

A document, in general, is a collection of many textual segments (ignoring multimedia content for simplicity) called *paragraphs* that are in turn composed of many *sentences*. If a document contains Z number of paragraphs then it does not mean that all of them discuss the same topic [337]. Similarly, it is not mandatory that, among a set of segments (paragraphs or sentences) relevant to a given topic (Y where $Y < Z$), all of them are opinionated about the topic. For example, a blogpost discussing the topic of *Afghan War* may also contain some sections about *basic teachings of Islam*, *Iraq war*, *terrorism*, *lack of jobs and justice* in third world countries and all of them might not be opinionated. Therefore if a topic-relevance or opinion finding approach considers this blogpost as a single monolithic document then the non-relevant or non-opinionated portions of this document will affect the overall ranking of the document in topic-relevance and opinion run. On the other hand, if we filter the documents to have a potential set of textual segments (like sentences or passages, etc.), it might give us a more accurate ranking of the documents.

Above discussion suggests that documents should better be processed on smaller levels like sentence or passage level for finding better opinion-topic association. The purpose of the work in this chapter is to present our opinion finding approach which exploits WordNet semantic relationships [267] to estimate

sentence-level opinion-topic associations within documents. The documents with more relevant opinionated sentences are given higher OTA score. Besides this, our approach also exploits some document level features based on basic and simple heuristics. Previous opinion finding approaches suggest that a document with higher numbers of adjectives and adverbs is likely to be more opinionated than a document with less number of these parts of speech [62]. Similarly, Yang et al. [173] have suggested that a document containing large number of first-person subject and object pronouns will prove to be more subjective than a document with less number of these pronouns. As described in chapter 4 that subjectivity expresses reality from an individuals point of view, and a natural way to express one's own point of view when writing is to use a first-person perspective. We also included second-person subject and object pronouns, since our hunch was that they too would be instrumental in teasing out subjective texts. Our approach for opinion detection takes into account all of these heuristics by proposing few document level features (explained below in detail).

6.3   Opinion Finding Features

Our opinion finding approach presented in this chapter is a combination of five document-level features and a sentence-level feature. These heuristics features have already been used by few approaches but the way these features have been formulated is different. The only sentence-level feature (i.e., opinion-topic association (OTA) feature) exploits the semantic relations of WordNet [230] to find sentence-level opinion-topic associations in all sentences of a document. Its computation involves expansion of the given query with relevant and opinionated terms. Below, we discuss each of the opinion finding feature used in our approach in detail.

6.3.1   Document Subjectivity

Generally it is assumed that more a document contains subjective terms, more are the chances for it to be an opinionative document [144]. To benefit from this relation, we used the lexicon *SentiWordNet* [100] to compute subjectivity of terms in a document which eventually leads to the calculation of subjectivity of that document.

As already described in chapter 4, SWN assigns three numerical scores ($Obj(s)$, $Pos(s)$, $Neg(s)$) to each synset s of the WordNet describing how objective, positive or negative the terms within a synset are. The range of three scores lies in interval $[0, 1]$ and sum of all the scores equals to 1. It is also very important to note that a term can belong to multiple synsets of SWN and might have different subjectivity values in different synsets. The total number of synsets a term appears in represents the total number of senses for that term. For example, the term *burn* has total 15 senses with positive and negative score of 0.0 in synset burn#v#12 sunburn#v#1 while a positive score of 0.0 and negative score of 0.75 in synset bite#v#2 burn#v#4 sting#v#1. Therefore while looking for a term's subjectivity score in SWN, it is better to use average subjective scores of the terms if we are not using any sense disambiguation method as is the situation in our case. We calculate the average subjectivity score of a term by adding positive and negative scores for all the senses of that term and then divide the total score by total number of term's senses (see equation 6.1). Finally we calculate the average document subjectivity score $Subj(d)$ (in equation 6.2) by summing up subjectivity scores of the terms present in the document. It is very important to mention it that we just used verbs, adjectives and adverbs of a document for calculating the subjectivity of the document.

$$Subj(w) = \frac{\sum_{s_i \in senses(w)} (Neg(s_i) + Pos(s_i))}{|senses(w)|} \quad (6.1)$$

$$Subj(d) = \frac{\sum_{w_i \in d} Subj(w_i)}{|d|} \quad (6.2)$$

$Subj(w_i)$ is the average subjectivity score of a document term w_i in SWN as computed by equation 6.1, $|d|$ is the total number of words in document d and $|senses(w)|$ is the set of word senses found in SWN. In equation 6.1, s_i is the i-th sense of the term w and belongs to set of all senses $senses(w)$ for document term w.

6.3.2   Document Emotiveness Component

According to Zhou et al. [406], deceptive messages are used to be more expressive and expressiveness can be measured with the help of a measure called *emotiveness* which is actually the ratio of adjectives plus adverbs to verbs and nouns. Assuming expressiveness an important clue of opinionativeness of a doc-

ument, we compute $Emot(d)$ as the emotiveness of a document d as given in equation 6.3 where $|X|$ is the total number of X in the document d.

$$Emot(d) = \frac{|\{w \in d | w \in Adjectives\}| + |\{w \in d | w \in Adverbs\}|}{|\{w \in d | w \in Verbs\}| + |\{w \in d | w \in Nouns\}|} \qquad (6.3)$$

6.3.3 Document Reflexivity

People make a lot of use of pronouns like *I, Me, Myself, We, Ourselves* etc. while expressing their opinions. For example, use of *I* in *I think, as far as I am concerned* etc. An example of a real world opinion, posted as a comment on a famous blog[1], is given below to demonstrate the use of these pronouns while expressing opinions:

*He's a very charmig man, **I** have no doubt about it, but **I** haven't heard no opinion from him about Glass-Steigel and Nafta. Also, **I** dont know if this is the right moment to talk about those issues, he seems to be more willing to help Obama than defending his administration. Is there any footage about this? **I**'ve tryed google but haven't found nothing worthy.*

Pronouns such as *I, Me, Myself, We, Ourselves, etc.* give a sense of subjectiveness to the words around them; therefore, we consider them an important clue of opinionatedness of a document. We prepare a list of such pronouns and name it as R. We represent document reflexivity feature as $Refl(d)$ which is computed as given in equation 6.4.

$$Refl(d) = \frac{1}{|d|} |\{w_i \in d \cap R\}| \qquad (6.4)$$

where $|\{w_i \in d \cap R\}|$ is the number of occurrences of reflexive pronouns w_i found in document d and $|d| =$ Total number of words in the document d.

This heuristic has been used by few approaches in the literature [62, 172, 173, 387] but with different formulations. The idea is that any document with larger number of such words will be more opinionative relative to the one with less number of such words.

---

[1]`http://www.huffingtonpost.com/`

6.3.4   Document Addressability

The comment section of a blogpost is where the discussions between readers of the post and authors of the post are held. This discussion results to the generation of opinionated content. Readers write their comments in the comment section and sometimes they address other commentators using pronouns like *You, Yours,* etc., while writing their comments or cite the comment of others in their comment; hence creating an environment of discussion. We make use of these addressive pronouns to estimate the opinionatedness of a document. This idea have been exploited in the past by others too [62, 172, 173, 387]. We prepare a list of such pronouns and label it as A. The addressability feature of a document d is represented as $Addr(d))$ and is given below:

$$Addr(d) = \frac{1}{|d|} |\{w_i \in d \cap A\}| \tag{6.5}$$

where $|\{w_i \in d \cap A\}|$ is the number of occurrences of addressive terms w_i found in document d and $|d|=$ Total number of words in the document d.

6.3.5   Common Opinion Phrases

This component looks for opinion expressions in a given document. The basic idea is that if a document contains many opinion expressions then it is more opinionated than another which contains less number of opinion expressions. For this purpose, we have prepared a list of 100 opinion expressions (called as list P) for English Language with the help of many online blogs[2]. This list contains expressions like *What the hell it is, oh my god, it is thought that, that is not entirely true,* etc. This feature is computed by searching and counting the occurrences of the elements of this list in the given document. The mathematical expression for $Phrs(d)$ is given in equation 6.6:

$$Phrs(d) = \frac{1}{|d|} |\{w_i \in d \cap P\}| \tag{6.6}$$

where $|\{w_i \in d \cap P\}|$ is the number of occurrences of common phrase term w_i found in document d and $|d|$ is the total number of words in document d.

---

[2]http://www.huffingtonpost.com/,   http://www.youtube.com/,   http://www.thedailybeast.com/

6.3.6 Opinion-Topic Association (OTA)

Motivation behind the proposal of this feature has already been described in section 6.2. Each sentence of a given document is assigned an OTA score by measuring the opinion-topic associations within the sentence. A sentence is assumed to contain this opinion-topic association if it contains a query term (or its related terms) surrounded by few opinion terms (like *good, bad, beautiful,* etc.) and is assigned a higher OTA score relative to a sentence with no or less number of related or opinion terms within it. The OTA score of a sentence is computed by matching the given sentence and expanded query semantically. This semantic matching includes use of two semantic measures (i.e., Path and Lesk measures) [267]. Path measure uses the *is-a* relationship of WordNet while Lesk measure uses the gloss definitions of the WordNet for matching. Gloss definitions have been used in the previous work but their role is limited to prediction of semantic orientations of opinionated terms [99, 100, 312].

Before going into details, we summarize the process we follow to calculate the OTA score of a document d:

◇ Sentence boundaries in each document of the collection are identified to split the sentences[3].

◇ The given query is expanded twice. In first phase of query expansion, query is expanded with relevant terms with the help of Wikipedia and a search engine. In second phase of query expansion, query is expanded with opinion terms using TREC provided qrels,

◇ We extract a list of all compound words (like *red hot, eye-popping, etc.*) from WordNet. This list is used to search and mark compound words in expanded query and the set of relevant documents for this query,

◇ Disambiguation of the terms in the given query and a sentence (extracted from a relevant document) is done using Lesk measure,

◇ Stop words are removed from the sentence,

◇ Nouns (of the given query and a sentence) are matched using Path measure while verbs, adverbs and adjectives are matched using Lesk measure,

---

[3]Using the Sentence Splitter developed by Manchester University, UK and is available at `http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector`

⋄ OTA score of the current sentence is calculated on behalf of Path and Lesk matching between query and sentence,

⋄ OTA scores of all sentences in a document are calculated,

⋄ OTA scores of all sentences in a document are sum up to calculate the OTA score of the document.

This process is shown in figure 6.1 and the details of the process of OTA component are given below:
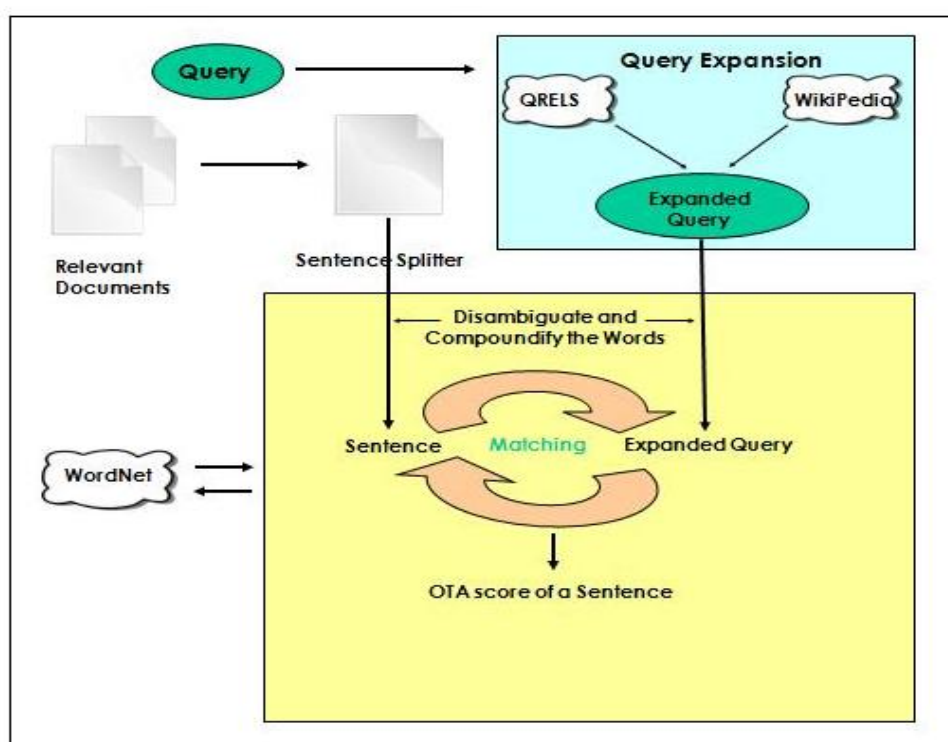
Figure 6.1: *Basic working of the OTA component*

Query Expansion

We propose a novel method of query expansion in which the original query is populated with two kinds of additional terms. First kind of terms to be added are relevant terms and second type of terms added are opinionated terms. The reason behind this bi-dimensional (i.e., relevant dimension and opinionated dimension) query expansion is to assign higher scores to sentences that contain

both relevant and opinionated terms. Below we describe the way we expand a query.

⬦ **Query Expansion with Relevant Terms:** In our two phase query expansion method, we use the query as a base to enrich it with relevant and opinionated terms. In first phase, we use Wikipedia for expanding the query with relevant terms. In Wikipedia based query expansion, we extract a list of *proper nouns and named entities* (often found as hyperlinked text within a Wikipedia document) from the Wikipage corresponding to the given topic (retrieved using the query)[4]. Later on, we manually filter this list of proper nouns and entities for choosing only most relevant entities. At the end of this phase of query expansion, we have a list of relevant terms[5] including the terms of original query.

⬦ **Query Expansion with Opinion Terms:** Relevance assessment results are used for second phase of query expansion by using opinionated documents (i.e. the documents labeled as 2, 3 or 4 in results). We consider it a particular case of relevance feedback[6] in which the user identifies the documents that satisfy his/her information need. We limit the number of chosen relevant opinionated documents to ten and label this small collection of ten documents as $O(Q)$, where Q represents the given query. First of all we remove stop words from collection $O(Q)$ and then a list of *verbs, adjectives and adverbs* is extracted from these documents. Once we have this list of verbs, adjectives and adverbs prepared, we remove duplicates, manually filter for very common terms like (know, do, live etc). Then we compute *document frequency (df)* and *collection frequency (cf)* of all terms in $O(Q)$. Later on, we rank all of these terms according to the ranking function given in equation 6.7 and choose top ten terms to be part of the final query with the terms already extracted from Wikipedia in first phase of query expansion.

$$Subj(t) = df(t) \times cf(t) \tag{6.7}$$

---

[4]If a corresponding Wikipedia page is not found then we use a popular search engine to search for a set of related web documents (using title of the topic) and a list of related proper nouns and named entities is prepared from the snippets of the top ten relevant documents to expand the query.

[5]This list is later on used for selection of relevant passages in the phase *Relevant Passage Selection* from the top $1,000$ retrieved documents

[6]The top ten opinionated documents chosen for relevance feedback are excluded from results while evaluation of the system to avoid biased results

and

$$cf(t) = \frac{|\{t|t \in O\}|}{|T_O|} \tag{6.8}$$

$$df(t) = \frac{|\{d|t \in d\}|}{|D|} \tag{6.9}$$

where $|\{t|t \in O\}|$ is the count of term t in collection $O(Q)$, $|T_O|$ is the total number of terms in the collection $O(Q)$, $|\{t|t \in d\}|$ is the number of documents in which term t appears and $|D|$ is the total number of documents in the collection $O(Q)$.

Sentence-Query Semantic Matching

Once the query has been expanded with relevant and opinionated terms, we use two similarity measures [267], Path and Lesk, of lexicon WordNet [230] to find sentence-level opinion-topic associations. We use Path measure to match nouns of both query and the given sentence while Lesk measure of used to perform matching between verbs, adverbs and adjectives. Below, we explain both measures in detail.

⋄ Path measure is formulated to compute semantic relatedness of word senses by counting nodes in the verb and noun *is-a* hierarchies of WordNet. For example, the path between the concepts *shrub#n#1* and *tree#n#1* is *shrub#n#1* - *woody_plant#n#1* - *tree#n#1*. Hence, only one node exists between concepts *shrub#n#1* and *tree#n#1* which indicates that both concepts are closely related. Since a longer path length indicates less relatedness, the relatedness value returned is the multiplicative inverse of the path length (distance) between the two concepts (see equation 6.10).

$$Relatedness = \frac{1}{distance} \tag{6.10}$$

If the two concepts are identical (e.g., car#n#1 and auto#n#1 are identical and both belong to the same synset), then the distance between them is one; therefore, their relatedness is also 1.

⋄ Lesk measure [189] finds the overlap between glosses of the words being compared as well as words directly linked to them. The major objective of Lesk measure is to count the number of words that are shared between two glosses. The more overlapping the words, the more related the senses

are. For example, if we want to find a matching between words *pine* and *cone* then according to the *Oxford Advanced Learner's* Dictionary, the word *pine* has two senses:

○ sense 1: kind of evergreen tree with needleshaped leaves,

○ sense 2: waste away through sorrow or illness.

The word *cone* has three senses:

○ sense 1: solid body which narrows to a point,

○ sense 2: something of this shape whether solid or hollow,

○ sense 3: fruit of a certain evergreen tree.

By comparing each of the two gloss senses of the word *pine* with each of the three senses of the word *cone*, it is found that the words *evergreen tree* occurs in one sense in each of the two words. So these two senses are then declared to be the semantically most similar when the words *pine* and *cone* are matched together. Similar is the situation when Lesk measure is used for sense disambiguation of words.

The secret behind using different measures for matching different types of terms (i.e., Path measure for nouns and Lesk measure for verbs, adverbs and adjectives) lies in the nature and scope of these measures. Path measure is more precise because it uses the hierarchical relations of WordNet. We use it for matching nouns (i.e., relevant terms) of a sentence and query because we want to have more precise relevancy matches between sentence words and query terms to assign higher score to more relevant sentences. For rest of the terms (i.e., verbs, adverbs, and adjectives), we use Lesk measure to accommodate all possible semantically related words [278]. Besides matching of a sentence and the given query, we use Lesk measure to resolve the words' contextual sense ambiguities [189].

Below, we formulate the process of semantic matching between a sentence and the given query. Formulation of the Path measure for a sentence S_j is shown in equation 6.11.

$$P(SNoun_{ij}) = \frac{\sum_{a=1}^{|N_{Q'}|} Path(SNoun_{ij}, Q'Noun_a)}{|N_{Q'}|} \qquad (6.11)$$

where $Q'Noun_a$ represents the a-th noun of the expanded query while $Path(SNoun_{ij}, Q'Noun_a)$ represents the path similarity score between noun $SNoun_i$ of j-th sentence S_j of a document d and query noun $Q'Noun_a$. Path similarity score for noun $SNoun_{ij}$ is finally normalized by total number of nouns ($|N_{Q'}|$) in the expanded query Q'. Equation 6.12 computes the Path score $PathNn(S_j)$ for the sentence S_j

$$PathNn(S_j) = \frac{\sum_{i=1}^{|N_{S_j}|} P(SNoun_{ij})}{|N_{S_j}|} \tag{6.12}$$

where $|N_{S_j}|$ is the total number of nouns in the sentence S_j while $Path(SNoun_{ij})$ is shown in equation 6.11.

Just like the Path measure, we describe the formulation of Lesk measure. Equation 6.13 shows the formulation of Lesk measure for an adjective of a sentence S_j.

$$L(SAdj_{ij}) = \frac{\sum_{a=1}^{|Adj_{Q'}|} Lesk(SAdj_{ij}, Q'Adj_a)}{|Adj_{Q'}|} \tag{6.13}$$

where $Q'Adj_a$ represents the a-th adjective of the expanded query Q' while $Lesk(SAdj_{ij}, Q'Adj_a)$ represents the Lesk similarity score between word $SAdj_i$ of j-th sentence of a document d and query adjective $Q'Adj_a$. Lesk similarity score for word $SAdj_{ij}$ is finally normalized by total number of adjectives ($|Adj_{Q'}|$) in the expanded query.

Similarly, equations 6.14 and 6.15 represent formulation of Lesk measure for adverbs and verbs respectively.

$$L(SAdv_{ij}) = \frac{\sum_{a=1}^{|Adv_{Q'}|} Lesk(SAdv_{ij}, Q'Adv_a)}{|Adv_{Q'}|} \tag{6.14}$$

$$L(SVb_{ij}) = \frac{\sum_{a=1}^{|Vrb_{Q'}|} Lesk(SVb_{ij}, Q'Vb_a)}{|Vrb_{Q'}|} \tag{6.15}$$

Once Lesk scores for adjectives, adverbs and verbs has been computed, we combine these scores to compute the Lesk score of a sentence:

$$LeskAdj(S_j) = \frac{\sum_{i=1}^{|ADJ|} L(SAdj_{ij})}{|Adj_{S_j}|} \tag{6.16}$$

where $\left| Adj_{S_j} \right|$ is the total number of adjectives in the sentence S_j while $Lesk(SAdj_{ij})$ is shown in equation 6.13.

$$LeskAdv(S_j) = \frac{\sum_{i=1}^{|ADV|} L(SAdv_{ij})}{|ADV|} \tag{6.17}$$

where $|ADV|$ is the total number of adverbs in the sentence S_j while $Lesk(SAdv_{ij})$ is shown in equation 6.14.

$$LeskVb(S_j) = \frac{\sum_{i=1}^{|VB|} L(SVb_{ij})}{|VB|} \tag{6.18}$$

where $|VB|$ is the total number of verbs in the sentence S_j while $Lesk(SVb_{ij})$ is shown in equation 6.15.

Final Lesk score of a sentence S_j is computed by combining Lesk scores of adjectives, adverbs and verbs as shown in equation 6.19:

$$Lesk(S_j) = LeskAdj(S_j) + LeskAdv(S_j) + LeskVb(S_j) \tag{6.19}$$

Finally, OTA score $OTA(S_j)$ for the sentence S_j is computed as shown in equation 6.20:

$$OTA(S_j) = LeskAdj(S_j) + LeskAdv(S_j) + LeskVb(S_j) + PathNn(S_j) \tag{6.20}$$

We represent the OTA score of a document d as $OTA(d)$ which is computed by summing up the OTA scores of all sentences within that document i.e.

$$OTA(d) = \frac{\sum_{j=1}^{N} OTA(S_j)}{|N|} \tag{6.21}$$

where $|N|$ is the total number of sentences in the document d and $OTA(S_j)$ is the OTA score of the the sentence S_j.

Our opinion finding approach used a parameter-free approach to combines the score of all six features to get the final opinion score for a document. The scores of all individual features are further normalized (by their maximum values) to bring all of them in the range of [0..1]. Final opinion score of a document d is represented as $Opin(d)$ and is given in equation 6.22). The final opinion score $opin(d)$ is normalized (by dividing it by 6 which is the number of total features)

to bring its value in the range $[0..1]$.

$$opin(d) = Subj(d) + Emot(d) + Refl(d) + Addr(d)$$
$$+Phrs(d) + OTA(d) \qquad (6.22)$$

At the end of the opinion finding process, all relevant documents input from the previous stage are re-ranked by their final scores (i.e. $final(d)$) which is achieved by combining relevance score $rel(d)$ and opinion score $opin(d)$. Relevant score of a document $rel(d)$ is normalized by the highest relevant score in the topic-relevance baseline.

$$final(d) = opin(d) + rel(d) \qquad (6.23)$$

This combination (equation 6.23) is a parameter-free combination. The purpose of parameter-free combination of relevance and opinion scores is to keep it generalized for different topic-relevance baselines or data collections. However, we are aware that the combination of equation 6.23 needs more effective technique. We deal with this problem in chapter 8 where we combine opinion and relevance scores using machine learning technique.

6.4 Experimentation

We used TREC Blog 2006 data collection [212] for evaluation of our approach with 50 topics of year 2006. The details of the data collection are given in chapter 3.

For pre-processing of the data collection, we remove unnecessary HTML tags like *Script, Style and Image* etc., from the data collection. We also remove the hyperlinks present in a document because most of the noisy data (like calendars, ads, etc.) lies in the form of links in a web document. Even there is a possibility that we can lose valuable data too but loss of valuable data is much lesser than the amount of noisy data we will get rid of.

We perform experimentation in two phases. In the first phase, we look at the probability distributions of different features in opinionated and non-opinionated documents. We perform evaluation of our approach in second phase of experimentation.

6.4.1   Feature Analysis

In this section, we analyse the effectiveness of proposed features by looking
at their probability distributions over topics of TREC Blog 2006 data collec-
tion. The probability distribution comparison between opinionated and non-
opinionated documents shows that features proposed can be helpful to distin-
guish between these genres of documents. In the probability distribution figures
of features, the topics are represented on x-axis while y-axis shows the value
the corresponding feature averaged over all opinionated and non-opinionated
documents for that topic.

Even the probability distribution curve of *subjectivity* (figure 6.2) feature in
opinionated documents does not distance itself from its counterpart in non-
opinionated documents, but difference is sufficient enough to estimate the role
of this feature in identification of opinionated documents.
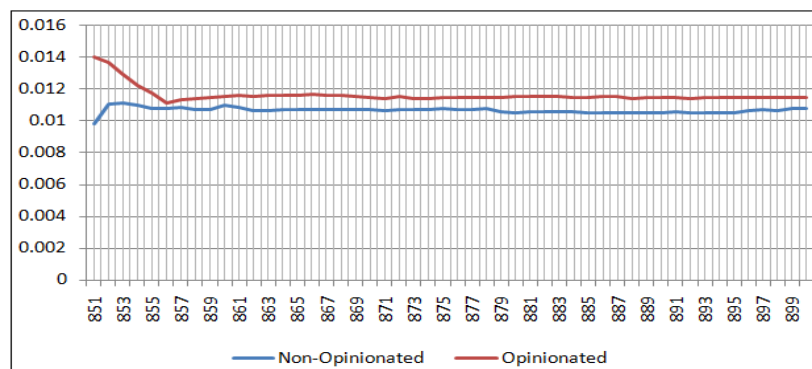
Figure 6.2: *Comparison of Probability Distribution for "Subjectivity" Feature between Opin-
ionated and Non-Opinionated Documents*

A careful analysis of these distributions reveals that probability distributions of
features *reflexivity* (see figure 6.3) and *addressibility* (see figure 6.4) have exactly
similar shape of curves in both opinionated and non-opinionated documents. In
addition, the difference in positions of curves for both features in opinion and
non-opinionated documents can be easily followed that shows their capability
to distinguish opinionated documents from non-opinionated documents.

Figures 6.5 and 6.6 show the probability distributions for *emotivity* and *common
phrases* features respectively in both opinionated and non-opinionated docu-
ments. Curves for both features almost lie in same value range and the distance
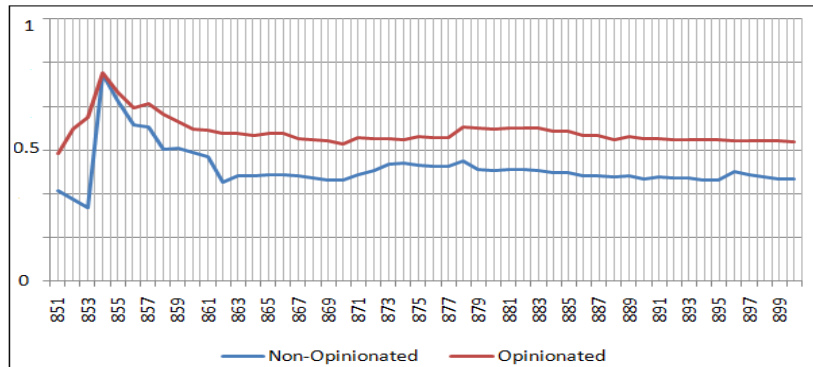
Figure 6.3: *Comparison of Probability Distribution for "Reflexivity" Feature between Opinionated and Non-Opinionated Documents*
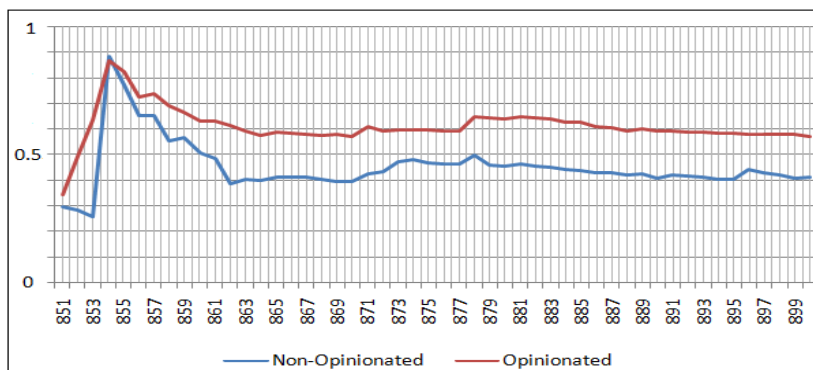
Figure 6.4: *Comparison of Probability Distribution for "Addressibility" Feature between Opinionated and Non-Opinionated Documents*

between curves for opinionated documents and non-opinionated documents for both features show that both of these features are good indicators of opinionatedness.

Similarly, probability distribution curve for OTA feature tells the same story as for rest of the features. OTA curves in opinionated and non-opinionated documents place themselves at enough distance from each other to prove the effectiveness of OTA feature for opinion finding task.

One strange observation that is shared by all probability distributions shown above is a sudden fall or peak appearing for topic number 854. The cause of this fall or peak is the large average size of documents for this topic which is more than average size of the documents for other topics.
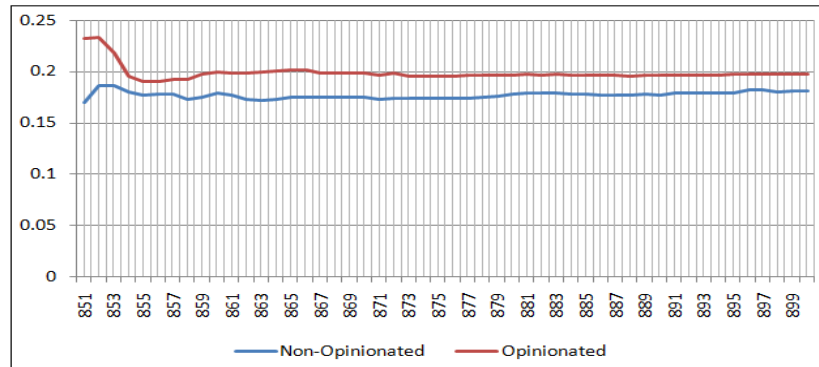
Figure 6.5: *Comparison of Probability Distribution for "Emotivity" Feature between Opinionated and Non-Opinionated Documents*
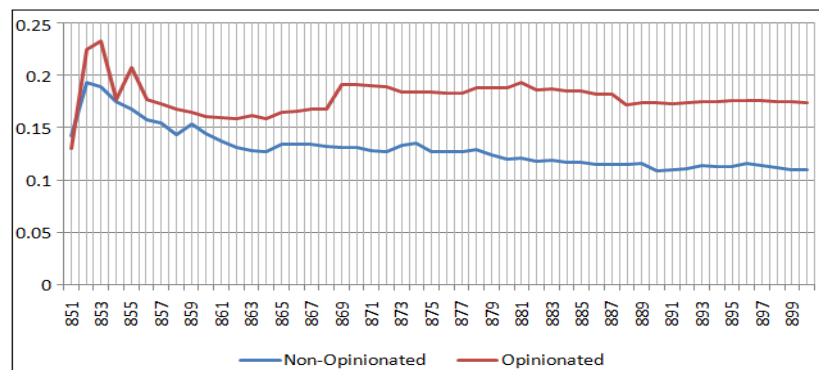
Figure 6.6: *Comparison of Probability Distribution for "Common Phrases" Feature between Opinionated and Non-Opinionated Documents*
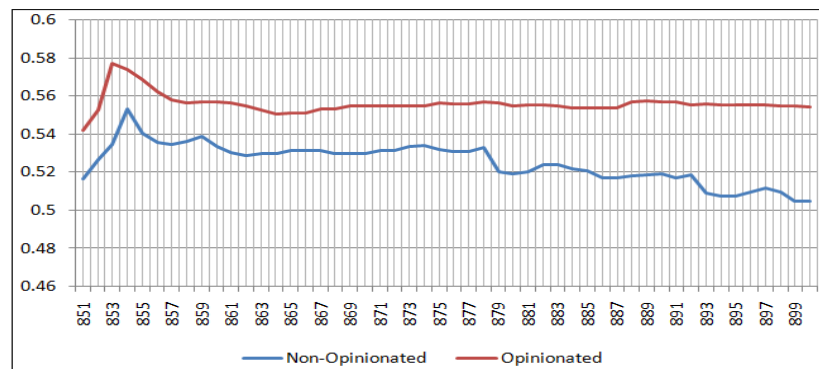
Figure 6.7: *Comparison of Probability Distribution for "OTA" Feature between Opinionated and Non-Opinionated Documents*

6.4.2   Evaluation of our Approach

We use *OKAPI BM25* model [329] for retrieving top *1000* relevant documents for each topic. Each document is given a relevance score represented by $rel(d)$. This baseline has a relevance MAP of *0.2210* over topics of TREC 2006 (i.e., from topic 851 to 900). Table 6.1 shows the details about baseline produced.

	Topic Relevance	Opinion Finding
MAP	0.2210	0.1689
P@10	0.5200	0.3420

Table 6.1: *Baseline MAP and P10*

Experiments[7] were performed in three different setups (see table 6.2) and using three different strategies on same data collection. The details about these experimental setup are given below.

Sentence-Level Setup (SLS)
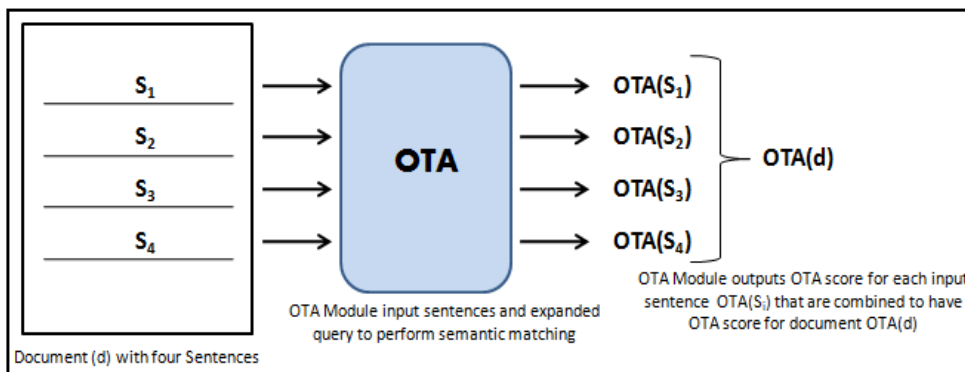
Figure 6.8: *OTA Configuration for Sentence-Level Setup*

The experimentation for first setup was performed as as it has been described in section 6.3. The objective of this setup is to evaluate the effectiveness of proposed features with OTA feature being computed on all sentences of a document. Figure 6.8 describes the way OTA features are computed for this setup. The results for first step are shown in table 6.3.

---

[7]We use $WordNet :: Similarity$ [24, 267] for computation of Path and Lesk measures with default normalization for Lesk measure.

SLS with Selected Sentences
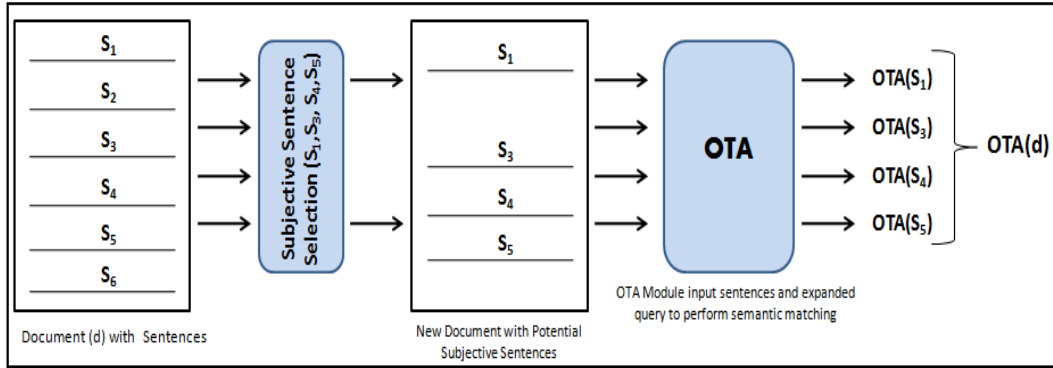
Figure 6.9: *OTA Configuration for Sentence-Level Setup with Selected Sentences*

Few modifications were made in first setup (SLS) to analyze their impact on results. The principle objective is to observe the performance of our approach by computing OTA feature only for potential subjective sentences of a document. We performed the following modifications in SLS:

⋄ At Document-Level Features We removed *emotivity* feature from this setup because of its nominal impact on results of first setup.

⋄ At OTA Level In addition, a sentence selection component was introduced in our opinion finding approach (in OTA component) i.e. only subjective sentences in a document were selected and provided to OTA component for semantic matching with the expanded query. A sentence is considered a subjective sentence if it contains one or more adjectives [139].

SLS with Selected Passages

In third setup of experiments, we selected a set of relevant passages from each document of the collection such that each document is left only with relevant passages. The basic objective for this experimental setup is to analyze the results of our approach by computing OTA feature for sentences of only relevant passages of a document.

⋄ At Document-Level Features we remove *emotivity* feature as we did for second setup and

Setup	Subj(d)	Emot(d)	Refl(d)	Addr(d)	Phrs(d)	OTA(d)
Sentence-Level Setup (SLS)	Yes	Yes	Yes	Yes	Yes	Yes
SLS with Selected Sentences	Yes	No	Yes	Yes	Yes	Yes (with a sentence selection module introduced)
SLS with Selected Passages	Yes	No	Yes	Yes	Yes	Yes (with a passage selection module introduced)

Table 6.2: *Experimental Setup Descriptions*
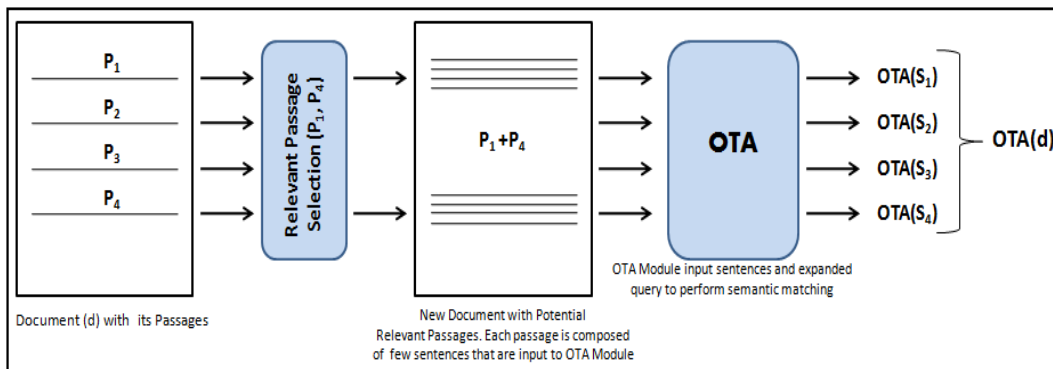
Figure 6.10: *OTA Configuration for Setup-3*

◇ At OTA Level We propose a *Relevant Passage Selection* algorithm for selecting only relevant passages from a document and removing non-relevant passages. The details of this algorithm are given below.

1. **Passage Identification:** There are three ways passages can be identified in documents [56, 168]: Discourse Passage (passages based on document mark-up), Semantic Passages (based on shift of topics within a document) and Window Passages (based on fixed or variable number of words). Regarding the structure of the blog documents in our collection, we decided to identify passages based on their mark-up.

2. **Selecting Relevant Passages:** Deciding criteria for selection of a relevant passage is not an easy task because we have few options like *to select all passages having title of the query in it* or *to select all passages having any query term of the expanded query through*

Wikipedia. Thinking not to miss any relevant passage, we decide to go with second option. Therefore, we choose all such passages in relevant opinionated documents which have at least one occurrence of any of the query term which is part of the expanded query through Wikipedia.

Results

After calculating scores for each individual component, we add document relevance score $rel(d)$ and document opinion score $opin(d)$ score to have final score of the document d. Finally documents are re-ranked using their final scores $final(d)$ 6.23.

Table 6.3 shows the opinion finding (OF) *MAP (Mean Average Precision)* and *P@10* results for our own baseline obtained using Okapi BM25. A significant improvement of almost 29% is observed in opinion finding MAP over baseline results. It is very important to note that we evaluated our approach with residual data collection because our approach involves the process of relevance feedback. A comparison of our approach's opinion finding MAP results with other approaches is not as such possible because of the use of different topic relevance baselines and according to Macdonald et al. [53], the performance of the opinion finding approach is very much dependent on the topic-relevance baseline being used. However, if we compare the percentage improvement of O.F. MAP over baseline with previous work [52, 260] then our approach performed better.

Opinion Run	MAP	P10	Improvement
Baseline	0.1689	0.3420	-
Sentence-Level Setup (SLS)	0.2177*	0.5120*	28.89%
SLS with Selected Sentences	0.2198*	0.5127*	30.13%
SLS with Selected Passages	**0.2243***	**0.5200***	**32.26%**

Table 6.3: *O.F. MAP and P@10 for three Experimental Setups. An asterisk (*) shows the significant improvement over baseline.*

Table 6.3 lists the results for our three setups. Surprisingly the results for second setup did not make a big difference as was expected. The major cause of such results may be less than expected performance of POS Tagger which might have not performed well because of absence of proper punctuations, good use of grammar rules and capitalization etc. within the sentences.

Results for third setup are the best results that highlight the importance of passages for the opinion finding in blogs. It can also be noticed that we got some improvements in second and third setup over first setup but these are very marginal improvements. However these results encourage us to further explore the role of passages for opinion finding in blogs.

We believe that the results given by our approach can further be improved if the use of Path and Lesk measures could be optimized. Path measure seems to be more precise because it uses the *is-a* relationships of WordNet. On the other hand, Lesk measure compares the gloss definitions of terms for matching. The matching of gloss definitions is a bit unreliable because of ambuiguity of WordNet glosses [243]. The use of eXtended WordNet (XWN)[8] can be helpful in this regard.

6.5   Limitations of our Approach

Our approach has performed well but it is subjected to many limitations and we discuss major ones in this section. However, we have worked to eliminate some of these limitations in our work presented in next chapters.

⋄ The use of semantic relations for opinion-topic association is interesting and useful but becomes less effective when dealing with a data collection of gigantic size. Matching each noun, verb, adjective, and adverb of a sentence with corresponding types of terms using Path and Lesk measure takes too long and if this process is to be repeated with thousands of sentences then the task becomes a bit impractical. Sufficient amount of memory and processing power is needed. In addition, Lesk measure depends on the gloss definitions of concepts present in WordNet for matching that are sometimes too short to well describe a concept that creates doubts about the reliability of the matching results. Therefore, there is a need of more reliable and robust IR model which is not only effective but also performs well. In chapter 7, we have exploited the robustness of language models by proposing a passage-based langugae modeling approach for opinion finding task. This language modeling approach also takes support of proximity-based bi-dimensional query expansion technique.

---

[8]http://xwn.hlt.utdallas.edu

⬦ During individual feature analysis, we observed that almost all features were equally capable to distinguish the opinionated documents from non-opinionated documents. Therefore, we used a parameter-free method to combine scores of various features to compute the final opinion score of a document. This approach of combining scores lacks fine-tuning. Using a machine learning approach in this regard would be better approach. Similarly, linear combination of relevance and opinion score also requires a better fusion technique. These limitations are target of our work presented in chapter 8 where we use machine learning algorithm to evaluate each opinion finding feature proposed for that work and to find a good combination of features.

⬦ The manual selection of relevant entities from Wikipedia is another drawback of our approach. It should be replaced by an effective approach for entity retrieval. This limitation has been tackled in chapter 10 while presenting our work for entity retrieval in news articles.

6.6   Chapter Summary

In this chapter, we presented our opinion finding approach which aims at improving opinion finding results by focusing on the problem of finding opinion-topic associations with the help of semantic relations of WordNet(Path and Lesk measures). This approach also takes support of bi-dimensional query expansion. Basic motivation behind this approach is the assumption that a sentence containing relevant and opinionated terms could be more opinionated and hence the document with such sentences.

6.6.1   Findings

⬦ Selecting a subset of subjective sentences from a document for opinion finding task is useful.

⬦ Passages are another very effective granularity level where the results of opinion finding task could be improved.

6.6.2 What needs to be improved?

Even our approach has given exceptionally good results by giving an improvement of 32% over O.F. MAP of baseline but still there are many issues where we need to improve our work. Major problems with our approach are:

◇ Path and Lesk measures seem to be effective but our work lacks any empirical evidences to support this claim. Another major problem with using these semantic relations is the requirement of performing intensive computations that results in heavy cost of processing time and power.

◇ A balance is required while combining Path and Lesk scores of a sentence. Suppose two sentences s_1 and s_2 where sentence s_1 contains k number of opinionated terms with no topic-relevant terms and sentence s_2 contains l opinion terms with p topic-relevance terms. Using our approach, sentence s_1 could be assigned a higher score if opinion terms of sentence s_1 and given query match closely which should not be the case because these opinionated terms could not be topic-relevant. Our approach should include a mechanism that should rather deal with both type of sentences differently or should have some evidence about relevancy of opinionated terms. In our approach presented in chapter 7, we use proximity-based query expansion technique to deal with this problem.

◇ Our approach linearly combines scores of different features. This combination if done by an effective technique, like using some machine learning algorithm, could be very effective. We have tried to tackle this problem in chapter 8 where we use better formulations of these features and use machine learning techniques to combine them.

◇ Our approach is required to use a standard baseline like the one offered by TREC for TREC Blog 2008 track for a fair comparison with other approaches. Therefore, rest of our work use these standard baselines.

Chapter 7

Passage-Based Opinion Detection in Blogs

Inconsistencies of opinion, arising from changes
of circumstances, are often justifiable.
Daniel Webster

7.1 Introduction

In this chapter, we present our work for opinion detection using a passage-based approach. The limitations of our work presented in chapter 6 drove us to concentrate on more robust IR models for the task of opinion detection. Therefore, in this chapter we propose a *Language Modeling (LM) approach* for opinion detection. The basic idea is to combine the opinion score of terms with the language model to adapt it for opinion retrieval. Furthermore, we propose a method of query expansion with two types of terms (i.e., relevant terms and opinionated terms). This expansion of the given query is little different from the one introduced in chapter 6 because it also involves proximity of relevant and opinionated terms. Experimental results not only prove the effectiveness of our approach by giving a significant improvement over baseline but also show that opinionated information retrieval is less dependent on features based on term frequency which is considered one of the most important feature in topic-based information retrieval.

This chapter starts with the motivation for current work (section 7.2). In section 7.3, we describe our passage-based language modeling approach in detail. In section 7.4, we evaluate our system under TREC Blog track evaluation framework and compare its performance with the best published results for topics of TREC Blog 2006.

7.2   Motivation

Like the work presented in chapter 6, the basic aim of work proposed in this chapter is to focus on the problem of finding opinion-topic associations in documents for the task of opinion detection. The approach [235, 238] presented in chapter 6 exploits the semantic relations of WordNet to find the sentence-level opinion-topic associations within documents. The results of work in chapter 6 suggest that passages are better processing units for finding opinion-topic associations for opinion detection task. In addition to this, it seems more practical to process blog documents on passage level because sentence splitting is a very challenging task especially when we are dealing with blogs. Lack of punctuations, lack of capitalization and grammar mistakes make the task of identifying sentence boundaries more difficult and hence affects the performance of the system. Even if split well, we may lose the context of a sentence while working on sentence-level. And last but not the least, blog documents are more logically structured in the form of passages. A blogpost is split into many passages and normally a comment is contained within the boundaries of a passage. Therefore, it becomes more feasible (and logical) to process blog documents on passage level.

Indeed, passage identification and utilization in information retrieval has been the focus of research for quite some time [1, 55, 56, 167, 205]. Utilization of passages has been shown to be highly beneficial for a variety of information retrieval tasks: classical ad hoc retrieval [55, 56, 167, 168, 205], question answering [76, 146] and query expansion [50], etc. However, we cannot find any work that specifically uses passages for the task of opinion mining. The approaches adopted by [144] and [388] use passages for topic-relevance retrieval. Yang et al. [144] adopts a passage-based procedure for topic-relevance retrieval and sentences for the task of opinion mining. However, the work of Lee et al. [388] has more resemblance to our work. They use passage-based language model for topical-relevance retrieval of documents. For opinion finding task,

they prepare a query-specific lexicon using the best passage extracted using the complete-arbitrary passage approach [248] from the top N relevant documents. But our work adopts a different approach here. In our case, relevant passages are selected from the top opinionated documents instead of relevant documents. This selection strategy gives an upper hand to our approach because relevant passages in an opinionated document are likely to be more opinionated than relevant passages in a relevant document. Hence, the lexicon or list of opinionated words prepared from opinionated passages is more reliable than the others. Another point where our language modeling approach is different from others is its adaptation for opinionated information retrieval and this is done by combining the opinion score of terms in the model (see equations 7.15 to 7.20).

7.3   Passage-based Opinion Finding

In Information Retrieval, passage-based retrieval has played very important role. Our work is heavily influenced by the work conducted in the past for passage-based adhoc retrieval documents [32, 33]. In passage-based ad hoc retrieval, whether we can return the relevant passages as a result [7], or simply mark the entire document as relevant if it contains (some) relevant passage(s) [56, 167, 205]. The focus of our work presented here is on the latter (i.e., on using passage-based methods for retrieving documents). Indeed, the merits of passage-based document retrieval have long been recognized [241, 303]. Perhaps the most prominent one is that using passages rather than whole documents to induce document ranking is more effective for detecting long (or heterogeneous) relevant documents with many parts that contain no query-relevant information, as in the case of examples from above.

In this section, we describe our passage-based language modeling approach for opinion finding. We start with bi-dimensional query expansion which involves the use of Wikipedia as a knowledge resource and relevance feedback technique. Query expansion with relevant terms helps to select the relevant passages of a document that are later on used to build passage-based language model. Experimental details of our work are given in section 7.4.

7.3.1 Query Expansion

We propose a two phase query expansion method in which the title of the given query (topic) is used as a base to populate the query with two types of terms. This method of query expansion is bit different from the one proposed in chapter 6 because in current methods, we also use proximity measure to expand the query which has proved its effectiveness for opinion related tasks [308]. In the first phase, the query is populated with relevant terms (the terms extracted from the Wikipage concerning to the title of the given topic). This includes mostly the hyperlinked proper nouns and named entities. In second phase, the query is expanded with opinionated terms (the terms surrounding the Relevant terms in the documents marked as opinionated during opinion feedback process (discussed below in detail). This includes only verbs, adverbs and adjectives. Examples of both types of terms for *Topic-851* titled as *March of the Penguins* are given below in table 7.1. Figure 7.1 explains the process of query expansion in pictorial form. It is very obvious that expansion of the query with opinion terms does not necessarily require the selection of relevant passages (i.e., can be done at document level) but we have intentionally described it in context of passages to maintain a coherency with passage-based language model to be described later. Below, each step of the query expansion process is described in detail.

Relevant	**Opinionated**
Luc Jacquet	Hilarious
Academy award	Enjoy
Antarctica	Emotional
Bonne Pioche	Delightful
Documentary	Boring
National Geography	Absolutely

Table 7.1: *Example of few Relevant and Opinionated Terms for Topic-851 with title March of the Penguins*

Query Expansion with Relevant Terms

In first phase, we use Wikipedia for expanding the query with relevant terms. In Wikipedia based query expansion, we extract a list of *proper nouns and named entities* (often found as hyperlinked text within a Wikipedia document) from the

Wikipage corresponding to the given topic (retrieved using title of the query)[1]. Later on, we manually filter this list of proper nouns and entities for choosing only most relevant entities. At the end of this phase of query expansion, we have a list of relevant terms[2] (L_{REL}) including the title of the query.

Query Expansion with Opinion Terms

The objective of this phase of query expansion is to enrich the query with important opinionated terms. This objective is achieved in two steps as described below:

1. *Selecting Potential Relevant Passages:* In first step, we use the list L_{REL} and relevance assessments (*qrels*) to remove irrelevant passages from the collection of top $1,000$ documents for a particular query. After this stage, documents are left with only relevant passages. Basically, this stage performs three tasks:

   ◇ Passage Identification: Generally, there are three ways passages can be identified in documents [56, 168]: *discourse passage* (passages based on document mark-up), *semantic passages* (based on shift of topics within a document) and *window passages* (based on fixed or variable number of words). Looking at the structure of the blog documents in our collection, we decided to identify passages based on their mark-up (i.e., using $<P>$ and $<DIV>$ tags).

   ◇ Selection of Potential Relevant Passages: Deciding criteria for selection of a relevant passage is not an easy task because we have got a few options like *to select all passages having only original query terms in it* or *to select all passages containing any one of the term from the list L_{REL} obtained through Wikipedia*. Thinking not to miss any relevant passage, we decide to go with second option. Consequently at the end of this stage, we are left with a document collection with each document containing only relevant passages.

2. *Extracting Opinion Terms:* In the second stage, we extract a list of opinion words (L_{OPIN}) from top 10 documents ($O(Q)$) among the $1,000$ relevant

---

[1]If a corresponding Wikipedia page is not found then we use Google Search Engine to search for a set of related web documents (using title of the topic) and a list of related proper nouns and named entities is prepared from the snippets of the top ten relevant documents to expand the query.

[2]This list is later on used for selection of relevant passages in the phase *Relevant Passage Selection* from the top $1,000$ retrieved documents

documents marked as opinionated (labeled as 2, 3 and 4) in the TREC qrels. We assume it a special case of external feedback in which a user is asked to identify and mark the top 10 opinionated[3] (we call it *Opinionated Feedback*) and top 10 non-opinionated documents ($\bar{O}(Q)$) from the list of relevant documents presented to him (we call it *Non-Opinionated Feedback*). Non-Opinionated Feedback is used in weighting the query terms.

Once $O(Q)$ and $\bar{O}(Q)$ have been identified for a given query Q, we remove the stop-words (articles, conjugations and prepositions, etc.) from all documents in $O(Q)$ to make their further processing easier. Then we mark the occurrences of each relevant term of list L_{REL} within all the relevant passages of these documents. A window of four words (i.e., 2 on the left and 2 on the right) is created around each occurrence of a relevant term and is filtered to include only verbs, adjectives and adverbs. This process is repeated with each opinionated document in $O(Q)$. Finally at the end of this step, we have a list (represented as L_{OPIN}) which contains all opinion words found close to relevant query terms (L_{REL}) within relevant passages of opinionated documents. Duplicates are removed from L_{OPIN} and opinion (or subjectivity) score of each term is calculated using popular lexical resource SentiWordNet (SWN) [100]. In SWN, each synset of the WordNet is assigned a subjectivity score (in the form of positive and negative score) and an objective score such that sum of both equals 1. All terms in the list L_{OPIN} having subjectivity score of zero (or if they do not exist in SWN) are removed from the list. A term may appear in more than one synsets of SWN. Therefore, we compute the subjectivity score ($Subj(t)$) of a term t by summing up its positive and negative scores in all synsets it appears and by normalizing it by total number of synsets it appears in (see equation 7.1).

$$Subj(t) = \frac{\sum_{s_i \in senses(t)} (Neg(s_i) + Pos(s_i))}{|senses(t)|} \qquad (7.1)$$

In equation 7.1, $Neg(s_i)$ is the negative score of the sense s_i of term t as found in SWN, $Pos(s_i)$ is the positive score of the sense s_i of term t as found in SWN and $|senses(t)|$ is the total number of senses for term t in SWN.

---

[3]The top ten opinionated documents ($O(Q)$) chosen for relevance feedback are excluded from results while evaluation of the system to avoid biased results.
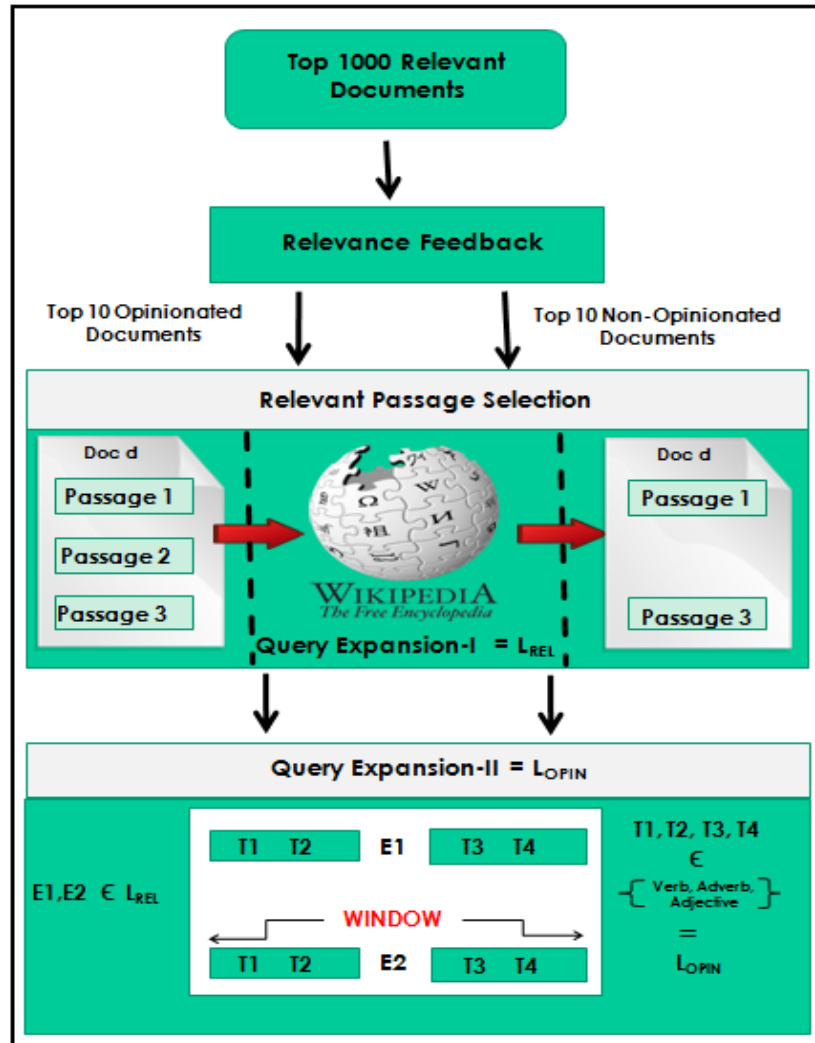
Figure 7.1: *Query Expansion with relevant and opinionated terms*

7.3.2   Term Weighting

During the query expansion process, we have expanded the original query with two types of terms (i.e., relevant (L_{REL}) and opinion terms (L_{OPIN})). The terms in relevant list (L_{REL}) and opinion list (L_{OPIN}) are weighted using two different schemes. For opinion terms, we combine the subjectivity score of the terms with their number of occurrences in the collection $O(Q)$. For relevant terms, only frequency evidence is used for weighting. As described above already that subjectivity is calculated using SWN and the parameters used for frequency

calculation are collection frequency (cf), passage frequency (pf) and document frequency (df). Therefore, we calculate the cf, pf and df for all terms using formulas given in equations 7.2, 7.3, and 7.4.

$$cf(t) = \frac{|\{t|t \in O\}|}{|T_O|} \qquad (7.2)$$

$$df(t) = \frac{|\{d|t \in d\}|}{|D|} \qquad (7.3)$$

$$pf(t) = \frac{|\{g|t \in g\}|}{|P|} \qquad (7.4)$$

In above equations, $|\{t|t \in O\}|$ is the count of term t in collection $O(Q)$, $|T_O|$ is the total number of terms in the collection, $|\{t|t \in d\}|$ is the number of documents in which term t appears and $|D|$ is the total number of documents in the collection $O(Q)$, $|\{g|t \in g\}|$ is the number of passages in which term t appears and $|P|$ is the total number of passages in the collection.

For opinion terms, the opinion score of each term is calculated in two different ways: First, it (labeled ALL in results table 7.3) is calculated using equation 7.5 shown below; second, it (labeled $FREQ$ in results) is calculated as shown in case two of equation 7.5. If analyzed carefully then it can be noted that case-I of equation 7.5 (i.e., ALL) gives equal importance to evidences of frequencies and subjectivity but case-II of equation 7.5 gives more value to subjectivity.

$$Opin_{func}(t) = \begin{cases} if func = ALL \ then \ (cf * pf * df) * Subj(t) \\ if func = FREQ \ then \begin{cases} (cf * pf * df) \ if Subj(t) \geq 0.5 \\ \text{Term is dropped if } Subj(t) < 0.5 \end{cases} \end{cases}$$
$$(7.5)$$

For example, a term with subjectivity value of 0.4 is dropped even if it is one of most frequent terms occurring in the collection. However, we believe that the combination of frequencies of terms with their subjectivity in equation 7.5 will do good to retrieve opinionated documents. To make the selection of more appropriate terms (that would really help to differentiate opinionated documents from non-opinionated documents) possible, we propose to use the top 10 non-opinionated documents $(\bar{O})$ already marked during external feedback process. All terms present in L_{OPIN} are checked for their existence in irrelevant docu-

ments and concerning frequencies (cf, pf and df) are calculated. Opinion scores of the terms for documents in $\bar{O}$ are calculated similarly using equation 7.5 and final score of a term t present in L_{OPIN} is calculated using equation 7.6. It's obvious from equation 7.6 that a term which is not present in non-relevant document(s) but it is present in opinionated document(s) will be assigned a higher final score. So as a result of using equation 7.6, we are giving higher scores to terms which are uniquely present in opinionated documents and not in both (i.e., opinionated and non-opinionated).

$$Score_{Opin}(t) = Opin(t)_{Rel+Opin} - Opin(t)_{NonRel} \qquad (7.6)$$

All terms are ranked using their final opinion score and then top 30 terms are selected to be part of the final query. A term is removed from the list L_{OPIN} if $Opin(t)$ results in a negative value.

Same process is repeated for relevant terms of the list L_{REL} (i.e., relevance score of each term is computed using small document collections $O(Q)$ and $\bar{O}$ as we did for opinion terms). The expression used for assigning weights to relevant terms is given below in equation 7.7. Eventually final scores of relevance terms are computed using equation 7.8.

$$Rel(t) = cf(t) * pf(t) * df(t) \qquad (7.7)$$

In equation 7.7, the value of $Rel(t)$ lies in range $[0..1]$.

$$Score_{Rel}(t) = Rel(t)_{Rel+Opin} - Rel(t)_{NonRel} \qquad (7.8)$$

All relevant terms are ranked according to their final relevance score. A term is given a final relevance score of zero if $Score_{Rel}(t)$ results in a negative value and top 10 relevant terms are selected to be part of the final list of relevant terms. At the end, we merge both lists (i.e., L_{REL} and L_{OPIN}) to form a final list of query terms that contain both relevant and opinionative terms.

7.3.3   Passage-Based Language Model

There are two ways passages have been used for ad hoc retrieval: First, returning the passages as result of the query. Second, returning the documents as a result of the query while attributing a score to the documents on behalf of its passage(s). We will focus on second case in our work.

We have described earlier that we are using language modeling for realization of our passage-based approach for the task of opinion detection. A statistical language model is a probability distribution that captures the statistical regularities of language generation. It determines how likely a given string is in a language, given a model of language generation. Query-likelihood Model [245, 274, 340] is one of the most frequently used language model in research work and in this work too. In the query-likelihood model, we estimate the probability of a query being generated by a probabilistic distribution over a fixed vocabulary induced by a document. For a query q and a document d this generation probability is often denoted $P(q|d)$. The posterior probability $P(d|q)$ [324, 340] is used in order to rank documents, which can be written using Bayes' rule as

$$P(d|q) = \frac{P(q|d)P(d)}{P(q)} \tag{7.9}$$

Since $P(q)$ is not dependent on the document and in lack of prior information $P(d)$ is assumed to be uniformly distributed, the ranking task reduces to estimating $P(q|d)$. For estimating the probability $P(q|d)$, we use Unigram Language Model, which were shown to be quite effective [183, 229, 274]. Unigram language models assume that terms are independent of each other. In our work, we use three passage-based documents scoring functions that are realized using a Unigram Language Model shown as below:

$$Score_{AVG}(d) = \frac{1}{|S|} \sum_{i=1}^{|S|} P(q|g_i) \tag{7.10}$$

$$Score_{MAX}(d) = max_{g_i \in d} P(q|g_i) \tag{7.11}$$

$$Score_{LINEAR}(d) = \sum_{i=1}^{|S|} P(q|g_i) \tag{7.12}$$

Where $Score_{AVG}(d)$ is the average of scores of all passages within a document d for a given query q, $Score_{MAX}(d)$ is the score given to document d for a query q on behalf of one of its passages having maximum score, and $Score_{LINEAR}(d)$ is a linear addition of scores of all passages; $|S|$ is the total number of passages within the document d, g_i is the i-th passage and $P(q|g_i)$ is the probability of

generating query q from passage g_i which can also be written as shown below:

$$P(q|g) = \prod_{t_i \in q} P(t_i|g) \qquad (7.13)$$

Using above equation can lead to a sparse matrix. To avoid this situation, we need to use a kind of smoothing model for better results. We use *Mixture of Language Models (MIX)* for this purpose. In this model, we assume that each word in the query q (actually expanded from passages) is generated from a mixture of three language models: the Collection Model, the Document Model and the Passage Model itself.

$$P(t_i|MIX) = \lambda_1 P(t_i|g) + \lambda_2 P(t_i|d) + \lambda_3 P(t_i|c) \qquad (7.14)$$

Where $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $P(t_i|g)$ is the probability of generating query term t_i from passage g, $P(t_i|d)$ is the probability of generating query term t_i from document d and $P(t_i|c)$ is the probability of generating query term t_i from the whole collection c. All three are given below in equations 7.16, 7.18, and 7.20. This computation of score of a term t_i in a passage $(Scr(t_i, g))$, document $(Scr(t_i, d))$ or collection $(Scr(t_i, c))$ also involves the use of its final score $Score(t_i)$ (equations 7.6 and 7.8) because we want these scores of the term t_i to be function of its score in $O(Q)$. This is how we combine the opinion scores of the terms with the language model.

$$Scr(t_i, g) = \frac{C(t_i, g)}{|T_g|} \times Score(t_i) \qquad (7.15)$$

$$P(t_i|g) = \frac{Scr(t_i, g)}{\sum_{\forall t \in g} Scr(t, g)} \qquad (7.16)$$

$$Scr(t_i, d) = \frac{C(t_i, d)}{|T_d|} \times Score(t_i) \qquad (7.17)$$

$$P(t_i|d) = \frac{Scr(t_i, d)}{\sum_{\forall t \in d} Scr(t, d)} \qquad (7.18)$$

$$Scr(t_i, c) = \frac{C(t_i, c)}{|T_c|} \times Score(t_i) \qquad (7.19)$$

$$P(t_i|c) = \frac{Scr(t_i, c)}{\sum_{\forall t \in c} Scr(t, c)} \qquad (7.20)$$

Where $C(t_i, g)$, $C(t_i, d)$ and $C(t_i, c)$ are the counts of term t_i in passage g, document d and collection c respectively. $|T_g|$, $|T_d|$ and $|T_c|$ are the total number of terms in a passage g, document d and collection c respectively. $Score(t_i)$ is the score of each term t_i computed using equation 7.6 for opinionated terms and equation 7.8 for relevant terms.

At the end of opinion finding phase, each document is assigned an opinion score which is linearly combined with the document relevance score $(rel(d))$. The relevance score $rel(d)$ is already normalized by the highest relevance score in a topic-relevance baseline of a given topic. This addition of opinion score and relevance score of a document d results in final score for the document (i.e., $final(d)$). Finally the documents are re-ranked according to this final score $(final(d))$. The combination of relevance and opinion score we use is a parameter-free combination. The purpose of parameter-free combination of relevance and opinion scores is to keep it generalized for different topic-relevance baselines or data collections. However, we are aware that combining relevance and opinion scores of documents requires more effective technique. We deal with this problem in chapter 8 where we combine opinion and relevance scores using machine learning technique.

7.4   Experimentation

For experimentation purposes, we use TREC Blog 2006 collection [212] with topics of year 2006. We show the effectiveness of our passage-based approach by achieving an improvement in opinion finding MAP over our baseline. It is to be recalled that each query is expanded with sets of relevant (top 10 terms) and opinionated terms (top 30 terms) using knowledge resource Wikipedia and relevance feedback respectively. The motivation behind selecting less number of relevant terms than number of opinionated terms is to give less preference to relevancy because we are already using relevant score of a document while computing final score of a document.

7.4.1   Data Preprocessing

In this phase, we remove unnecessary HTML tags like *Script, Style and Image* etc. We also remove the hyperlinks present in a document because most of the noisy data (like calendars, ads, etc.) lies in the form of links in a web document. Even there is a possibility that we can lose valuable data too but loss of valuable data is much lesser than the amount of noisy data we will get rid of.

7.4.2   Topic Relevance Retrieval

The purpose of *Topic Relevance Retrieval* stage is to retrieve a set of relevant documents for each topic. The topic relevance baselines provided by TREC Blog track not only makes this task easier but also provides an opportunity to better evaluate opinion finding approaches working under the TREC Blog framework. The reason behind this step of providing baselines is too much dependency of performance of opinion finding approaches on topic relevance baselines [53]. By testing the opinion finding approaches over same baselines, it becomes easy to analyze the effectiveness of different opinion finding evidences.

In our case, we retrieve top $1,000$ documents for each query by using the strongest TREC baseline (i.e., baseline-4 during the phase of topic relevance retrieval). This baseline has a opinion finding MAP of 0.3022 over topics of TREC Blog 2006 (i.e., from topic 851 to 900). Rest of the details for baseline-4 are given below in table 7.2.

	Topic Relevance	Opinion Finding
MAP	0.4300	0.3022
P@10	0.7920	0.5240

Table 7.2: *Baseline-4 MAP and P@10 for topics of year 2006 [53]*

7.4.3   Results and Discussions

The documents are ranked by their final score which is the result of the linear addition of their opinionated score and topic retrieval score as mentioned in TREC baseline-4. Experiments were performed using different values of λ_1, λ_2 and λ_3 but best results are obtained with lambda values of $\lambda_1 = 0.5$, $\lambda_2 = 0.3$

and $\lambda_1 = 0.2$. The final results are given below in table 7.3. The metrics used for evaluation are MAP (Mean Average Precision) and P@10 (Precision at top 10 Documents). Table 7.3 shows the results for three different document scoring functions (shown in equations 7.10, 7.11, and 7.12 using two different query term weighting functions *ALL* and *FREQ*.

Ranking Function	Metric	ALL	FREQ
AVG $^{7.10}$	MAP	**0.3303***†	**0.2735**
	P10	**0.6340***	0.4980
MAX $^{7.11}$	MAP	0.3290*†	0.2636
	P10	**0.6340***	0.5280
LINEAR $^{7.12}$	MAP	0.2342	0.2418
	P10	0.5160	**0.5400**

Table 7.3: *O.F. MAP and P@10 for three Ranking Functions. An asterisk (*) shows the significant improvement over baseline and a † indicates the best reported results ever (to the best of our knowledge) [314]*

The results show an improvement of almost 9.29% (0.3022 vs 0.3303) in MAP over baseline results which is the best ever reported MAP over TREC Blog 2006 topics to the best of our knowledge. However, it should be noted that our approach uses relevance feedback which makes it difficult to justify the performance comparisons.

7.4.4   Comparison with Previous Approaches

The previous best reported MAP over topics of TREC Blog 2006 is 0.3221 [314]. We also discussed earlier two works ([144] and [388]) that also adopt a passage-based approach in their work. The results of our approach cannot be compared with [388] because they use topics of TREC Blog 2008 for their work while we used topics of TREC Blog 2006. However, [144] performed experimentation with topics of TREC Blog 2006 and reported an opinion finding MAP of 0.1576 which is much lower than our opinion finding MAP (0.3303). But the comparison of the performance of our work with [144] cannot be justified because of the use different topic relevance baselines and according to [53], the performance of opinion finding approach depends on the strength of the underlined baseline.

As far as *P*@10 results are concerned, unfortunately [314] does not report the *P*@10 results so we cannot compare *P*@10 results with [314]. TREC Blog 2008 overview paper [53] reports some results obtained with baseline-4 but those

results are reported on topics of TREC 2008. But still our *P@10* results (0.6340 using title field only) are comparable to the best *P@10* reported for topics of TREC 2008 (0.6400 using title field only) in [53] especially when [53] states that TREC 2006 topics are the most difficult topics among topics of year 2006, 2007 and 2008 and topics of TREC 2008 are the easiest. However, the *P@10* results in table 7.3 follow the same pattern as MAP.

7.4.5   Effect of Ranking Functions

If we look at the MAP results then it is very clear that the results for ranking functions *AVG and MAX* are far better than the results of *LINEAR (LIN)* ranking function. It should be noted here that both functions (i.e., *AVG and MAX*) are basically representing the score of one passage of a document while *LINEAR (LIN)* ranking function is basically representing all the passages of a document (or the whole document itself) which, in a way, confirms our point that it is not the whole document which can improve the performance of opinion retrieval but it may be relevant portions of the document (passages in this case) that might have opinions about the given topic. Even if we do not consider that *AVG* is being calculated over scores of all passages, the difference between results of *AVG and MAX* is very marginal for both (i.e., *FREQ and ALL*).

7.4.6   Effect of Term Weighting Schemes

We performed experimentation with two different term weighting schemes, *ALL and FREQ* (see equation 7.5). If we compare the results of *ALL and FREQ* then it is evident that *ALL* outperformed the *FREQ*. The reason behind low performance of *FREQ* is quite obvious because there are not a large number of terms in a document having the subjectivity score of over 0.5. There may be few terms that are more subjective in their nature but those are less frequently used. While in case of *ALL*, a better balanced formula is used which combines both (i.e., frequencies and subjectivity) together. In other words, it can be said that *FREQ* is more dependent on the frequency of highly opinionated terms while *ALL* is likely to depend on presence of any opinionated terms. Better performance of *ALL* function also suggests less dependency on term frequency in opinionated information retrieval which is in fact one of the most important feature in topic-based information retrieval. This was also proved by Pang et al. [265].

7.4.7 Limitations of our Approach

Results show that our approach has performed well but we still have some limitations in our work that should be removed. For example, the manual selection of proper nouns and named entities from Wikipedia makes this process a bit unreliable. To tackle with this problem, we present our work for entity retrieval in chapter 10.

Another major problem we found in our approach is the linear combination of topical relevance and opinion score of a document. We believe that if a more effective approach is used for combination of these two scores, we can achieve much better results. The same problem is also shared by our work presented in chapter 6, therefore, we tackle with this problem by using effective machine learning techniques in the work presented in chapter 8.

7.5 Chapter Summary

In this chapter, we presented our passage-based language modeling approach for the opinion finding task. This approach involves expansion of the given query with two types of terms (i.e., relevant terms and opinionated terms). The process of query expansion use the proximity between the relevant and opinion terms to make sure that opinion terms being selected are about the given topic.

7.5.1 Findings

⋄ This passage-based LM approach determines how a relevance-based LM approach can be adapted for opinion finding task by using the distribution of opinion terms in documents.

⋄ During experiments, it is found that processing a subset of relevant textual segments (passages in this case) of a document is better than processing of the complete document for opinion finding task.

⋄ Combination of term's occurrences and its subjectivity score (labeled as *ALL* in table 7.3) is more effective than just using occurrence-based evidences (labeled as *FREQ* in table 7.3) for opinion finding task. It also

testifies the differences between opinion-based retrieval and topic-based retrieval where term frequencies play very important role.

7.5.2 What needs to be improved?

The approach discussed in this chapter gives good results but still suffers due to some drawbacks that are discussed below.

⋄ This work does not evaluate the role of query expansion in opinion finding task and similarly many related questions remain unanswered. For example, it could have been very interesting to see how topic-relevant terms and opinion terms in expanded query are affecting the opinion finding results.

⋄ More effective techniques (like machine learning algorithms) are needed for combination of relevance and opinion scores of a document to further improve the effectiveness of this approach.

Some of these limitations have been tackled by our work presented in chapter 8.

Section II

Dealing with Topic Dependencies

Combining Topic-Dependent and Topic-Independent Evidences For Opinion Detection

It is not best that we should all think alike;
it is a difference of opinion that makes horse races.
Mark Twain

8.1   Introduction

In this chapter, we propose an approach which combines topic-independent and topic-dependent evidences for the task of opinion detection. We use very simple heuristic-based features for this task. Most of the features we used for this work are reformulations of the features proposed in chapter 6 that have been combined using machine learning technique. Our proposed approach proves its effectiveness by improving the O.F. MAP of five baselines provided by TREC. Apart from this, we also prove that adverbs and adjectives could be the best indicators of subjectivity of a document while computing the opinion score of a document.

This chapter is organized as follows: Next section (section 8.2) describes the motivation for this work with section 8.3 giving details of the features used for this work. Experimentation details, its results and discussions are given in section 8.4. We end the chapter with a short summary of the chapter.

8.2   Motivation

Bing Liu [202] defines *opinion mining* as, "the ability of recognizing and classifying opinionated text within the documents". This definition gives us an overview of a scenario where we are given a set of documents with a task of sorting out the opinionated documents from non-opinionated documents without any concerns with relevancy to a particular topic. However, this task becomes more difficult when it requires to find opinionated documents about a particular topic. Few opinion finding approaches try to overcome difficulty of this task by using topic-related information which helps to find associations between topic and opinions present in documents. These types of approaches are called *topic-dependent approaches* [247]. Using topic or domain related information can be helpful to improve the performance but doing so loses scope and generalization of the approach because their performance varies from domain to domain. On the other hand, topic-independent approaches [169, 400] are capable of retaining their generalization across the domains but generally do not perform well because they do not take support of the useful information provided by topic. It seems that there exists a trade-off between generalization and performance in this regard. In this situation, an ideal approach will combine topic-independent and topic-dependent evidences to keep positive points of both type of approaches (i.e., good performance and generalization). In this chapter, we propose an approach which combines topic-dependent and topic-independent evidences for the task of opinion finding. Motivation for the features we are using in this work have already been described in chapter 6.

We have already discussed topic-independent and topic-dependent approaches in chapter 4 while discussing challenges for opinion detection (section 4.4). Moreover, there are few approaches [120, 172, 173] that have used similar kind of evidences as we do (i.e., features like "reflexivity" and "addressability" discussed below in section 8.3) but these approaches were not able to give very good results. Yang et al. [172, 173] created an *IU lexicon* by extracting *n*-grams involving terms like *I, you, yours, me, etc.* from Cornell[1] movie data collection and positive blog training data. Zhou et al. [120] experimented with a proximity approach between title of the topic, opinion words like *like, love, hate, suck, nice, good, bad, awesome, awful, never, think and, feel* and words like *me, we,*

---

[1]`http://www.cs.cornell.edu/people/pabo/movie-review-data/`

I, they, you, he, she. In our work, we have polished these topic-independent
simple heuristics-based features and later on combined them with very basic
topic-dependent features (i.e., relevance score and relevance rank) using a very
effective machine learning algorithm. The experimental results show the effec-
tiveness of combination of features used.

Even our proposed approach uses same kind of features as many other have
already used but our approach can be distinguished from others on behalf of
following points:

1. Our approach does not require any external feedback for opinion finding,

2. We do not use any external data collection for training data,

3. Uses relevance evidence of the documents and therefore adaptable for
   baselines of different strengths,

4. It does not use any proximity measures.

In next section, we present our opinion finding approach.

8.3   Opinion Finding Features

We propose 43 content-based opinion features (see appendix B) but in this
section we will discuss only those that performed well during experimentation.
Below is the detail for each selected feature.

8.3.1   Topic-Independent Features

In this section, we will discuss topic-independent features we used for our work.
We have categorized the features according to their types.

Parts of Speech (POS)-based Features

We have figured out some POS-based features like average number of adjectives,
average number of adverbs, average number of verbs, average number of nouns
and emotivity etc. Most of the features are normalized by number of words in
the document. One of the most important POS-based feature is emotiveness
formulated in two different ways as explained below.

Emotivity (I)

Researchers have been exploiting presence of adverbs and adjectives in a document as an indicator of its emotivity [406]. Assuming it an important clue of opinionativeness of a document, we calculate Emotivity of a document by counting the numbers of adverbs and adjectives in a POS-tagged document.

$$Emot_I(d) = \frac{|t_i \in d : type(t_i) \in \{Adjectives, Adverbs\}|}{|t_i \in d : type(t_i) \in \{Verbs, Nouns\}|} \qquad (8.1)$$

Emotivity(II)

Emotivity (II) is a variation of feature Emotivity(I) obtained by normalizing Emotivity (I) by the total number of words in the document d and is given below:

$$Emot_{II}(d) = \frac{Emot_I(d)}{|D|} \qquad (8.2)$$

where $Emot_I(d)$ is *Emotivity(I)* as shown in equation 8.1 and $|D|$ is the total number of words in document d.

Subjectivity-based Features

Subjectivity-based evidences can be very effective. Therefore, we are using the popular lexical resource SentiWordNet(SWN) [100] for calculating the subjectivity of the terms present in a document. Various features were proposed like average number of positive (and negative, neutral) words in a document, number of negations (neither, nor, neither etc) in a document. Besides this, we have proposed four different Subjectivity functions discussed below. Subjectivity is a document level feature here so subjectivity of each term is summed up to calculate the subjectivity of the document. Subjectivity of a term t is calculated as:

$$Subj(t) = \frac{\sum_{s_i \in senses(t)} (Neg(s_i) + Pos(s_i))}{|senses(t)|} \qquad (8.3)$$

In equation 8.3, $Neg(s_i)$ is the negative score of the sense s_i of term t as found in SWN, $Pos(s_i)$ is the positive score of the sense s_i of term t as found in SWN and $|senses(t)|$ is the total number of senses for term t in SWN.

Using the subjectivity of a term t found in a document, we have proposed four
different subjectivity functions that are discussed below:

Subjectivity(I)

$$Subj_I(d) = \frac{\sum_{t \in d} Subj(t)}{|X|} \qquad (8.4)$$

Where $Subj(t)$ is the subjectivity score of a term t as calculated using equa-
tion 8.3. $|X|$ is the total number of document terms found in SWN.

Subjectivity(II)

$$Subj_{II}(d) = \frac{\sum_{t \in d} max(\sum_{t \in SWN} Pos(t), \sum_{t \in SWN} Neg(t))}{|X|} \qquad (8.5)$$

Where $|X|$ is the total number of document terms found in SWN. In this func-
tion, we prefer to give a positive score to a document if the number of positive
terms prevails in the documents and vice versa.

Subjectivity(III)

$$Subj_{III}(d) = \frac{\sum_{t \in d} Subj(t)}{|D|} \qquad (8.6)$$

Its almost same as *Subjectivity(I)* but in this case we normalize the function
with i.e., $|D|$ the total number of words in the document d.

Subjectivity(IV)
The idea behind the proposal of this feature is to just focus on strongly subjec-
tive terms of a document to calculate the subjective score of the document d.
Therefore, a threshold value of 0.5 is fixed to select only those terms from the
document which have strong sense of opinionativeness.

$$Subj_{IV}(d) = \begin{cases} \frac{\sum_{t \in d} Subj(t)}{|V|} & if\, Subj(t) \geq 0.5 \\ 0 & otherwise \end{cases} \qquad (8.7)$$

Where $|V|$ is the total number of document terms having subjectivity value
≥ 0.5.

Note: It is to be noted that subjectivity of a document d is computed using set
of unique words (only adjectives and adverbs) in it (i.e., term frequencies (tf)
of words are not taken into account) having $length > 2$.

Common Heuristics based Features

Reflexivity

The motivation behind reflexivity feature has already been described in chapter 6. However, its formulation for this work is little different from the one given in chapter 6. We represent reflexivity as $Refl(d)$.

$$Refl(d) = \frac{|\{t : t \in (d \cap R)\}|}{|R| + |A|} \tag{8.8}$$

Where R represents the list of reflexive pronouns we prepared and $|\{t : t \in (d \cap R)\}|$ is the number of occurrences of reflexive pronouns found in document d with $|R|$ and $|A|=$ Total number Pronouns in two lists we prepared (A is the addressability list discussed below).

Addressability

Similar to reflexivity feature, the detail about addressability feature has already been given in chapter 6. Like reflexivity, a little change has been done in formulation of addressability feature. We represent the addressability of a document d as $Addr(d)$ and it is given in equation 8.9:

$$Addr(d) = \frac{|\{t : t \in (d \cap A)\}|}{|R| + |A|} \tag{8.9}$$

Where A represents the list of addressive pronouns we prepared and $|\{t : t \in (d \cap A)\}|$ is the number of occurrences of A terms in document d with $|R|$ and $|A|=$ Total number Pronouns in two lists we prepared.

8.3.2   Topic-Dependent Features

We used two topic-dependent features which are relevance score and relevance rank. Both of these provide information about the relevancy of a document. Even these are topic-dependent features but these two features do not affect generalization of our approach because their impact is limited to relevance of a document only.

Relevance Score

To provide an evidence of its relevance, we use relevance score of a document (provided in baseline) as a feature.

Relevance Rank

Rank information complements the information delivered by relevance score.

8.4 Experimentation

Acknowledging the importance of Blogs as a rich source of opinions, we decided to use blogs as our test data collection and the best available choice was to use TREC Blog 2006 data collection [212]. We evaluate our approach using TREC provided five baselines for 50 topics of year 2007. Baseline-4 is the strongest of the baselines provided by TREC and therefore, we choose it to experiment with topics of year 2006 and 2008 too (50 topics for each). Apart from this, we use baseline-4 for evaluating individual features using topics of year 2007 (see section 8.4.1 below).

8.4.1 Individual Features

The evaluation of each individual feature (combined with relevance score and relevance rank) is performed using 5-fold cross validation for Support Vector Machines (SVM). We used TREC provided strongest baseline (i.e., baseline-4) for this task for topics of year 2007. Table 8.1 shows the results of opinion finding MAP for few important features for which good results were obtained or which helped to improve results when used in combination with other features (discussed in next sub-section 8.4.2).

Table 8.1 shows that Subjectivity(IV) performs the best among all others. It is followed by Subjectivity (III) feature which improves the O.F. MAP of baseline by a very little margin of 0.44%. The feature addressability is the third feature that managed to improve the baseline, even this improvement is very low to be considered. Other features in the table 8.1 fail to improve the baseline-4 but

Table 8.1: *O.F. MAP for individual features*

Feature	Sub-Feature	O.F. MAP
Baseline	-	0.3784
Emotivity	Emotivity(I)	0.3754
	Emotivity(II)	0.3209
Subjectivity	Subjectivity(I)	0.3750
	Subjectivity(II)	0.3728
	Subjectivity(III)	**0.3801** (+0.44%)
	Subjectivity(IV)	**0.3879** (+2.51%)
Addressability	Addressability	**0.3787** (0.07%)
Reflexivity	Reflexivity	0.3718

performed good enough to be considered as part of the combinations to be evaluated (see next section 8.4.2).

8.4.2   Combining Features

We compare the effectiveness of different combinations of features on TREC Blog 2006 Data Collection. We use a transformation function for combining different feature to get a better result which is given as [79]:

$$g(\chi, \Theta) = \frac{\chi}{\chi + \Theta} \tag{8.10}$$

Where χ is the feature to be transformed and Θ is the transformation parameter. Using this transformation function performs better than a linear combination of features. However more non-linear transformations can also be explored. Different combinations were evaluated and we report the results for best performing combination Comb-4 (the best combination found is shown in table 8.2).

Table 8.2: *Best Feature Combination (i.e., Comb-4)*

Feature	Type
Emotivity(I)	*POS Type*
Subjectivity(IV)	*Subjectivity Based*
Reflexivity	*Heuristic*
Addressability	*Heuristic*
Relevance Score	*Relevance Based*
Relevance Rank	*Relevance Based*

Evaluation of Combination Comb-4

First of all, we evaluate this combination using topics of year 2007 for baseline-
4. We use several machine learning classifiers for this purpose. A 5-fold cross-
validation is performed with classifiers Multinomial Logistic Regression (with a
ridge estimator [186]), Support Vector Machines (SVM) and Naive Bayes clas-
sifier (NB). The evaluation measures being used to report results are P@10
and MAP. All of the classifiers improved the results of the baseline with the
combination (Comb-4); however SVM performed the best by giving 5.5% of
improvement in O.F. MAP of baseline (see table 8.3). The results show that
combining subjectivity based evidences with reflexivity and addressability in-
cluding the relevance score of the document can be very effective when looking
for opinions in blogposts.

Table 8.3: *Results using baseline-4 for TREC Blog 2007 topics with combination-4 (Comb-
4). An asterisk (*) indicates statistical significance w.r.t. O.F MAP*

RUN	MAP	P@10	Classifier
Baseline-4	0.3784	0.5340	-
Comb-4-BL4-LR-2007	0.3955*(4.5%)	**0.5800***	**Logistic Regression**
Comb-4-BL4-SVM-2007	**0.3994***(5.5%)	0.5640	**SVM**
Comb-4-BL4-NB-2007	0.3835 (1.3%)	0.5760*	**Naive Bayes**

Knowing that SVM could give better results using Comb-4 for the task of opin-
ion finding, we perform experimentation using baseline-4 for topics of year 2006
(Comb-4-BL4-SVM-2006) and year 2008 (Comb-4-BL4-SVM-2008) and results
are shown in table 8.4. Similarly, we evaluate our approach for other baselines

Table 8.4: *Results using baseline-4 for TREC Blog 2006 and 2008 topics with combination-4
(Comb-4) and SVM. An asterisk (*) indicates statistical significance w.r.t. O.F MAP*

RUN	MAP	P@10
Baseline-4 (2006)	0.3022	0.524
Comb-4-BL4-SVM-2006	0.3124	**0.5940***
Baseline-4 (2008)	0.3822	0.6160
Comb-4-BL4-SVM-2008	**0.3893***	**0.6500***

provided by TREC using topics of year 2007 and SVM. The results are shown
in table 8.5.

Table 8.5: *Opinion Finding Results of different baselines for topics of year 2007. An asterisk (*) indicates statistical significance w.r.t. O.F MAP and a † indicates the best reported results ever (to the best of our knowledge) [307]*

Baseline	MAP	P@10
Baseline-1	0.2758	0.4540
Comb-4-BL1-SVM-2007	0.2978*	0.4980
Baseline-2	0.3034	0.5320
Comb-4-BL2-SVM-2007	0.3056	0.5400
Baseline-3	0.3488	0.5760
Comb-4-BL3-SVM-2007	0.3743*†	0.6120*
Baseline-4	0.3784	0.5340
Comb-4-BL4-SVM-2007	0.3994*†	0.5640*
Baseline-5	0.3805	0.5580
Comb-4-BL5-SVM-2007	0.3897	0.5840*

Adjectives and Adverbs: Better Indicators of Document Subjectivity

Table 8.3 shows the experimentation results when we consider only adverbs and adjectives while computing subjectivity of a document. Being curious about other parts-of-speech combinations, we compute the subjectivity feature with other combinations of POS. Experimentation results for same combination of features (i.e., Comb-4) as for run Comb-4-BL4-SVM-2007 with different combinations of POS for subjectivity features are shown in table 8.6.

Table 8.6: *MAP and P@10 Results with Different Combinations of POS for Subjectivity Feature*

Run	Combination	MAP	P@10
Comb-4-BL4-SVM-2007-VAA	Verbs+Adjectives+Adverbs	0.3946	**0.5760**
Comb-4-BL4-SVM-2007-VAJ	Verbs+Adjectives	0.3960	0.5640
Comb-4-BL4-SVM-2007	Adjectives+Adverbs	**0.3994**	0.5640
Comb-4-BL4-SVM-2007-VAV	Verbs+Adverbs	0.3961	0.5680

Even the results (for both MAP and P@10) given in table 8.6 do not differ by big margins but are good enough to suggest that using adverbs and adjectives might be a good POS combination while computing the subjectivity of documents (run Comb-4-BL4-SVM-2007).

Effects of Reformulated Reflexivity and Addressability

Similarly, we repeated the experiments of table 8.3 using same formulations of
features reflexivity and addressability as presented in chapter 6 (run Comb-4-
BL4-SVM-2007-CH6). A significant fall in performance is observed when we
use the formulations of chapter 6 for reflexivity and addressability features (see
Table 8.7).

Table 8.7: *Re-Computation of C4-SVM Results with Formulations of Chapter 6 for Reflex-
ivity and Addressability Features*

Run	MAP	P@10
Comb-4-BL4-SVM-2007	**0.3994**	0.5640
Comb-4-BL4-SVM-2007-CH6	0.3764	0.5500

8.4.3   Discussion

Table 8.3 shows the results for opinion detection task using three different clas-
sifiers for combination we have used. Results show that SVM gives the best
results for this combination of features while logistic regression model standing
second in rank and Naïve Bayes gives the least best results. Opinion finding
MAP achieved using SVM (i.e., 0.3994) gives an improvement over previously
best reported O.F. MAP of 0.3968 (to the best of our knowledge) of Santos et
al. [307] over topics of year 2007. The significance of the improvement in the
results of baseline was validated through t-test (with $p < 0.05$). It is to be
noted that for a fair comparison with previous work, we have to compare the
results of our approach with an approach using TREC provided baseline-4 for
the task of opinion detection and which stands results of Santos et al. [307] the
best previous results ever reported for TREC 2007 topics.

The chart in figure 8.1 shows the degree of improvement of our best run (SVM)
over baseline (BL) for each topic. We got improvement over 40 topics of total 50
topics and an average improvement of 12.26% was observed among improved 40
topics. The maximum improvement of 41.60% was noted for topic 936 (Grammy
Awards) and minimum improvement of almost 0.17% was noted for topic 917
(snopes). Table 8.8 shows the top 10 topics for which we could improve the
results with improvement mentioned in percentage over baseline.

An analysis of the topics with improved results reveals that generally a trend of
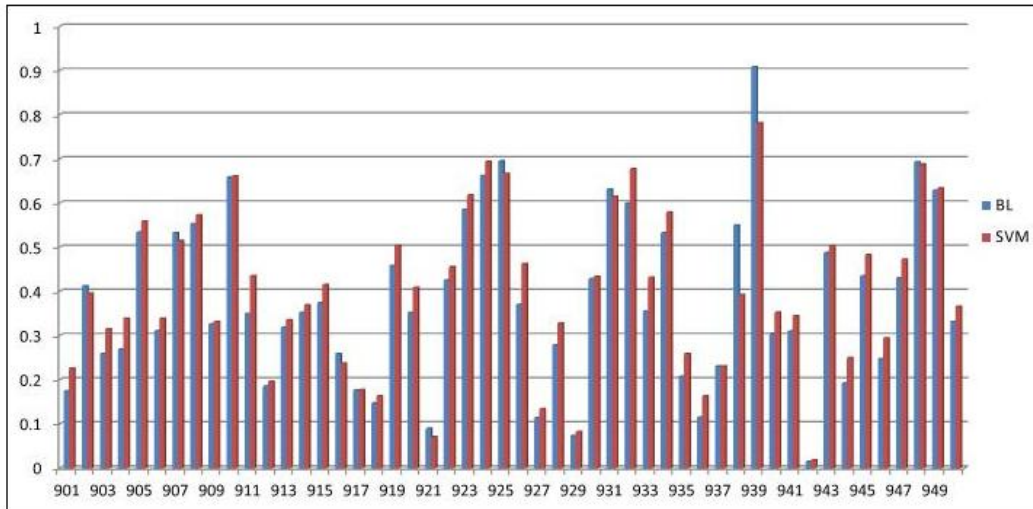
Figure 8.1: *O.F. MAP comparisons between Baseline-4 and Comb-4 using SVM for topics of year 2007 (run Comb-4-BL4-SVM-2007)*

Table 8.8: *Few topics for which the results were improved (run Comb-4-BL4-SVM-2007)*

Num	Topic	BL	SVM	%age Imp	Title
1	936	0.1149	0.1627	41.60	Grammy awards
2	944	0.1916	0.2496	30.27	Opera Software
3	901	0.1749	0.2255	28.93	Jstor
4	904	0.2693	0.3385	25.69	Alterman
5	926	0.3699	0.462	24.89	Hawthorne Heights
6	935	0.2073	0.2587	24.79	Mozart
7	911	0.3496	0.4351	24.45	SCI FI Channel
8	903	0.2585	0.3149	21.81	Steve Jobs
9	933	0.3552	0.432	21.62	Winter Olympics
10	946	0.2467	0.2939	19.13	Tivo

Table 8.9: *Few topics for which the results were not improved (run Comb-4-BL4-SVM-2007)*

Num	Topic	BL	SVM	%age Imp	Title
1	902	0.4121	0.3954	-4.05	Lactose Gas
2	907	0.5324	0.5148	-3.30	Brrreeeport
3	916	0.2588	0.2375	-8.23	Dice.com
4	921	0.0895	0.0706	-21.11	Christianity Today
5	925	0.6951	0.6669	-4.05	Mashup Camp

improvement was seen among the topics of general public interest like topics like Grammy Awards, Opera software, Alterman (columnist and author), hawthorne heights (musical band), and Steve Jobs, etc. All of these are whether public figures (well known in U.S.A at least) or products that concern public1 for daily life use. In blogs, people read or comment on blogpost they are interested in. Therefore, mostly topics of general public interest get more attention and such posts are more visited than others which are discussing less popular topics. This observation also confirms the results of Rijke [18] where he concludes that blogs with more number of comments tend to be more opinionated than blogs with less number of comments. Table 8.9 shows the few topics for which our approach could not improve the results at all. If we look at the topic titles, we note that such topics are not popular among public or their popularity is limited to a certain group of people.

Similarly, table 8.5 summarize the results of evaluation of Comb-4 using SVM for all baselines provided by TREC. The topics of year 2007 are used for evaluation with all baselines. Our approach managed to improve all baselines while beating previously published O.F. MAP results [307] for baseline-3 (run Comb-4-BL3-SVM-2007 0.3703 Vs 0.3743) and baseline-4 (Run Comb-4-BL4-SVM-2007 0.3968 Vs 0.3994). These results prove the effectiveness of our approach [53].

While discussing the results of table 8.7, it should be noted that in formulations of reflexivity and addressability for chapter 6, the number of occurences of pronouns (*I, me, my,* etc. for reflexivity and *you, yours, yourself,* etc. for addressability) are normalized by total number of words in the document while formulations for reflexivity and addressability, as shown in equations 8.8 and 8.9, are normalized by a constant (i.e., by total number of pronouns in both lists R and A we prepared). Formulations of chapter 6 for both of these features give very poor results as compared to the formulations used for this work. Most probable reason behind this poor performance seems to be the normalization factor used in formulations of chapter 6. It has been observed that it is mostly the "Comments Section" of a blogpost which is more likely to contain words like *I, me, mine,* etc. (reflexivity) or *you, yours,* etc. (addressability) used to emphasize one's opinion or to address others respectively while authoring a comment. Therefore, the ratio of such words to total number of words in a blogpost may result to very insignificant values and this problem becomes more severe when a long document contains many non-opinionated segments within it. This seems to be the only reason behind this poor performance of our approach when experimented with formulations of chapter 6 for reflexivity and addressability

features. Even our approach has given significant improvement over its base-
line and has beaten the previous best results but we can see few drawbacks
associated with it. One of the problems that we see in our approach is the use
of such features which might be specific to blogs or blog type documents (like
homepages) and may not work well with other genre of opinionated documents
(like news editorials). For example, use of pronouns (like *I, me, we, us,* etc.,)
produced good results for opinion finding in blogs but such pronouns may not
be as frequently used in news articles. This is just an intellection however which
if tested can be interesting addition to the contributions of this work.

8.5   Chapter Summary

In this chapter, we have proposed an approach which uses a set of topic-
independent opinion finding evidences combined with relevance-based topic-
dependent features to find opinion documents from TREC Blog 2006 collec-
tion. Our approach has performed well by giving a significant improvement
over TREC provided baselines and also improving previously best reported re-
sults for the task of opinion finding. Unlike our previous approaches, we have
avoided the use of query expansion technique which is generally discouraged in
IR research community.

8.5.1   Findings

⋄ Good performance of our approach affirms that very simple heuristics-
based features could prove to be effective when optimised well for opinion
finding task.

⋄ Adjectives and adverbs form a better combination of parts-of-speeches for
computing the subjectivity of a document.

⋄ More popular topics tend to be more opinionated.

⋄ Role of topic-related information can be minimized without significantly
affecting the performance of an O.F. approach.

8.5.2   What is lacking?

Despite its good performance, our approach can be further improved on
different fronts like

⋄ Role of different type of parts-of-speeches need to be explored further in
detail.

⋄ It could have been interesting to analyze the role of these features on
different granularity levels.

We plan to focus on removing these deficiencies of our work as part of our future
work.

Section III

Use of Social Evidences

Chapter 9

A Preliminary Investigation on Using Social Network Based Evidences for Opinion Detection in Blogs

Opinions founded on prejudice are always
sustained with the greatest violence.
Hebrew Proverb

9.1   Introduction

In this chapter, we propose our preliminary work for proposing a framework which combines several evidences (content-based and social network based evidences) for performing the tasks of sentiment detection, sentiment prediction and multidimensional ranking. *Sentiment Prediction* is a task of estimating sentiment of a blogger for a specific topic by using its social-evidences in the absence of enough content-based evidences. Proposing approaches for this task can be helpful for many other tasks like identifying communities of bloggers with the same sentiments (positive or negative) for a topic within blogosphere. Similarly, the task of multidimensional ranking allows us to rank and view the blogposts according to many different contexts i.e. relevancy, opinion score, trust, quality, polarities and with respect to age and gender of a blogger. For example, if someone wants to look at the blogposts published by female bloggers

on the topic of *abortion* ranked in descending order of their opinion score then multidimensional ranking task can be useful in such case.

This chapter is organized as follows: The elements of online social networks that can be helpful for opinion detection and prediction are discussed in section 9.3 with section 9.4 introducing a framework composed of these evidences. In section 9.5, we are discussing some challenges to be faced while utilising these elements of online social networks and section 9.6 presents a demo justifying the use of social networking evidences for opinion finding task. The related work for this contribution can be consulted in chapter 4.

9.2   Blogs: An Ideal Choice for Temporal Data Analysis

Statistical surveys show that blogs are proliferating at an ever-increasing rate. Million of bloggers publish posts on a variety of topics ranging from cooking recipes to international politics. For example, product reviewing is one of the most popular activity in blogosphere. Product review experts or people having already used a product share their expertise and experiences in their blogs and help their readers in decision of buying a product. A survey, conducted by Universal McCann[1] in July 2009, reports that 32% of the 200 million bloggers worldwide blog about opinions on products and brands. In addition, it was found that 71% of the 625 million active Internet users actually read blogs. Another very interesting finding reported in a survey conducted by Nielson[2] is that 78% of readers trust the opinion of other consumers about a product. These statistics suggest that blogs can be considered as a reliable source of opinions. Similarly discussion of political issues and agendas is another very popular activity of blogosphere. In fact, blogging has changed the way politicians used to convince public about their party agenda. Being well-aware of the popularity of blogs among general public, Howard Dean, one of the major contestants for the 2004 Democratic presidential nomination, created official campaign blog[3] to keep people informed of his political agenda and to stay informed of public opinion. Similarly, democracy advocates in both Iran and Iraq chose blogging

---

[1] http://www.universalmccann.com
[2] http://www.nielsen.com
[3] Edward Cone, The Marketing of the President 2004, Baseline Magazine, December 2003

as a technique for registering dissent[4].

Besides popularity, effectiveness is another important characteristic of blogs. The effectiveness of blogs can be estimated by their influence on real world events. One of the major reasons for Howell Raines resignation as editor of the New York Times in June 2003 was the heightened attention bloggers gave to Jayson Blair scandal [102].

Temporal nature of blogs is another very important characteristic of blogs. It forms the foundation for time-based trend analysis in blogosphere. Time-based trend analysis has utmost importance for many domains. For example, in order for businesses to make judicious decisions, it is very important for them to track customer opinions and complaints in a timely fashion. Here the blogosphere provides free large scale information sources from which businesses can quickly learn opinions and complaints from their customers. Due to its temporal nature, the blogosphere is much more dynamic than traditional Web pages [64]. For example, an announcement of a new product may instantly trigger intensive discussions in the blogosphere. Very often, it is exactly these dynamic trends that are valuable for businesses to track, understand, and predict the interests of their customers. Recently, many commercial blog and Web search engines have introduced services for temporal trend analysis. For example, for given keywords, BlogPulse[5] and IceRocket[6] generate trend curves over time in terms of the percentage of blog entries that contain the keywords. For a given tag, Technorati[7] provides curves that show the daily number of entries that adopt the tag. Google has just announced a new service called Google Trend[8] that, for given keywords, plots the search volume and news reference volume that are related to the keywords over time. Figure 9.1 shows a Google trend curve for keyword *French Strike* with highest peak in year 2010 (marked as E) because of on-going strikes against new pension laws approved by the French President.

One of the most important features of blogosphere is its networked structure. Existence of links between bloggers forms a social network of bloggers. A blog page may contain many types of links which generally includes:

⬦ Links to some old blog entries that might be relevant to the current blogpost.

---

[4]http://www.ragingcow.com
[5]http://www.blogpulse.com/
[6]http://trend.icerocket.com/
[7]http://www.technorati.com/tags/
[8]http://www.google.com/trends

⋄ Links to blog entries of other bloggers in the blogosphere or links to some external source like news article, editorial, etc.

⋄ A blog's *blogroll* refers to a list of links to other blogs that usually occupy a permanent position on the blogs home page. They provide a representation of a bloggers interests and preferences within the blogosphere. Bloggers are likely to use their blogrolls to link other blogs that have shared interests.

⋄ Links to other bloggers, having interest in the blogpost, present at the end of a blogpost are called *Trackbacks*.
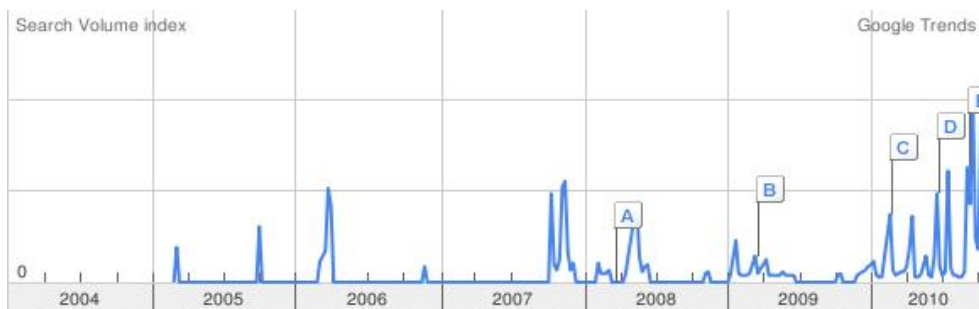
Figure 9.1: *Google Trend Curve for keyword French Strike*

In light of all major blog features discussed above, Jones et al. [159] declares blogs as an ideal choice to be chosen as a data collection for research purposes. Blogs have also inspired the researchers working in the domain of opinion mining because of the popularity, influence and reliability of opinions contained within blogs. Realizing the need for effective opinion finding systems and considering blogs an ideal data collection for this task, TREC started a blog track in year 2006 and its details can be seen in chapter 3. There exist a lot of approaches for opinion finding in blogs and some of them have performed very well. According to Ounis et al. [260], two types of approaches have been used for opinion mining in blogs: Machine-Learning approaches and Lexicon-based approaches. Both kind of approaches use the content-based evidences for finding opinions in blogs and no one actually benefitted from the social features of blogs.

9.3 Basic Infrastructure of Blogosphere

The basic structure of the blogosphere in the context of our proposed approach is shown in the figure 9.2. Four major elements of the blogosphere are discussed in following subsections.

9.3.1 Blogger's Profiles

Generally, the creation of a blog on a blog site (or a personal account on any other online social network) requires some personal information for the personal profile of the blogger. It's not obligatory that blogger should give all information about him/her-self but most of the time information about his/her age, gender and location can be found in bloggers profiles [141]. Therefore, we are considering only these three parameters for blogger's profile. These parameters can play very important role while predicting opinions or analyzing trends among bloggers for a certain topic or event. For example, views of women and men can differ on the issue of abortion. In addition, opinions and interests can be categorised according to bloggers age, gender and locations. Also these three parameters can be used to remove or locate biasness among the opinions like Kumar et al. [182] has provided the list of common interests shared among people of same ages.

9.3.2 Blogposts

Bloggers posts are called Blogposts. All blogposts of a blogger are published in a chronological order so that the latest one is always on the top. Each blogpost is followed by a comment section where readers of a blog can post their comments about the topic being discussed in the blogpost. This creates an environment of discussion in the blogosphere and gives rise to a rich source of opinions. In the figure 9.2, blogposts have been marked as $X_B P_1$, $X_B P_2$, etc., for blogger X and similarly for blogger Y and blogger Z.

9.3.3 Blogger's Network of Friends

The bloggers not only provide informational and valuable content in their blogs but also mention links to other valuable information. A blogroll is a list of links to blogs that the blogger likes. A blogroll is usually included in the blog's homepage sidebar. Blogrolls represent bloggers' interests and preferences. For example, a blogger who mostly writes about sports cars and other vehicles might provide links to other bloggers who write about cars, car repairs, car sports etc. The links to other blogs can be added on a blog (i.e. in blogroll) for two reasons:

⋄ The author of the linked blog is blogger's friend. We call these Friend's connections.

⋄ The blogger is interested in the linked blog and reads it's regularly. We call these Relevancy connections.

These links can be seen in the figure above. They are marked between nodes of bloggers like an arrow from blogger Y to blogger X shows that blogger Y has a link to blog of blogger X. Similarly links between blogger Y and blogger Z can be seen too. We assume these links as marking of their friends (or interests) within a blogosphere. Kumar et al. [182] found that there exists a correlation between having common friends in blogosphere and having common interests. The network of very close friends can be very helpful to predict the opinion of a blogger. Same kind of evidences can be used for calculating their scores for certain posts and then a partial score can be transferred to each of the nodes of their friends. This will go like this to friends of friends. In fact each node in the network represents a blogger. Each blogger would have a Total score of its node coming from its sub-nodes representing its user profile, its posts etc. Now a part of this total score can be transferred to his friends. How much part this would be depends upon the strength of the link between two bloggers.

9.3.4 Comments

Readers of a blogpost can post their comments for a blogpost as shown in the figure 9.2 above. Comments on different blogposts have been marked as C_1, C_2 . . . C_N with total number of N comments on a certain blogpost. Comments
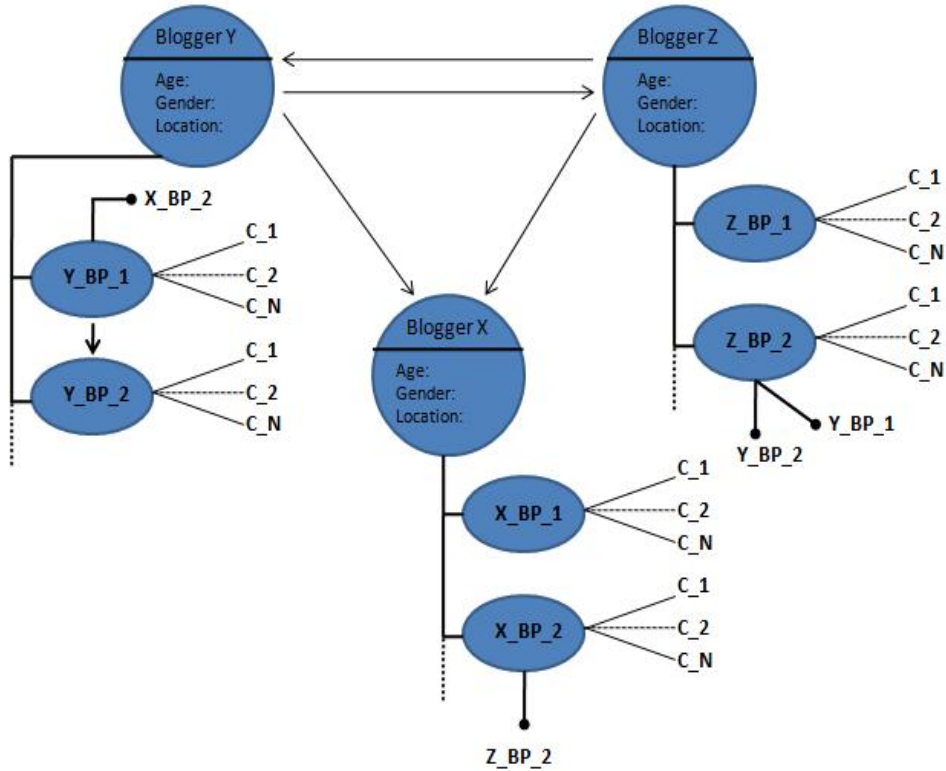
Figure 9.2: *A sample of blogosphere with 3 bloggers (X, Y and Z) highlighted with our
defined parameters*

make very important element of a blog which is rich of opinions and it was
demonstrated by work of Mishni et al. [232] where the authors demonstrated
that by indexing the comments with the blogpost, the recall of the blog search
is increased.

9.4 Framework

In our proposed framework, the blogosphere is represented as a directed graph
where each node (i.e. a blog) is connected to other nodes through edges which
are actually friend's connections (i.e. links in blogroll) or relevancy connec-
tions (links from blogpost/s of blog A to blogpost/s of another blog B). Each
node is represented by a set of variables. Before discussing about these vari-
ables, we would like to introduce two different scenarios our framework deals
with i.e. Topic-Independent scenario and Topic-Dependent scenario. In topic-
independent scenario, we do not consider any topic and all computations for

variables of a blog are done by taking into account all of its blogposts while in topic-dependent scenario, we are input a topic and all computations for variables of a blog are done in respect to the given topic i.e. we consider only topic-relevant blogposts for computing the variables for a blog node. The details of variables representing a node are given below:

◇ Blogger's profile: A blogger's profile includes age, location and gender of the blogger and is represented as PR in our framework. Profile data is entered by the blogger himself and remains consistent but we categorize it as a variable because locations may change with the time. In case if a blog is owned by a company, the time period from its date of establishment would be considered as its age. PR is independent of the given topic.

◇ Relevance: This variable is among the set of variables representing a blog node in topic-dependent scenario and basically represents the topic-relevance score of a blog node for a given topic. It is represented by letter R in our framework. It is a numeric variable whose value lie in range [0..1].

◇ Trust: Trust of a blog is the representation of the trust it gains as a result of computations of social and content-based evidences. This variable represents a blog node in both defined scenarios, however, the way its value is calculated is different in both scenarios. Trust is a topic-independent variable i.e. its value is calculated without any interference with the given topic. It is represented by the letter T. It is a numeric variable whose value lie in range [0..1].

◇ Quality: Quality of a blog is the representation of the quality standards it maintains. This score is calculated as a result of computation of both social and content-based evidences. Its calculated in two different ways in both defined scenarios. It is represented by QT in our framework. Quality is also a topic-independent variable. It is a numeric variable whose value lie in range [0..1].

◇ Opinion: This variable will represent the degree of opinionatedness of a blog or a blogpost. This opinionatedness can be measured with respect to a given topic in topic-dependent scenario and without considering any topic in topic-independent scenario. Content-based evidences can be used for calculating value of this variable. It is represented by letter O and its value lies in range [0..1].

⋄ Polarity: This variable also highlights itself in both scenarios and is computed using content-based and social evidences. Letter P is used to represent it in our framework. Polarity can be represented in both defined scenarios. The possible values for this variable are *positive, negative and neutral.*

It is very important to note that all blogposts of a blog are also represented by the same set of variables as a blog itself in both scenarios. All variables required for representation of a blog node are calculated by combining some blog-level and blogpost-level evidences. Both blog-level and blogpost-level evidences generally include social and content-based evidences. Once we have computed the values of blogpost-level evidences, these can be combined with blog-level evidences to calculate the values of different variables that represent a blog node in a blog network.

In topic-idenpendent scenario, a blog and its blogposts are represented by following set of variables, (PR, T, QT, O, P) where PR is profile of the blogger, T is the trust associated with a blog/blogpost, QT is the overall quality of the blog or blogposts, O for opinionatedness and P is the overall sentiment orientation of a blog/blogpost.

In the topic-dependent scenario, a blog node and blogposts are represented with following set of variables, (PR, R, T, QT, O, P, Q) i.e. (Profile, Relevancy with topic Q, Trust, Quality, Opinionatedness, Polarity, and Topic). Relevancy is the topic-relevance score for given topic Q, Trust score represents the trust value given to a blogger on behalf of different parameters and Quality is the measurement of quality of the posts for a blog. Opinionatedness represents the opinionated nature of a blog or blogpost i.e. higher the score for opinionatedness is, more the document is opinionated and vice versa. Polarity measure basically shows the emotional orientation of the current blogger predicted on behalf of various parameters (content-based or/and his network-based) for topic Q.

Figures 9.3 and 9.4 explain the topic-independent and topic-dependent scenarios respectively. In topic-independent scenario, the score for different variables of the blog Y is calculated via scores of its all blogposts (mentioned as Y_BP_1 and Y_BP_2 in figure 9.3) while in topic-dependent scenario only relevant blogposts are taken into account (marked in pink square i.e. blogpost Y_BP_2 in figure 9.4).
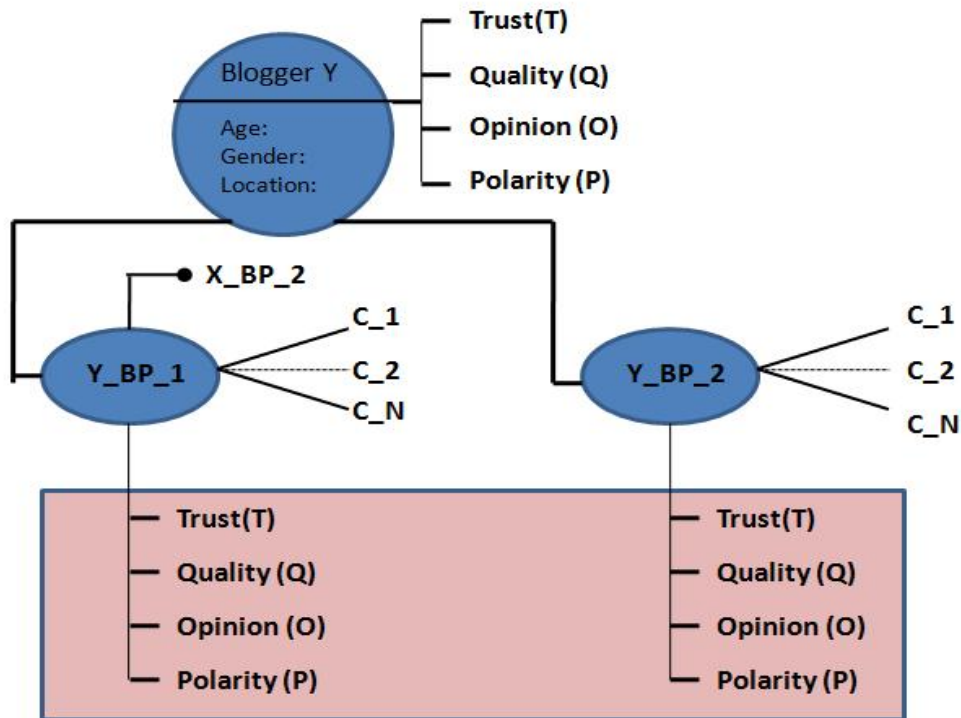
Figure 9.3: *Topic-Independent Scenario with its set of variables*

9.4.1 Tasks

Our proposed framework performs three major tasks that include "Opinion Detection", "Sentiment Prediction", and "Multidimensional Ranking" in topic-dependent scenario. In topic-independent scenario, this framework can handle many tasks too, for example, identifying the most trustworthy nodes in the network, identification of a blog with most quality blogposts, and identifying the bloggers with overall positive attitude towards different issues being discussed in blogosphere, etc. However in this work we will concentrate on the tasks in topic-dependent scenario because lot of further work is required to formalize the tasks in topic-independent scenario. Besides the tasks mentioned above, our framework has the capacity to provide lot of other services like automatic trend analysis, better personalized services, etc., but this work will restricts itself to the discussion of tasks defined for topic-dependent scenario.
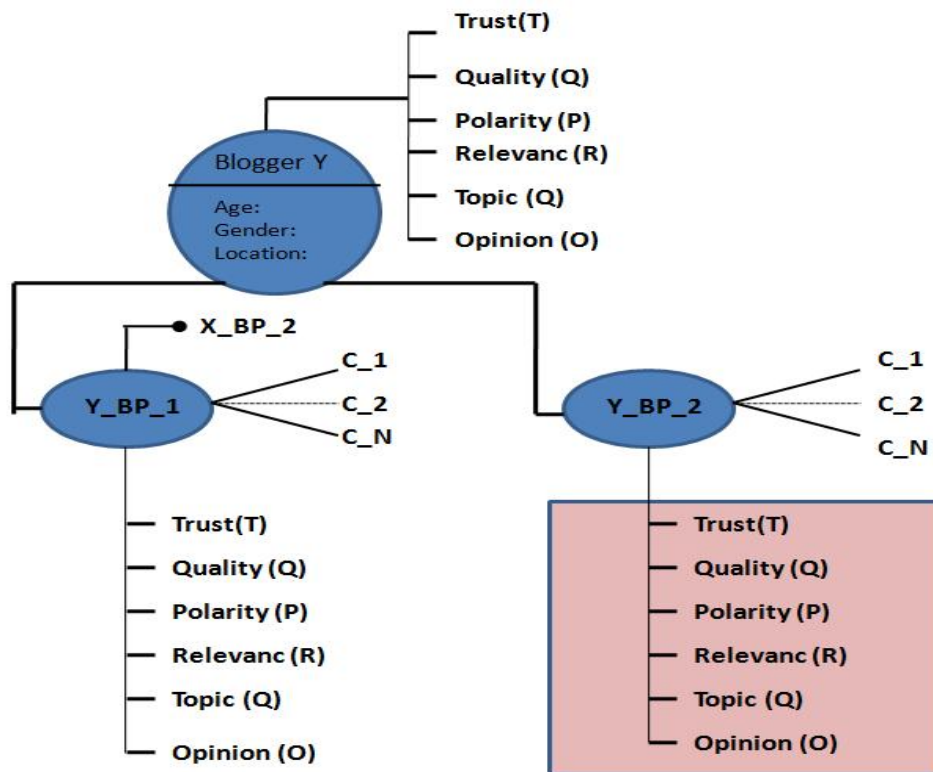
Figure 9.4: *Topic-Dependent Scenario with its set of variables*

Opinion Detection

We have already discussed the task of "opinion detection" in detail in chapter 3
and chapter 4. For completion of this task, any effective approach from the
literature can be adopted.

Sentiment Prediction

The purpose of "sentiment prediction" task is to estimate the sentiment (i.e.,
positive, negative or neutral) of a blogger for a given topic when not enough
relevant content is available to determine his/her sentiment through content-
based sentiment classification techniques. In short, this task finds the answer
for the question, "What might be the opinion of a particular blogger about topic
X?" This is where social network based characteristics of blogosphere (e.g., trust,
quality) can play their role. There are many crucial applications of this task like
predicting the election results before they are held, estimating the popularity of

a film of a specific genre among public before its release, predicting the public reaction before passing a law or amending it etc.

Multidimensional Ranking

Our proposed framework allows us to rank and view the documents according to many facets like relevancy, opinionative nature, trust, quality, polarity of opinions, time, age and gender etc. This is what defines "multidimensional ranking" task. For example, let us suppose a user who wants to retrieve all those blogposts that are being published from location *New York* on subject of *Social Health Care in Unites States* ranked in descending order by their opinionated and trust scores.

Out of various variables mentioned above, topic (Q) and profile (PR) are explicit variables, relevancy (R) and opinionatedness (O) can be computed through approaches as discussed in chapter 2 and chapter 3. Rest of three variables i.e. trust (T), quality (QT) and polarity (P) can be estimated using features as described below. All features have been divided in two classes i.e. content-based features and social-network based features.

9.4.2   Trust Estimation

Jennifer Golbeck [116] defines "trust" between two individuals as:

Alice trusts Bob if she commits to an action based on a belief that Bob's future actions will lead to a good outcome.

Before online social networks, trust was seen as an issue of information security. However, in the context of online networks, it highlightes more social aspects. Trust estimation in online social networking is one the most popular topic these days. Many algorithms [115, 116, 188] for trust estimation in online social networks have already been proposed by researchers.

Advogato[9] serves as a community discussion board and resource for free software developers [190]. Each user on this site is assigned a trust score with the help of a network flow model. Their method works towards computing global trust estimates relative to a set of good peers.

---

[9]http://advogato.org

Appleseed is an algorithm proposed by Ziegler and Lausen [408] for trust calculation. It normalizes the trust values for each person and thus is subjected to poor performance in a social network setting where the degrees of nodes can vary a great deal. Guha et al. [121] proposes a propagation model for trust and distrust. They represent the continuous ratings to binary values of trust and distrust. They observed a relatively low error rate in their calculations.

Jennifer Golbeck proposed a method to quantatively infer trust between users for a recommendation system in web-based social network [113, 114]. She suggested a method to combine the trust levels of all users in the path from a user to the target user to receive information, and applied it to the film recommendation system to run in an actual application. She also proposed a method to infer trust based on user's reputation instead of similarity of preferences in a semantic web-based social network.

To calculate trust value of a blog node or blogpost in our framework, one can adapt already existing approaches as discussed above and modify it (if needed) according to one's needs. However, we also propose some features that can serve as a base to propose an effective approach to infer trust score for a blog/blogpost. These features are discussed below in the form of social and content-based evidences.

Content-Based Evidences

⋄ Spamming is a common phenomenon spreading across the whole web. According to requirements of our framework, each blogpost should be given a non-spam score after analysis of its contents using some spam analysis algorithm [171, 271]. The trust put by spam analysis algorithm in blogposts is eventually used to calculate trust score of the blog.

⋄ Most of the web development and marketing experts[10] suggest that inactivity makes a blog look like a less dependable and credible source. According to marketing experts, recent posting equals frequent posting and if bloggers blog daily, the content is fresh and new! and this is what readers want, and this is how readers put their trust in a blog. In other words, more a blogger is active, more often he posts the blogposts and also comments on different blogposts he is interested in. So we can include this clue as a good indicator of Trust.

---

[10]http://www.websitebusiness.com/blog/2010/03/09/be-an-active-yet-relevant-blogger.html

Social-Network Based Evidences

◇ There is famous saying, *a person is known by his company he keeps* i.e. if a person is mostly found in a company of morally corrupt people, he is more likely to be taken as of the same genre while if he more associated with socially respected persons then he is more likely to be respected. Using this heuristic, we believe that if a blog is having connections with less-reliable blogs (i.e. with low trust scores), it should be given lower trust score.

◇ Similarly, trust of a blogger X can be increased if blogger X has more things common with the bloggers who have high trust values within the blogosphere. Just like EigenTrust [163] considers trust as a function of corrupt and valid files shared between peers, an metric which is a function of similarities with bloggers having high trust values can be useful. In short, bloggers with high trust values propagate their part of trust values to a blogger X if X has few things common with them.

◇ Biasness should be taken into account when looking for opinions. Some opinions might be synthetic and not natural because of some emotional attachment with same gender, age or location depending on the topic of the query [141]. However this evidence might require some external feedback. For example, if we have a query *Cricket in Olympics* to find people's opinion that whether cricket should be included in Olympic Games or not then its most probable that people from Cricket playing nations will go in favor of including Cricket in Olympics while others might resist it or might not take any interest at all. In this case, external feedback will be the name of cricket playing nations like Pakistan, England and Australia etc.

◇ One additional evidence that can be used for trust calculation is the popularity of the blogger himself in real world. We know that these days lots of celebrities are writing their own blogs. The words and statements delivered by celebrities are given more importance and celebrities are considered a more trustable source than newspapers and other sources. However this evidence also need external feedback i.e. framework should be given the list of celebrities and their sites to add extra trust score for blogposts from such sites.

9.4.3 Polarity Estimation

Content-Based Evidences

⋄ Any of the already existing sentiment classifications approaches like machine learning based approaches and lexical based approaches can be used.

⋄ In absence of relevant blogposts in a blog for a certain topic, polarity of related blogposts (less relevant) can be used as an evidence to predict the orientation of bloggers thinking. For example, polarity of all related blogposts for a topic *War on Terrorism* can be analyzed and an overall behavior can be estimated or predicted on behalf of polarity of these blogposts. However, an affective technique is needed here to check if a blogpost is relevant or related to another topic.

⋄ The overall polarity of comments of a blogger on a blogpost relevant to the given topic T can be considered as one of the clues of estimating the polarity of opinion a blogger.

Social-Network Based Evidences

• Friends can influence one's opinion or in other words they can come up with the arguments to change your mind about something. There exist considerable number of works [142, 161] to analyze the way people influence opinions of others in a network. In case of absence of enough content-based evidences for estimating the sentiment of a blogger for a particular topic, relevant posts posted by connected bloggers can be determined by the polarity tag associated with them and can be used to estimate the orientation of bloggers' opinion about that topic

• The user profile can play a vital role to predict the polarity of opinions about some issue in absence of other content-based clues. For example, if we are looking for opinions on the issue of *abortion* then gender will play a very important role in this case. Most probably the women will go against abortion and men might go in its favor. Similar is the case with location and age parameters of the user profile.

9.4.4 Quality Estimation

In this section we will describe the features that can be used to determine the quality of blogs or blogposts.

Content-based Evidences

- The links present in a blogpost content refer to other blogs. If a blogpost is citing trustworthy blogs i.e. the ones with high trust scores then its content can be classified as of good quality.

- Quality of information is also highlighted by the freshness of data [317]. The users want latest and up-to-date relevant information about the topic they are looking for. Therefore, if a blog is regularly updated by the blogger, it can be a good candidate for being declared a qualitative blog.

- Blogs are written by people of all ages. In our framework, age of the blogger is accessible through his profile. It is intuitive to expect more quality content from a mature person than a child of twelve years. Therefore, the age of the blogger can also be considered while computing quality of the blog content.

- The number of comments by the author of the blogpost himself also indicates the extent to which he is trying to defend his/her point of view. Therefore, it can also be a good feature for quality estimation of a blogpost.

Social-Network Based Evidences

- A popular blog is trustful. Normally, the popularity is measured in terms of times a blog has been referenced by others. Therefore, more the number of times it has been referenced, more reliable it is and so has more quality content.

- The number of in-links also indicates the popularity and authenticity of the blogpost in blogosphere and should be taken into account when estimating the quality.

- Similarly, the popularity of out-bound links from within the blogpost can be important to estimate quality. The popularity of the out-links can be calculated using PageRank etc.

- The percentage of out-links to itself i.e. to his previous blogposts indicates that bloggers is sure of the authenticity of information he has been providing. Therefore, it can be used as an indicator too.

9.5   Challenges

9.5.1   Privacy Issues

From the birth of online social networks, the issue of privacy has hindered the creation of innovative ideas. The bloggers range from young children to old retired persons. Young kids in their blogs reveal too much about themselves, their families etc. by posting their and family photos, mentioning the school names etc. In most of the cases, this disclosure of private information reaches a level that is troubling for parents. In addition to writing about their personal lives, people also express their views on political, religious and social issues in blogs. Difference in opinion on some contentious issues like war on terror, abortion, global warming, etc can create social problems for real world communities. Knowing or reading someone's opinion about something may be acceptable for most of the bloggers (because after all they are publishing it) but using his/her reviews or personal life for research purposes may not be pleasant for them at all. The most common model of social networking sites is based on presentation of the participant's profiles (in this case bloggers) and visualization of their network of relations with others [119]. However the type and visibility of personal information changes across different types of sites. For example, in some sites anyone can view any information and in others, such information may be restricted. On few sites users disclose large amount of information (like Facebook users where 90.8% profiles contain images, 87.8% of users reveal their birthdates, almost 40% publish their phone numbers etc) [119] and according to a study of 1.3 million bloggers at *LiveJournal.com* [182], 52% of the bloggers mention their age in profile, 68% of them express their at least one interest. Talking about risks of delivering private information to other, any personal information revealed can be exploited by many ways by anyone. The nature of privacy attack can be very severe if the information disclosed is extensive and intimate. In case of blogs, it becomes more important because of the sensitivity of topics being discussed. Online social networks sites users and consumer's privacy concerned authorities have been showing their concerns over this issue again and again [39, 313].

9.5.2   Absence of Data Collections

Another related challenge is the availability of data collection. Because of all above serious privacy concerns, research community is still deprived of such data collection that comprises all information of users' profiles with their blog posts. The only solutions left is extracting the users few details (like name, gender, age, location etc) from the web documents manually which is very hard task. Even we can find some work in past related to this problem but the percentage of accuracy is very low.

9.5.3   Language Complexities

Third problem is inherent to blogs i.e. language of blogs. Generally, the language within blogs is very informal i.e. no language grammar rules are followed. This makes the use Natural Language Processing (NLP) Techniques difficult in this case [224, 237]. Most common problems that are faced while processing blog data are:

- No capitalization

- No grammar rules

- Use of abbreviated words

- Spelling Mistakes

- Poor use of punctuations

- Ads

For example, one of the most commonly used short form is *WTH(i.e., what the hell)* represents the expression of unhappiness on author's part. One of the solutions is to prepare a list of such phrases that are shortened and then just replace them with their actual words [235] but today the use of such abbreviations is so long that this solution seems infeasible and also lot of overlapping occurs between different abbreviations that can lead to calculation of wrong semantics.

9.6   Time-Based Data Analysis

As described earlier that this work is in its early stages and therefore, lack any experimental results; however, in this section we demonstrate the importance of social

networking evidences for opinion detection task with a very simple example. For
this purpose, we use topic number 853 from the TREC Blog06 data collection which
is entitled as "State of the Union" relating to state of the union address of Presi-
dent Bush of year 2006. We identify the major entities (describing the issues) from
Wikipage of this topic[11] to perform a time-based analysis of the importance given to
these issue by general public. Keeping things simple, we choose frequency of an entity
as its importance criterion. Using the document polarity attribute, we can also see
in figure 9.6 that how people's opinion vary in its polarity as time progresses. This
change in opinion may depend upon time, events or demographic profile of the people
holding an opinion. Determining this cause behind the change in opinion can be a
very interesting task and we leave it for our future work.

For example, let's look at profiles of different entities for topic number 853 that
are basically different issues USA president talks about in his address like illegal
migration, same-sex marriage and energy crisis etc. This time-based profiling of issues
shows the level of importance given to these issues by general public and how this
opinion changes over time (or gender, region, and age etc., in case all this information
is known). Figures 9.5, 9.6, 9.7, and 9.8 show the temporal profiles for issues *Energy
Crisis* and *Hurricane Katrina* with respect to their importance (Y-axis) in different
documents (shown over X-axis). We have assumed the frequency of an entity the
criterion of its importance.
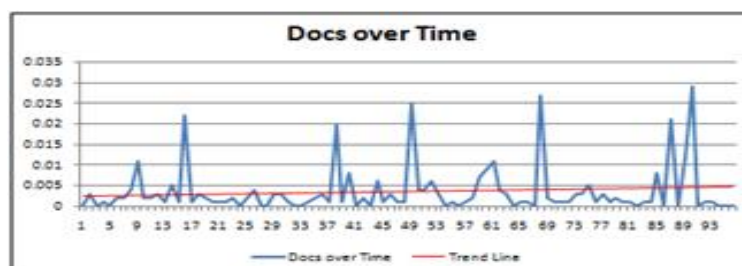
Figure 9.5: *Entity Profile over time- Energy Crisis*

A careful analysis of figures shown above reveals that people are more concerned
about issue of *Hurricane Katrina* than issue of *Energy Crisis*. In figure 9.7, it can be
seen that the topic of *Hurricane Katrina* is discussed more (with higher average of
importance values) for longer periods of days shown over X-axis. Discussion among
people about *Energy Crisis* starts just one day before the state of the union address
of President Bush when the print and electronic media might be reporting talk shows
discussing issues that president will talk about but other more social issues like *same-*

---

[11]http://en.wikipedia.org/wiki/2006_State_of_the_Union_Address

Figure 9.6: *Entity importance over time in all types (Positive, Negative, Neutral, Relative) of documents- Energy Crisis*
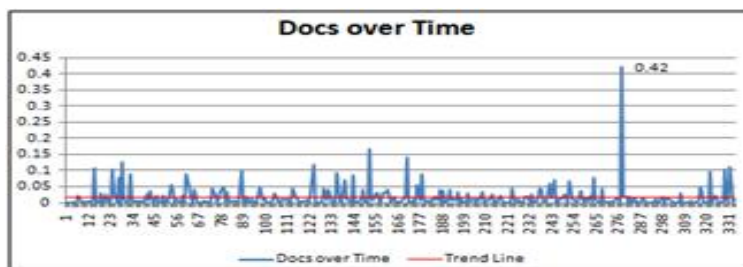
Figure 9.7: *Entity Profile over time- Hurrican Katrina*
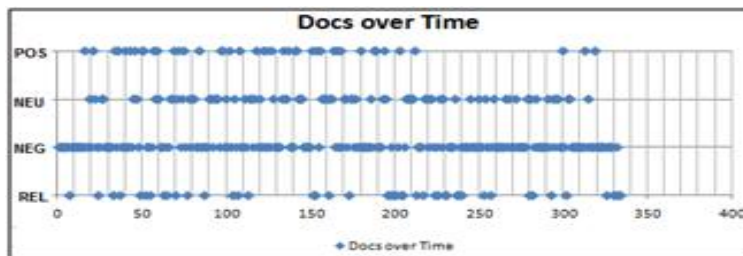
Figure 9.8: *Entity importance over time in all types (Positive, Negative, Neutral, Relative) of documents - Hurrican Katrina*

sex marriage, *illegal migration* were given more importance by general public and were more discussed in blogs.

The contribution of this chapter is basically our preliminary work for framework which exploits content and social evidences of blogosphere for opinion related tasks. We describe the infrastructure of this framework and a set of evidences for the tasks defined. This work is in preliminary stages and still requires mathematical modeling and experimentation results.

9.7   Chapter Summary

In this chapter, we have described the basic structure of blogosphere and discussed
many social networking features of blogosphere that can help researchers for the
tasks of opinion detection, opinion prediction and multidimensional ranking of blog
documents. Basic elements of blogosphere have been discussed with their possible
contributions in the proposed framework. This chapter also discusses the challenges
that researchers might face while realizing this framework. We have concluded that
research work in this area is limited because of these challenges. Because of unavail-
ability of data collection, no extensive experimentation has been performed. However
a demo performed over TREC Blog06 collection is presented. Formal modeling of
the framework proposed and extensive experimentation is kept as part of our future
work.

Section IV

Entity Ranking

Time-Aware Entity Retrieval

Opinion is the medium between knowledge and ignorance.

Plato

10.1   Introduction

Entity Retrieval (ER) is a recent search task which goes beyond the classic document search. It allows users to find more than just web pages (i.e., people books, movies, etc.). One of the promising application of ER in the commercial world is *news search*[1].

In this chapter, we address the problem of ranking entities in news applications. We introduce the original task of *Time-Aware Entity Retrieval (TAER)* which takes into account the evolution of entities over time in a news topic thread. To evaluate the effectiveness of systems performing such task, we develop an extension of the TREC-2004 Novelty corpus [320] by annotating relevance at the level of entities. We then develop features and ranking models for the original TAER task. Novelty retrieval studied the retrieval of novel information in documents.

This chapter is organized as follows. Section 10.2 describes the motivation and major findings of our work. Section 10.4 introduces the dataset we created for evaluating time-aware entity search and defines the task we address. Section 10.5 introduces and motivates several features extracted from documents and entity history in order to rank entities. Section 10.6 presents an experimental evaluation of the aforementioned features, and 10.7 describes an additional task, i.e., entity profiling, with some preliminary results. The chapter ends with a conclusions section.

---

[1] http://news.yahoo.com, http://news.bbc.co.uk, http://news.google.com

10.2   Motivation

News Retrieval has also been the focus of much attention in the IR research community (e.g., [94, 108]), but to our knowledge there have been no ER tasks defined for news. A possible application consists in enriching the user interface by placing retrieved entities next to the news article to user is currently looking at. One example mock-up interface is shown in Figure 10.1 where important entities are shown just next to the news article matching the user query. A study of possible methods for identifying such entities is the focus of our work presented in this chapter.
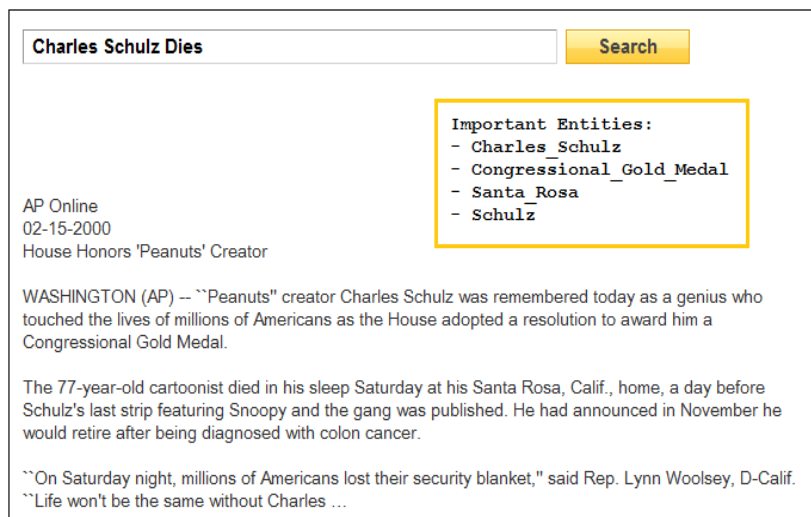
Figure 10.1: *A possible user interface for Entity Retrieval in news articles*

Dealing with ER in news is particularly interesting as news articles are often focused on entities such as people, companies, countries, etc. It is also a challenging task because unlike standard ER tasks there is the time dimension involved. Given a news topic, the decision about which entities should be retrieved or not changes with time. Not all frequently appearing entities should be considered relevant to the topic (e.g., news agencies) and new important entities may appear only later in the story (e.g., witness of a murder).

We propose an approach which takes into account both information from the current news articles as well as from the past relevant articles in order to detect the most important entities in the current news. Specifically, we design *local* and *history* features exploiting appearance of entities in the text.

Our main findings presented are:

- sentence novelty is worse than pure sentence relevance as an indicator of entity relevance.

- entities that become relevant have a high probability of remaining relevant the next article and the entire news thread.

- the relevant history of an article (e.g. the previous relevant articles) can be exploited as a source of information for TAER.

10.3   Research Problem Context

Entity Ranking has already been discussed in detail in chapter 5. However the different components (sentence retrieval in news, time-based IR, etc.) needed for this task are active areas of research in the IR community and are discussed below:

10.3.1   Novel Content Retrieval

With respect to the news domain, the TREC Novelty Track defined a task that takes into account the chronological order of documents. How to identify sentences that carry novel information with respect to previous retrieved content has been evaluated in [122, 320, 322]. The best performing approach (in terms of F-Measure) at the TREC Novelty Track 2004 used a variation of TF-IDF in order to detect new sentences [108]. It has been shown that exploiting the presence of entities can improve the effectiveness of novel sentence retrieval. Li and Croft presented an approach based on the presence of named entities [193–195]. Zhang and Tsai [403] employed named entity recognition and part-of-speech tagging to propose a mixed method for novel sentence retrieval. Compared to previous work on Novel Content Retrieval we aim at exploiting the time dimension of the collection in order to perform the different task of finding relevant entities in the documents.

10.3.2   Time-based Information Retrieval

Time-based Information Retrieval is an active related research area. The time dimension can be exploited in several search tasks [9]. Some authors [192**?**] have proposed an adaptation of language models to incorporates temporal expression in order to enable text search based on time. Diaz [94] proposed models for classifying whether web search queries are news-worthy or not over time. He used information

from previous clicks predicting the usefulness of displaying news in the result page at a given point as topics and interests develop over time. Alonso et al. [8] studied how the time dimension can enhance the presentation of query results. Berberich et al. [34] proposed an extension of the inverted index for temporal search (i.e., text search over temporally versioned document collections). Compared to previous work on Time-based Information Retrieval we focus on retrieving entities instead of documents.

10.4   Time-Aware Entity Retrieval

Consider the following user scenario: a user types a query (or topic) into a news search engine and obtains a list of relevant results, ordered by time. Furthermore, the user subscribes to this query so that she continues to receive the latest news on this query (or topic) in the future. We are interested in ER tasks related to this user scenario. Standard ER could be used to show to the user the most interesting entities *for the query*. The temporal dimension is not needed here.

However, if the user is observing a current document, it can be interesting to show the most relevant entities of the document for her query (or topic). This prompts a first task definition:

- Time-Aware Entity Retrieval (TAER): Given a query and a document relevant to it, and possibly a set of previous related documents (the *history* of the document), retrieve a set of entities that best summarize the document.

This is a newly defined task that can be useful, for example, in news verticals for presenting the user more than just a ranked list of documents. In the news context we define the task for most considered entity types: persons, locations, organizations, and products. More formally, we define a "news thread" relevant to a query as the list of relevant documents $D = [d_1 \ldots d_n]$. Then, given a document d_i we define its history as the list of relevant documents $H = [d_1 \ldots d_{i-1}]$ chronologically ordered pre-dating the document d_i. Given an entity e, we note as $d_{e,1}$ the first document in which the entity occurred in the news thread. Note that such a document is not necessarily the first document in D as entities may appear only in subsequent documents. Additionally, we will note as $d_{e,-1}$ as the last document in H which contains e.

10.4.1   A Dataset for Evaluating ER Over Time

The TREC Novelty Track in 2004 consisted on a collection of news articles and a set of topics for evaluating retrieval of novel information over ranked lists of documents for each topic. The systems had to retrieve information (i.e., sentences in this case) relevant to the topic and not yet present in the retrieved results [320]. That is, a novel sentence 1) is relevant to the topic and 2) contains new information compared to the sentences retrieved before it. A time-stamped list of documents is provided for every topic reflecting the temporal flow of the story the topic is about.

We selected the 25 'event' topics from the latest TREC Novelty collection (2004). We annotated the documents associated with those topics using state of the art NLP tools [14, 397] in order to extract entities of type person, location, organization, and product based on WSJ annotations. The annotation system detected 7481 entity occurrences in the collection: 26% persons, 10% locations, 57% organizations, and 7% products.

Six human judges assessed the relevance of the entities in each document with respect to the topic grading each entity on the 3-points scale: Relevant, Related, Not Relevant. An additional category, i.e., 'Not an entity', was used to mark entities which had been wrongly annotated by the NLP tool. A total of 21213 entity-document-topic judgments were obtained in the collection[2]. Examples of judged entities over two documents for a specific topic are shown in Table 10.1. The topic is about the event of the Peanuts' author death. Over the entire news thread some entities are relevant all the time (e.g., Schulz), some appears only after some time (e.g., his wife), and others are always present but with different relevance status (e.g., the city of Santa Rosa sometimes commemorates him and is therefore relevant, while other times it is just the place where the news has been written and is therefore not relevant). We can see entities of different types (e.g., persons, cities) and that some are highly relevant and stay relevant over different documents (e.g., Schulz). Other entities may change relevance status from related to relevant (e.g., Santa_Rosa) as they play a critical role in the current article, or to not relevant (e.g., Snoopy) as they are just mentioned in the current article. Moreover, some annotations do not represent named entities and are judged accordingly (e.g., center).

We performed double assessments on six topics in order to check the assessors' agreement obtaining an average Cohen's Kappa [185] of 0.5900. Looking at agreement rates

---

[2]The evaluation collection we have created is available for download at: `http://www.L3S.de/~demartini/deert/`

Table 10.1: *Example entities and their judgments in the first two articles for topic N79. Some entities keep their relevance status and others change it. Entities could appear only in some articles. Some annotations may not represent entities.*

Topic N79: Charles Schulz Dies		
	APW19991001.0198	APW19991027.0332
Schulz	Relevant	Relevant
Peanuts	Relevant	-
Charlie_Brown	Related	Related
Snoopy	Related	NotRelevant
Santa_Rosa	Related	Relevant
center	-	NotAnEntity

in other relevance judgment settings (0.49 on 40 topics at TREC 2006 Legal Track [26], 0.34 on 100 documents for opinion detection [259], 0.55 on sentence relevance at TREC 2004 Novelty Track [323]) we can see how entity relevance in a news corpus is less subjective than in other settings such as, for example, opinion detection.

10.4.2 Analysis of the Dataset

The collection we produced consists of an average of 31.2 relevant news articles per topic distributed over time. Each document in the collection contains on average 26.5 annotated entities among which 7.6 were judged relevant. On average each topic contains 63.4 entities which have been marked relevant at least once over the topic timeline.

We now investigate the relation between entities, sentence and relevance. Let n_s, r_s indicate that a sentence s is novel or relevant respectively. Let t_e indicate the type of entity e, and let us denote by r_e the fact that e is relevant, and $\bar{r}_e$ otherwise.

On average, a sentence contains 1.46 entities, a relevant sentence contains 1.88 entities, and a novel sentence contains 1.92 entities which indicates the presence of more information. The unconditional probability of a relevant entity in a sentence $P(r_e)$ is 0.411 (the sample is over sentences and then over entities in the sentence). The probability of finding a relevant entity in a relevant sentence $P(r_e|r_s)$ is 0.547 with a 95% bootstrap confidence interval of $[0.534 - 0.559]$, well above $P(r_e)$. The probability of a relevant entity in a novel sentence $P(r_e|n_s)$ is 0.510 $[0.491 - 0.531]$ which is below the probability in a relevant sentence.

This gives the following high level picture. Relevant sentences contain slightly more entities than non-relevant ones. Novel sentences contain slightly more entities than relevant (but not-novel) sentences; however, entities is novel sentences are more likely

to be irrelevant than in non-novel sentences.

In Table 10.2 we look at relevance probabilities per entity type (e.g., the probability of person entity being relevant would be noted $P(r_e|t_e = person)$). We show again that sentence novelty is less important than sentence relevance *regardless of the entity type*. Organization entities are more likely in a relevant sentences than the rest (63% of those appearing in relevant sentences have been marked relevant).

Table 10.2: *Probabilities of relevance for different entity types with 95% confidence intervals.*

$P(r_e|t_e = person)$	0.406 [0.391-0.421]
$P(r_e|t_e = person, r_s)$	**0.560** [0.533-0.588]
$P(r_e|t_e = person, n_s)$	0.496 [0.451-0.541]
$P(r_e|t_e = organization)$	0.479 [0.471-0.487]
$P(r_e|t_e = organization, , r_s)$	**0.631** [0.616-0.646]
$P(r_e|t_e = organization, n_s)$	0.587 [0.564-0.612]
$P(r_e|t_e = product)$	0.179 [0.164-0.194]
$P(r_e|t_e = product, r_s)$	**0.237** [0.210-0.265]
$P(r_e|t_e = product, n_s)$	0.189 [0.151-0.228]
$P(r_e|t_e = location)$	0.284 [0.271-0.297]
$P(r_e|t_e = location, r_s)$	**0.403** [0.379-0.427]
$P(r_e|t_e = location, n_s)$	0.397 [0.363-0.432]

With respect to pairs of entities co-occurring in the same sentence, we see that the probability that both entities have been assigned the same relevance judgment is 0.42. The probability for each possible pair is presented in Table 10.3. It worth noting that the most probable event is that two entities co-occurring in a sentence are both relevant. This result shows how entity co-occurrence might be a good indication for finding relevant entities.

$P(e_1, e_2)$	Relevant	Related	NotRel	NotAnEntity
Relevant	0.24	0.08	0.03	0.07
Related	0.08	0.07	0.03	0.03
NotRel	0.03	0.03	0.07	0.05
NotAnEntity	0.07	0.03	0.05	0.04

Table 10.3: *Probabilities of relevance for entities co-occurring in a sentence*

As compared to a classic document collection, in a news corpus the time dimension is an additional available feature. How useful is the information from past news articles? The probability of an entity being relevant in a document given that it was relevant the first time it appeared ($d_{e,1}$) is 0.893 [0.881 − 0.905] which shows how in most cases an entity which is relevant at the beginning of its appearance stays relevant for the rest of the news thread. It is also important to observe just the previous document

where the entity appeared. The probability of an entity being relevant in a document given that it was relevant the previous time it appeared is 0.701 [0.677 − 0.726]. Conversely, the probability of a relevant entity changing relevance status form one story to the next is 0.3. Another characterization of this is the probability of an entity being relevant in a document given that it was relevant in the i-th document of its history. This is shown in Figure 10.2 for relevant, related and not-relevant entities. We can see that relevant entities are the most stable over time while related entities tend to change relevance status over time (either to relevant or to not-relevant).
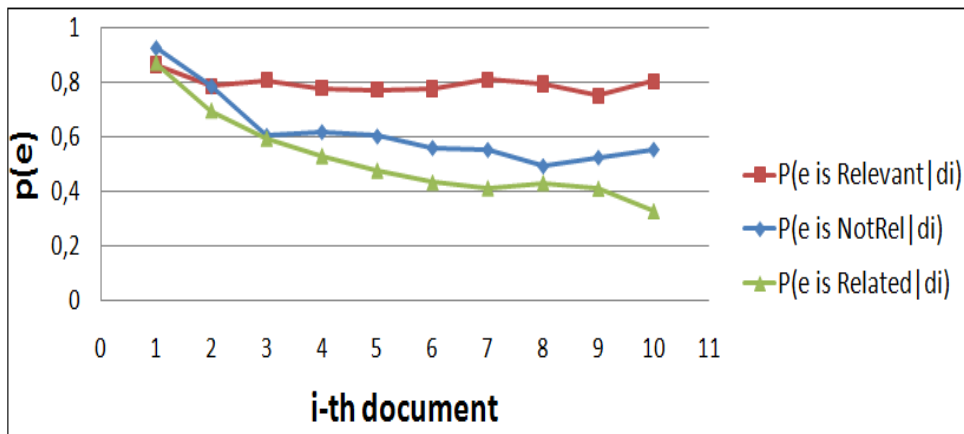
Figure 10.2: *Probabilities of entity relevance given its relevance in the i-th document*

10.5 Models for Time-Aware Entity Retrieval

In the TAER task we are given a query q and we want a ranking function that sorts the set of entities e_i occurring in document d according to their relevance. As we have seen in Section 10.4.2, relevant entities often appear in relevant sentences. More interestingly, the past articles seem to be a good evidence for entity relevance. For such reasons in the following we present a set of features that can help ranking entities in news articles both considering attributes from the current article as well as from the news published in the past on the same topic.

10.5.1 Local Features

As we aim at retrieving entities described in a document d, the first thing to do is to exploiting entity occurrences in d. The first feature we consider is the frequency of an entity e in a document d, noted $F(e, d)$. In the following we will use this

feature as our baseline. As we have seen in Section 10.4.2, relevant sentences contain more relevant entities. Therefore, a natural extension of the baseline is obtained considering the relevance score of the sentences where e appears with respect to q. We computed the BM25 scores [300] of sentences with respect to a disjunctive query consisting of all the terms in the topic title. We can therefore rank entities according to the average or the sum of BM25 score of the sentences where e appears in d (noted $AvgBM25s(e,d)$ and $SumBM25s(e,d)$ respectively). Key entities are often those performing certain actions in a news story. After running a dependency parsing over the sentence collection, it is possible to consider if an entity appears as a subject of a sentence as this is generally the person or thing carrying out an action. Hence, we define the $F_{subj}(e,d)$ as the number of times an entity e appears as subject of a sentence in the document d.

In the news writing style it is a common practice to summarize the story at the beginning of the article and provide more details in the following. Thus, we expect to find key entities toward the beginning and less important entities afterwards. We additionally propose two position-based features that take into account where in document d an entity e appears. Let $FirstSenLen(e,d)$ be the length of the first sentence where e appears in document d and $FirstSenPos(e,d)$ be the position of the first sentence where e appears in d (e.g, the fourth sentence in the document).

10.5.2 History Features

We now introduce a number of features that take into consideration the document history H. As defined for the current document, we can obtain a simple feature just by counting the occurrences of an entity in the past. Let $F(e,H)$ be the frequency (i.e., the number of times it appears) of the entity e in the history H.

As documents may have different length and, thus, contain more or less entities, it is possible to refine the previous feature taking this into account. Instead of counting each entity occurrence a simpler variation considers the number of documents in which the entity e has appeared so far. We thus define $DF(e,H)$ as the document frequency of e in H.

More than just looking at the entire set of past documents we can also consider specific documents. When a news story begin the key entities are already present. We thus define $F(e,d_{e,1})$ as the frequency of entity e in the first document where the entity appeared. As we have seen in Section 10.4.2, the previous document is also an important evidence of entity relevance. We define $F(e,d_{e,-1})$ as the frequency of entity e in the previous document where the entity appeared.

In news stories important entities interact with many other ones. We can compute $CoOcc(e, H)$, the number of other entities with which the entity co-occurred in a sentence in the set of past documents H. Finally, we leave the study of the influence of recency in the effectiveness of these features (i.e., do more recent documents provide better evidence as compared to older ones?) for future work.

We can have an initial analysis of such features by checking how entity relevance probability changes with the features value. Figure 10.3 shows the probability of an entity being relevant given different values of the features described above. We see that all are correlated with relevance over their entire domain.
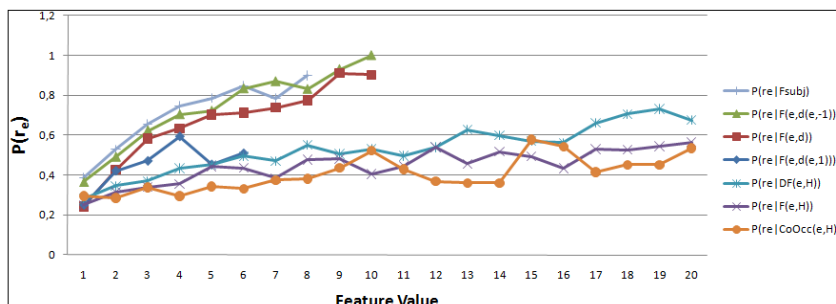
Figure 10.3: *Probability of an entity being relevant given different feature values for several features*

10.6 Experimental Evaluation

In this section we present the experimental evaluation of the features proposed for the TAER task.

We compare the effectiveness of different features and some feature combinations using several performance metrics. In order to evaluate the complete entity ranking produced by the proposed features, we compute Mean Average Precision (MAP). For completeness, as we aim at showing the user few entities, we check for early precision as well. We report values for Precision@3 (P@3), Precision@5 (P@5), and we test for statistical significance using the t-test. Because there were defined three levels of relevance when evaluating the test collection, we need to fix a threshold for binarising the relevance. In the following we consider related entities as non-relevant. As future work we will study effectiveness of our approach on Related entities. Many of the features we use are based on entity frequency, hence entity scores in the ranking will have many ties. For this reason, the evaluation measures we have computed are aware of ties, that is, they consider the average value of the measure for all possible

combinations of tied scores [225].

10.6.1 Evaluation of single features

Local Features Table 10.4 shows effectiveness values obtained when ranking entities in a document according to local features, where no single feature performs better than the simple frequency of entities in the document. For comparison, a feature that assigns the same score to each entity would obtain a MAP value of 0.42 with a ties-aware measure. The feature $F(e, d)$ obtains the best MAP value (0.60). The second best local features is $SumBM25s$ (0.52 MAP) which takes into consideration relevance of sentences where the entity appears. On the other hand, the features looking at the first sentence where the entity appears in the news article (FirstSenLen, FirstSenPos) do not perform well (0.45 and 0.43 MAP respectively). In order to ex-

Table 10.4: *Effectiveness of local features for TAER.*

Local Features	P@3	P@5	MAP
All Ties	.34	.34	.42
F(e,d)	**.65**	**.56**	**.60**
FirstSenLen	.37	.36	.45
FirstSenPos	.31	.31	.43
F_{subj}	.49	.44	.50
AvgBM25s	.27	.30	.41
SumBM25s	.50	.44	.52

ploit the position of the first sentence where an entity appears we need to deal with the problem of headers in news articles (e.g., news agency codes): as articles have different header lengths, it is not easy to detect the beginning of the article body. Additionally, three different news agencies contributed articles to the collection each of them having different formatting standards. For example, the agency NYT can have articles where the title and body do not start before the tenth sentence while for others (e.g., XIE) the interesting content can start already at the third sentence. The transformations of FirstSenPos that we explored did not improve performances of this feature.

History Features Table 10.5 presents the performance of TAER using history features. In general, history features perform better than local features and the highest performance is obtained by ranking entities according to its frequency in the past documents ($F(e, H)$). All history features but $F(e, d_{e,1})$ significantly improved over the baseline in terms of MAP. In terms of early precision (P@5) only $F(e, H)$

and the similar feature $DF(e, H)$ improve over the baseline. Moreover, features using the entire history H are performing better than features looking at single documents in the past.

It is also interesting to note that, when identifying relevant entities for a document, the frequency of the entity in the previous document in the story $F(e, d_{e,-1})$ is a better evidence than the frequency in the current document. This may be an indication of how people read news: some entities become relevant to readers after repeated occurrences. If an entity appears in the current and previous documents it is more likely to be relevant.

We additionally weighted the scores obtained from different documents in H with both the document length and BM25 score of the document with respect to the query. This approach did not improve the effectiveness of the original features without per-document weighting.

Table 10.5: *Effectiveness of history features for TAER and improvement over $F(e, d)$. In brackets the % improvement over $F(e, d)$. *(**) indicates statistical significance w.r.t. $F(e, d)$ with $p<0.05(0.01)$*

History	P@3	P@5	MAP
$F(e, d)$	.65	.56	.60
$F(e, d_{e,1})$	.58 (-11%)	.53 (-6%)	.56 (-7%)
$F(e, d_{e,-1})$	.64 (-2%)	.56 $(\pm0\%)$	.62* $(+3\%)$
$DF(e, H)$	.63 (-3%)	.57* $(+1\%)$	.65** $(+8\%)$
$F(e, H)$	**.66** $(+1\%)$	**.59**** $(+5\%)$	**.66**** $(+10\%)$
$CoOcc(e, H)$	.62 (-5%)	.57 $(+1\%)$	.65** $(+8\%)$

Given these results we conclude that the evidence from the past is very important for ranking entities appearing in a document. Thus, we expect effectiveness of methods that exploit the past to improve as the size of H grows. That is, the more history is available the better we can rank entities for the current news.

The y-axis of Figure 10.4 plots the average MAP for all the documents with history size $|H|$ using the feature $F(e, H)$.

For $|H| < 20$ the effectiveness of $F(e, H)$ increases together with $|H|$ up to values of 0.7. Results for higher values of $|H|$ show no clear trend due to the fact that there are just a few datapoints.

Influence of non-relevant documents. TREC 2004 Novelty Track topics also contained 157 irrelevant documents which are close matches to relevant ones (6.28 on average per topic) [320]. We checked the correlation between the performance of $F(e, d)$ and the number of irrelevant documents present in the topic. Pearson's correlation coefficient
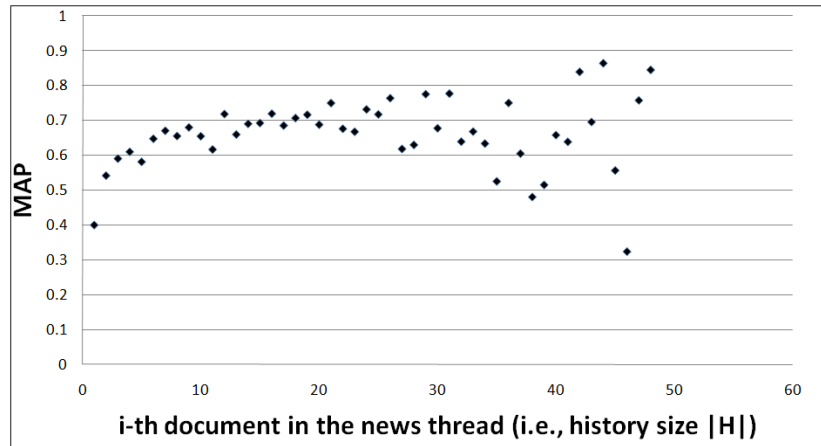
Figure 10.4: *Mean Average Precision values for documents having a certain history size*

between the number of irrelevant documents and AvgPrec is -0.234.

10.6.2 Feature combination

So far we have presented different features for ranking entities that appear in a document. Combining them in an appropriate manner yields a better ranking of entities; however, because the distribution of relevance probability is different among features, we need a way for combining them. The following experiments rank entities in a document according to a score obtained after combining several features together. We consider linear combination of features (transformed with a function as explained in [79]). Finally, we will consider a combination of all the features using machine learning techniques.

Linear Combination of Features   Let the score for an entity e and a vector $\vec{f}$ of n features be

$$score(e, \vec{f}) = \sum_{i=1}^{n} w_i g(f_i, \theta_i) \, , \qquad (10.1)$$

where w_i is the weight of each feature and g is a transformation function for the feature f_i using a given parameter θ_i. Since we are only interested in the ranking we can eliminate one weight parameter by fixing $w_1 = 1$ [79]. In this chapter we employ a transformation function of the form:

$$g(x, \theta) = \frac{x}{x + \theta} \qquad (10.2)$$

as suggested in [79], where x is the feature to transform and θ is a parameter. We also tried a linear transformation but it did not perform as well. More complex non-linear transformations could also be explored.

In order to combine features we then need to find a parameter θ_i for the function g and a weight w_i for each feature f_i. In Figure 10.5 we show how some of the functions we employed fit the distribution of probability for different features. The probability values are normalized in a way that the plot starts from the point $(x = 1, y = 1)$. The same is done for the g function using a multiplicative constant $z = (1 + k)$.

(a) (b) (c)
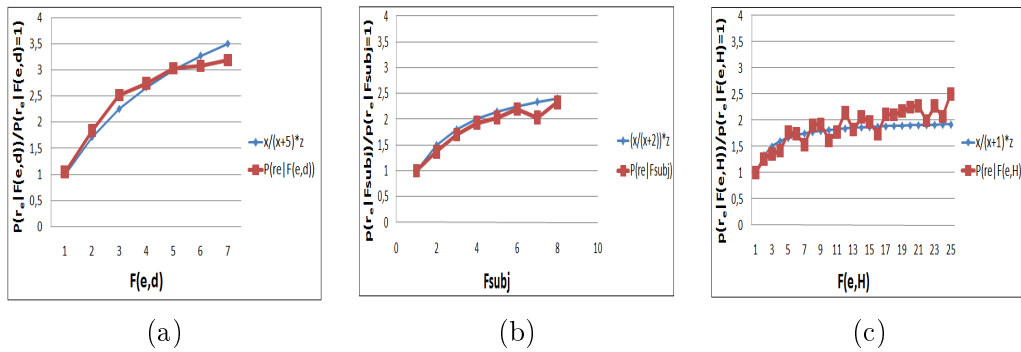
Figure 10.5: *Normalized probabilities of an entity being relevant for a given feature value and the selected g function normalized with a constant z.*
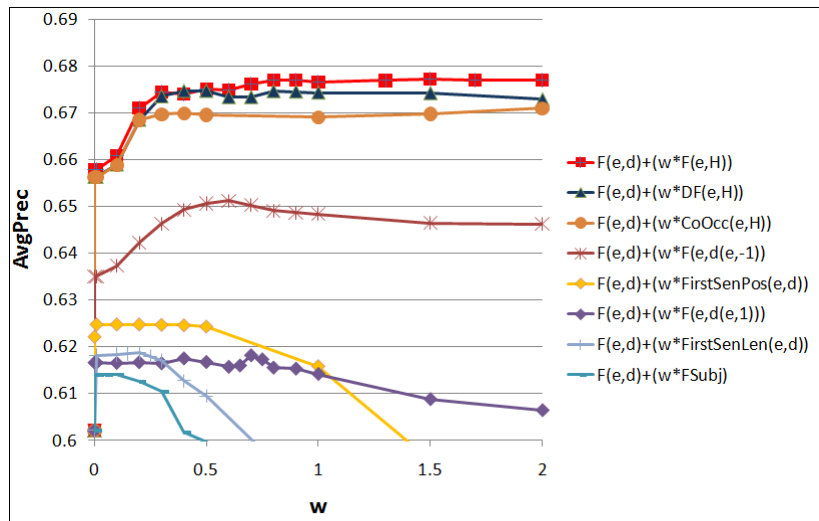
Figure 10.6: *Mean Average Precision values for different values of w when combining features with $F(e, d)$*

We tested two and three features combinations, where the variables θ_i, and the combination weights w_i have been tuned with 2-fold cross validation of 25 topics training

to optimize MAP. In order to find the best values we used an optimization algorithm that performs a greedy search over the parameter space [292]. Figure 10.6 presents MAP obtained for different values of w_1 when different features are combined with $F(e, d)$. In some cases the combination performs better than the original baseline with the best performing features being robust to all values of w. Features from the local document such as F_{subj}, $FirstSenLen$, and $FirstSenPos$ show performance improvements only for small combination weights whilst features from the history have a higher robustness to high values of w. The two features that looks at individual documents in the history ($F(e, d_{e,-1})$ and $F(e, d_{e,1})$) decrease their performance as w increases. On the other hand, features looking at the entire set of past documents H are most robust.

Table 10.8 summarizes the results for all the features using 2-fold cross validation. Combining $F(e, d)$ with another feature is able to outperform the baseline for some range of the weight w that can be learned on a training set. For some features ($AvgBM25s$, $SumBM25s$) the original baseline score is not improved by the combination. The best effectiveness is obtained when combining $F(e, d)$ and $F(e, H)$ obtaining an improvement of 13% in terms of mean average precision. Other features, when combined with the baseline, also obtain high improvements performing as good as the combination with $F(e, H)$ ($CoOcc(e, H)$ having 12% and $DF(e, H)$ having 13% improvement in terms of MAP). The feature $F(e, d_{e,1})$, which is poorly performing as individual feature (see Table 10.5), obtains a limited improvement of 2% in terms of MAP. These results also hold for early precision measures.

In order to combine three features we need to find suitable values for two different weights w_1 and w_2 (we tune parameters and report the performance using 2-fold cross validation). The results for two different combinations of features with $F(e, d)$ are presented in Table 10.6. Results show that combining the baseline with two features from the history we can reach an improvement of 15% in terms of MAP over the baseline.

Table 10.6: *Effectiveness of two features combined with $F(e, d)$.* (**) indicates statistical significance w.r.t. $F(e, d)$ with $p < 0.05(0.01)$. †(††) indicates statistical significance w.r.t. $F(e, H)$ with $p < 0.05(0.01)$.*

f_1, f_2	P@3	P@5	MAP
$F(e, d_{e,-1})$ $F(e, H)$	$\mathbf{.70}^{**\dagger\dagger}(+8\%)$	$\mathbf{.62}^{**\dagger\dagger}(+11\%)$	$\mathbf{.69}^{**\dagger\dagger}(+15\%)$
$CoOcc(e, H)$ $F(e, H)$	$.69^{**\dagger\dagger}(+6\%)$	$.62^{**\dagger\dagger}(+11\%)$	$.68^{**\dagger\dagger}(+13\%)$

Using Machine Learning for combining features. In order to combine two or more features together we used machine learning techniques. We performed 2-fold cross validation training a multinomial logistic regression model with a ridge estimator [186] with default parameters for ranking entities in each document. Results show that when combining any feature with $F(e, d)$ using logistic regression the results are comparable to those obtained with manual tuning (Table 10.8).

Table 10.7 presents a combination of every local and history feature. The com-

Table 10.7: *Effectiveness of two features combined with $F(e, d)$ using logistic regression. The list of features in presented in Tables 10.4 and 10.5. In brackets the % improvement over $F(e, d)$. * (**) indicates statistical significance w.r.t. $F(e, d)$ with $p < 0.05(0.01)$. †(††) indicates statistical significance w.r.t. $F(e, H)$ with $p < 0.05(0.01)$.*

Features	P@3		P@5		MAP	
Local	.65	$(\pm 0\%)$	.58*	$(+4\%)$	.63**	$(+5\%)$
History	.65	$(\pm 0\%)$	.60**	$(+7\%)$	.66**	$(+10\%)$
All	**.70****††	$(+8\%)$	**.63****††	$(+12\%)$	**.69****††	$(+15\%)$

Table 10.8: *Effectiveness of features when combined with $F(e, d)$. Bold values indicate the best performing run. In brackets the % improvement over $F(e, d)$. * (**) indicates statistical significance w.r.t. $F(e, d)$ with $p<0.05(0.01)$. †(††) indicates statistical significance w.r.t. $F(e, H)$ with $p<0.05(0.01)$*

Feature	P@3		P@5		MAP	
$FirstSenLen$	.65	$(\pm 0\%)$	.57*	$(+2\%)$	.62**	$(+3\%)$
$FirstSenPos$	**.67****	$(+3\%)$	.58*	$(+4\%)$	.62**	$(+3\%)$
$FirstSenPosTrans$	.67**	$(+3\%)$	**.58****	$(+4\%)$	**.64****	$(+7\%)$
F_{subj}	.65	$(\pm 0\%)$	.56	$(\pm 0\%)$	.61	$(+2\%)$
$AvgBM25s$	.65	$(\pm 0\%)$	.56	$(\pm 0\%)$	.60	$(\pm 0\%)$
$SumBM25s$	.65	$(\pm 0\%)$	.56	$(\pm 0\%)$	.60	$(\pm 0\%)$
$F(e, d_{e,1})$	.65	$(\pm 0\%)$	.57**	$(+2\%)$	.61**	$(+2\%)$
$F(e, d_{e,-1})$	.68**†	$(+5\%)$	.60**	$(+7\%)$	.65**	$(+8\%)$
$F(e, H)$	**.70****††	$(+8\%)$	**.62****††	$(+11\%)$	**.68****††	$(+13\%)$
$CoOcc(e, H)$	.68**††	$(+5\%)$	.61**††	$(+9\%)$	.67**††	$(+12\%)$
$DF(e, H)$	.69**††	$(+6\%)$	.61**††	$(+9\%)$	.68**††	$(+13\%)$

bination of local features performs better then the baseline and then most of the single local features (see Table 10.4). Finally, when all the features are combined (local+history) we obtain the best effectiveness which is anyway not better than the combination of the three best features (i.e., $F(e, d)$, $F(e, d-1)$, and $F(e, H)$ see Table 10.6). Such improvements are anyway negligible if compared with the best 2 features combination, that is, $F(e, d)$ and $F(e, H)$ obtaining a MAP of 0.68. Therefore, we can see how these two simple features perform very well and that it is difficult to

improve over such approach.

10.7 Building Entity Profiles

In this section we present an initial discussion about an additional task (i.e., the Entity Profiling task) providing some statistics on the test collection we have built.

In a search interface, we may wish to show to the user relevant entities in the entire document history and not just entities from the current document. This prompts a second task definition:

- Entity Profiling (EP): Given a query and the set of related documents, create for each entity a plot showing the user the temporal development of entity relevance (i.e., which entities become relevant and which become not relevant over time). This is related to new user interfaces being proposed in commercial systems[3] and can help the user understanding which are the key entities in the story even if they do not appear in the news article she is reading.

Given that for a single event (a topic in the TREC collection) there are many entities (31 documents per topic and 27 entities per document) appearing, the question is for which entities should we build and show a profile? Figure 10.7 presents the distribution of document frequencies for entities, where 67% of entities appear only in one document. For such entities it does not make sense to build a time-based profile as there is no evolution of their relevance.
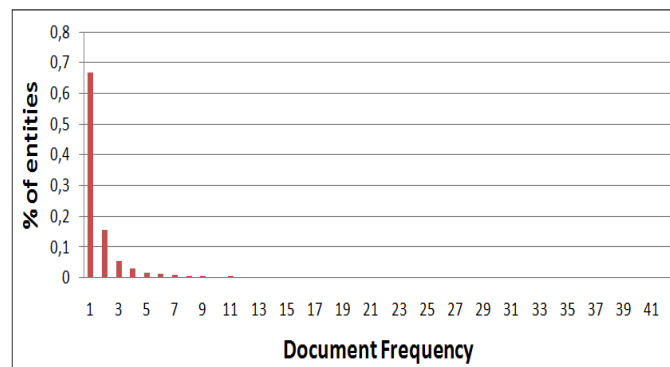
Figure 10.7: *Entities with given document frequency in the topic*

As we have already stated, relevant entities tend to stay relevant. It is therefore also

---

not interesting for the user to see entity profiles which are flat, that is that do not change over the story line. In Figure 10.8 we can see that half of the entities which are relevant at least once are relevant any time they appear.
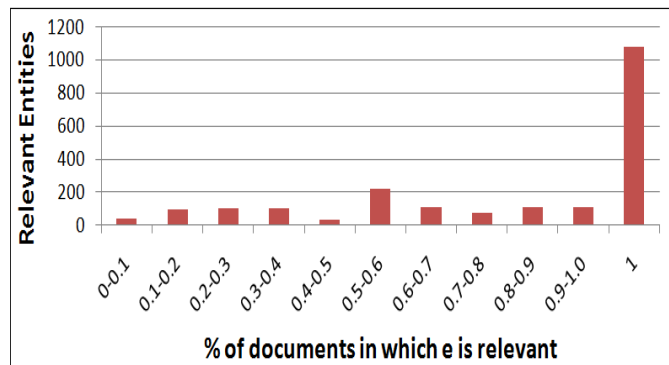
Figure 10.8: *Duration of relevance for entities relevant at least once*

We would therefore build entity profiles on entities which are relevant at least once and which do not have always the same judgment. There are 708 entities like this over the 25 topics in the collection.

A system has then to decide for each entity and each day when it appears in news whether to trigger ("ON") the entity or not ("OFF"). In order to evaluate such system we can define a true positive as the situation where the entity is relevant and the system returns "ON", true negative when the entity is non-relevant and the system returns "OFF", false positive when the entity is non-relevant and the system returns "ON", and false negative when the entity is relevant and the system returns "OFF". A system answering always ON will then get Precision of 0.56 and Recall of 1. A system exploiting the history and answering ON for entities appearing more than once in the current document or having $F(e, H) > (|H|/t)$ would get Precision of 0.60 and Recall of 0.54 for $t = 5$. Such result shows that a simple baseline performs well and that there is a Precision/Recall tradeoff.

Focus of our future work will also be an alternative to the Entity Profiling task. We imagine queries of the type: *Will entity e be relevant in the future?*. The task can be defined as predicting appearance (and relevance) of an entity e in future documents given that 1) e has appeared in the past (as relevant) and 2) e does not appear today. Analyzing relevance assessments we can see that 7% of entities appear at least twice as relevant with a gap (i.e., they do not appear in a particular day) in between. Being able to predict entity relevance would enable retrieval systems to extend the produced TAER result set including entities which are not present in the current news article which are anyway important for the overall story.

10.8   Chapter Summary

In this chapter, we have addressed the problem of entity search and ranking in news streams. For this purpose, we defined an original entity search task and further created a time-stamped test collection for evaluating it. We have tested several combinations of proposed features obtaining an overall statistically significant improvement of 15% in terms of mean average precision over the baseline that considers the frequency of entities in the document.

10.8.1   Findings

- Relevant sentences are more useful to determine the relevancy of an entity than novel sentences. In fact novel sentences introduce more entities than non-novel sentences, but many of these are not relevant. This is a counter intuitive finding, which challenges our view of novel sentences as introducing relevant entities. Our interpretation is that when an entity is first introduced (in a relevant and novel sentence), the reader cannot yet decide if the entity is truly relevant or not; only after repeated occurrences does the entity become relevant to the reader.

- We have proposed features both from the current document and from previous ones in the document's history in order to find relevant entities in a given document. We have experimentally shown that past frequency of entities is the most important of the features explored so far, more important than entity frequency in the current document another important feature.

- Position of the entity in the document (e.g., its first occurrence) is a weak indicator of its relevance, and it is specially difficult to use due to the different headers and introduction sentences present in different sources.

- Relevant entities (RV) tend to keep their same status over timeline of news articles while related (REL) and not related (NR) change their relevance status with the passage of time.

10.8.2   What needs to be improved?

Additionally, we have provided some preliminary observation on the Entity Profiling task concluding that an important challenge is the selection criteria of entities for which to build such profiles. As future work, besides testing our features on different time-aware document collections, we aim to develop and evaluate techniques for the Entity Profiling task. In addition to this, our future plans involve the extraction of opinions about the relevant entities of a topic.

Conclusions and Directions for Future Work

The greatest deception men suffer
is from their own opinions.
Leonardo da Vinci

11.1 Conclusions

The work presented in this thesis focused on different problems of the field of opinion mining and proposed approaches to solve few of them. Basically, our contribution include: 1) Finding opinion-topic associations in documents, 2) Analyzing role of topic (or domain) dependent and topic-independent evidences for opinion finding, 3) Identifying and ranking relevant entities to find opinions about them 4) Proposing a framework for opinion mining in the blogosphere by exploiting the social network based evidences.

1. Our first contribution corresponds to a major challenge of the field of opinion mining which is to find opinion-topic associations (OTA) in documents. The basic task is to associate correct opinions to corresponding topic-related textual segments. We have developed two kinds of approaches for this task, first on sentence level and second for passage-level. Both approaches take support of bi-dimensional query expansion. In our sentence-level approach [238], we exploited

semantic-relationships of WordNet to determine the degree of opinion-topic associations in each sentence of a document by matching sentence and query words. This sentence-query matching is done using *Path* and *Lesk* measures of WordNet. As a result of this matching, each sentence is assigned an OTA score. OTA score of a document is function of OTA scores of all its sentences. OTA score of a document is combined with other document level features like reflexivity, addressibility, and subjectivity, etc. Experimentation was performed with TREC Blog 2006 test data collection and results showed an improvement of almost 29% over its baseline. The results showed the effectiveness of these basic heuristic features and WordNet's semantic relations for the task of opinion finding.

The conclusion drawn from this experimental work is that basic heuristics like features and semantic relations of WordNet could be useful for opinion finding task. However, we also conclude that the performance could be further improved provided all features are well formulated and semantic relations used are precise and reliable. The semantic relatedness measures used in our work i.e., Path and Lesk have performed well but Lesk measure seems unreliable because of the way it is computed. We think that a more refined Lesk score could be useful for finding more precise similarity scores.

Further experimentation on passage-level in [238] motivated us to use passages as basic processing unit for the task of opinion finding. Therefore in our second contribution for finding opinion-topic association, we adapted a passage-based language modeling (LM) approach for opinion finding task. It has been observed that almost all of the opinion finding approaches depend heavily on the opinion score of the opinionated terms or we can say that their basic model revolves around the opinion score of the terms. This fact triggers us to propose an approach which estimates the language model of a passage by using the opinion score of the terms and the results proved the effectiveness of this approach. We obtained significant improvement over the strongest baseline of TREC Blog track (baseline-4) and beat the previous best results reported for TREC Blog 2006 topics. The results of our approaches show that passage-level processing is better for finding opinion-topic associations in documents. However, it should be noted that our language modeling approach is supported by bi-dimensional query expansion.

2. Topic-dependent opinion finding approaches are effective in performance but their performance varies from domain to domain [284]. On the other hand,

topic-independent opinion-finding approaches retain their generalization but generally such approaches suffer from lower performance. Therefore an ideal opinion-finding approach is the one which is both generalized and effective (i.e., which combines both approaches). In this thesis, we have proposed a mixed approach which combines topic-independent features (like reflexivity, addressibility, emotiveness, subjectivity, etc.) with topic-dependent features (like relevance rank and relevance score). Our approach proved its effectiveness by improving all TREC provided baselines and also improved the previous results for baseline-3 and baseline-4 for topics of year 2007. From experimental results of this approach, we can conclude that even simple topic-independent and topic-dependent features can perform well if they are formulated well and experimented with good combination of features. In this work, we also found out that combination of adverbs and adjectives could be better to compute the subjectivity feature proposed for opinion finding task.

As we discussed already that there are some approaches that have already used this type of heuristics-based evidences but they did not perform equally well as our approach did. The difference lies in the way we computed these evidences and the way they were combined. There are many approaches that have used many complex techniques for the task of opinion finding. Proposing a complex approach does not assure an outstanding performance. In contrary, it adds further burden to the processing of huge data collections. In this situation, an ideal solution is to have very simple set of features that are well formulated and are combined using some effective technique to perform well. Apart from simplicity, our approach also keeps its generalization intact. These characteristics of our approach with its good performance are enough to prove the effectiveness of our approach.

3. Blogosphere is not only a rich source of opinions but also its networked structure enriches it with many social network evidences that can be exploited for extracting or predicting opinions from blogs. There exist few works who have exploited this networked structure of blogosphere for opinion-related tasks but they are more restricted to identification of the most influential and opinionated blogs within it [142, 161, 325]. Using blogosphere's social evidences for opinion finding and opinion prediction tasks remains an open challenge. Our contribution for thesis includes a preliminary work in this regard. We have proposed a framework which exploits content and social structure of the blogosphere to perform the tasks of opinion detection, sentiment prediction and multidimensional

ranking. Our framework uses trust and quality measures (besides content-based evidences) for tasks defined. These features (like trust, quality, freshness, etc.) are already being used in IR domain and have proven their effectiveness for several tasks.

4. Entity-based opinion detection is relatively new subject in opinion detection. It deals with the task of finding opinions about all entities relevant to the issue as expressed in the given query. Generally, this task is performed in three steps: 1) finding relevant entities for the given query, 2) ranking the identified relevant entities according to their relevance, and 3) finding the opinions related to those entities. However, identifying a set of relevant entities is a big challenge for this task. In this thesis, we proposed a very effective approach for ranking entities in a news corpus which gets a significant improvement over its baseline. We prepared a novel entity labeled corpus with temporal information out of the TREC 2004 Novelty collection. We have experimentally shown that past frequency of entities is the most important of the features explored so far, more important than entity frequency in the current document another important feature. Position of the entity in the document (e.g., its first occurrence) is a weak indicator of its relevance, and it is especially difficult to use due to the different headers and introduction sentences present in different sources.

11.2 Future Work

11.2.1 Using Passage Coherency

To further improve the results for passage-based opinion detection, we would like to analyze the role of internal and external coherences of passages in a document in our future work. Internal coherency of a passage means that each sentence in that passage must lead up logically to the next sentence. Let us take an example of a passage given below to explain the idea of internal coherency of a passage.

*One reason that I feel ice cream should be banned is that ice cream **contains too many calories. Excessive calories** lead to **heart disease. Heart disease** is the most common killer among Americans. Thus, ice cream should indeed be banned*

In the above example, it can be observed that how each sentence's controlling idea leads up logically to the next sentence. Similarly external coherence of passages means that a passage is logically connected to the next passage in a document. We plan to measure these coherences from two different perspectives (i.e., topical coherence

and opinionated coherence). Topical coherence will look for how a passage maintains its control over the topic internally or how a document maintains it from passage to passage i.e. (external passage coherence). Similarly opinionated coherence will look for control in its opinion. We hope that passage coherency can play its role to improve results for opinion detection. Also it could be very helpful to analyze the variation of opinions from individual to individual who are posting their comments on a particular blogpost.

11.2.2 Entity-Based Opinion Detection

Bing Liu, Professor Dept. of Computer Science University of Illinois at Chicago and a famous researcher in the field of opinion detection, said in an interview[1] to *textAnalyticsNews.com*, on April 20, 2009:

Sentiment analysis is not simply the problem of determining whether a document, a paragraph or even a sentence expresses a positive or negative sentiment or opinion. It is also about entities. Without such information, any sentiment is of little practical use. So one should not only talk about sentiment analysis of documents, paragraphs or sentences, but also about the entities that sentiments have been expressed upon. Here an entity can be a product, service, person, organization, event or topic.

Public opinion holds lot of importance in a civil society. It is not only responsible for opinion changes in a society or a country but it affects the international events and policies. Public opinion changes with time and events. Therefore, it becomes very important to analyze change in public opinion about a certain entity with respect to time, related events and other related entities. Other interesting related tasks can be to forecast the change in public opinion about a certain entity at a certain time, occurrences of events like public manifestations or likely winners of a football tournament, entrance of a new related entity in temporal profile of an entity, etc. As our future work, we hope to extend our entity ranking work to propose an approach for the tasks as mentioned.

11.2.3 Domain-based Opinion Vocabulary Analysis

The opinion vocabulary of an author change with the topic, author's expertise of the langauge he is writing in, author's age, author's background, etc. As our future work, we would like to analyze the change in opinion vocabulary with respect to topic.

---

[1]http://social.textanalyticsnews.com/news/%E2%80%9C-challenge-still-accuracy-sentiment-p
E2%80%9D

For example, let say we have two sets of documents, D_1 with documents containing negative opinions about topic X and D_2 being a set of documents with negative opinions about topic Y. Extracting the list of most opinionated terms from both sets of documents and then comparing them using semantic relations of WordNet can reveal interesting information about vocabulary of opinions in a particular domain and across domains. In addition to this, use of social features like bloggers age, gender, location, etc., would enable us to analyze use of opinion vocabulary from various dimensions.

11.2.4   Social Framework for Opinion Detection

Improving the framework we have proposed for opinion related tasks is part of our future work. This includes mathematical modeling of our framework and experimentation with a suitable data collection to use social evidences we have proposed.

11.2.5   Automatic Weight Balancing Function

As part of our future work, we plan to focus on another major problem of opinion detection. According to Macdonald et al. [52], sometimes even the best opinion finding approaches fail to improve relatively stronger baselines. This observation was the basic motivating factor to use the strongest baseline of TREC Blog (i.e. baseline-4) in our experimentation. Combining topic-relevance and opinion scores to obtain final score of a document has a large impact on the performance of opinion-finding systems. Developing a function that can take topic-relevance values and opinion values of documents and suggest proper weights for both of them for their ideal combination would be interesting to work on.

Summarization of Opinion Finding Approaches by TREC Blog Participants (Chapter 4)

Table A.1: *Table Summarizing Document-Level Approaches of TREC Blog Track*

Task 1	Baseline Adhoc Retrieval Task
Task 2	Opinion Finding Task
Task 3	Opinion Polarity Task
Task 4	Blog Feed Distillation Task

Title of Paper	Tasks	External Collection	Lexical Resource	Query Exp.	ML Classifiers	Rel. FB	3rd Party SW	Retrieval	Opinion Find
Cas-ict.blog.final	2	X	General Inq. (GI)	Using Adj terms using GI and Corpus both	X	X	FirteX for Retrieval	Language Model	Language Model
Cmu.blog.final (Topic classification)	2	Movie Review Data Pang and Lee (http://www.cs.cornell.edu/People/pabo/movie-review-data/) + Customer Review Data . Hu and Liu (http://www.cs.uic.edu/~liub/FBS/FBS.html)	manually created list of verbs and adjectives	using web	Baysean Logistic Regression toolkit	X	Minipar Parser for POS	Indri toolkit with passage retrieval	Sentence to passages, Knowledge transfer
Indianau.blog.final (results of 4 modules added using Similarity Merge and Weighted Sum)	2	BlogPulse and Technocrati blogs	Lexicon,collocations,Morphology. Lexicon using 700 blogs from Blogpulse and Technocrati	using WIDIT	X	X	WIDIT	VSM with Okapi BM25	4 modules: 1)Opinion Term Module 2) Rare Term Module 3)IIU Module 4)Adj-Verb Module
Nii.blog.fmace	2	X	Sentimental words of Opinion Finder	X	X	X	Indri	Language Model	Language Model
Robert-gordonu.blog.final	2	X	X	List of subjectiveer terms taken from different review sites	some binary classifier not mentioned	X	X	Lang. Model	Lang. Model and shallow parsing techniques
Uamsterdam.blog.final (Linear addition of modules)	2	blogspot.com for SPAM likelihood module	General Inq.	X	SVM used in the Spam likelihood component of opinion finding	Blind in retrieval	X	Lang. Model with Blind RF	Modules: opinion Expr+Post Quality+ Spam Likelyhood+ Link based Authority
Uarkansas.blog.final (linear addition of rel. Score + Opnn score)	2	X	X	X	SVM	X	Lucene for indexing with TF*IDF; LibSVM	Lucene with TFIDF	Features
Ucaliforniaki-sc.blog	2	Yahoo Movie Reviews, Epinion Digital Camera Review data, Reuter Newswire data	SentiWordNet (SWN)	adjectives from SWN, bigrams and trigrams from training data	Logistic Regression Classifier	X	Lemur	Lang Model	Logistic Regression Classifier
Umd.blog.ent.legal.qa.final (used Wilcoxon test)	2	X	Wilson Subj. List	X	X	X	Indri for passage indexing	Use own algo for passage selection	PMI
Univ. of maryland	2	X	Manually created list of positive and negative words	X	SVM	X	Lucene for indexing with TF*IDF	Lucene with TFIDF	(Proximity, Positive+negative words, Lucene rel. Score)
Univ. Of illinois (chi square used too)	2	Ratetall.com, Wikipedia	X	Using wikipedia, web and RL Feedback	SVM for Sentence Classification	Yes	Concept Sim+ Term Simi	X	Number of Subjective sentences
Upisa.blog.fiynal		X	SWN to tag the collection with opinion words and then searching is done using proximity approach	X	X	X	Opinion Classifier	Mixed Ret+Opinion	Mixed Ret+Opinion

TREC Blog 2007

Title of Paper	Tasks	External Collection	Lexical Resource	Query Exp.	ML Classifiers	Rel. FB	3rd Party SW	Retrieval	Opinion Find
Cas-ict.blog.final	1,2,3	X	General Inq.	X	Drag Push Classifier	X	FirteX for Retrieval using BM25 and Language Model. They say LM is better so it was used	BM25+LM	Dragpush
Cas-nlpr.blog (blog page segmentation using SVM)	1,2	Movie Review Data,Customer Review Data and data from www.yelp.com for sentence training	X	X	Naïve Bayes Regression method to score sentences	X	Indri for indexing	Document Retrieval and Sentence Retrieval	Naïve Bayes Regression for sentence opinion scores. Scores for doc. Topical and sentence opinion are added using SVR
Dalianu.blog.final	1,2,3	X	Manually built list of 2000 sentimental words frequent in the corpus	Sentimental words around topic word	SVM for polarity	tried but did not perform well	Indri	Indri with query expansion	Sentimental words around topic word. Information Gain+SVM for polarity on sentence level
Fub.blog.final	1,2	X	Automatically learned from corpus using DFR+ Subj	X	X	X	LingPipe for removing non-english docs. Terrier for	Terrier with model PL2 + DPH	List of sentimental terms using DFR to give an
Fudan.blog.final	1,2	Movie review data on http://www.cs.cornell.edu/Peo ple/pabo/movie-review-data/	X	using Pseudo RF. Top K terms are ranked using SVM	CME classifier for sentence subjectivity. Trained on movie review data on http://www.cs.cornell.edu /People/pabo/movie-review-data/	Pseudo RF	X	Lucene	CME Classifier to predict subjectivity of each sentence and then SVM was used to calculate opinion score for the post based on sentence analysis
Indianau.blog.final	1,2	Movie review data on http://www.cs.cornell.edu/Peo ple/pabo/movie-review-data/ for IIU module and acroymns on http://www.netlingo.com/acro nyms.php	Wilson Subj. List	using WIDIT	X	X	WIDIT	Okapi BM25	4 modules: 1) High Freq Module 2) Low Freq Module 3)IIU Module 4)Opinion Acronym module 5) Wilson's Lexicon Module
Kobeu.blog.final	1,2	Amazon.com for extracting terms	List of terms from amazon.com	X	SVM and KNN	X	Lucene	Lucene with VSM using TF*IDF	SVM and KNN used
Niu.blog.final	1,2	X	List of terms found in opinionFinder	X	Cross Entropy	X	X	Generative Model for retrieval	Generative Model for retrieval
Ntu.blog.final	1,2,3	X	General Inq.	X	SVM	X	Lemur, LIbSVM	Lemur with TFIDF	Sentences selected with Topic words inside. Also sentences around such sentences.
Robert-gordonu.blog.final	1,2	X	X	X	SVM	X	X	Lucene	Regression based SVM with Information Gain measure
Acxiom Corporation	1,2	X	X	with Pseudo RF	SVM with Topic categoziation	Yes	Lemur	Lemur	Modules

Title of Paper	Tasks	External Collection	Lexical Resource	Query Exp.	ML. Classifiers	Rel. FB	3rd Party SW	Retrieval	Opinion Find
Uamsterdam-weerkamp.blog.final	1,2,3	AQUAINT-2 news corpus for query expansion using RF	List of opinion terms of OpinionFinder	Yes using Relevance Model Concept with AQUAINT-2 news corpus	X	Yes	Indri	Indri with Language Model with Query Rewriting	1)List of opinion terms from OpinionFinder system 2) Comment based approach. Polarity with number of positive,negative words
Uglasgow.blog.ent.final	1,2	to build lexicon but not mentioned	Built from various resources	X	X	X	Opinion Finder	Field-based model from DFR (Divergence from Randomness) Framework	1) Dictionary based Statistical app with Lexicon from various resources using Divergence 2) NLP based approach usng OpinionFinder
Uic-zhang.blog.final	1,2,3	Rateitall.com, Wikipedia,epinions.com	Manually built	Using wikipedia, web and RL Feedback	SVM	Yes	X	Concept based retrieval by Concpt Sim+ Term Sim using Okapi 25	chi-square as feature selection, SVM on sentence level, SVM for polarity task too
uwaterloo.blog.final	1,2	X	List of subj Adjectives by Hatzivassiloglou and McKeown	X	X	X	Wumpus IR system	Wumpus IR system with BM25	30 opinion words around topic words on both sides. They hypothesise that presence of subjective adjectives within fixed-size windows around query term instances in a document is a useful feature for finding opinions directed at the query concept.
Wuhanu.blog.final	1,2,3	X	SWN	X	X	X	X	BM25	SWN terms > 0.4. Number of positive and negative words for polarity
TREC Blog 2008									
Beijing University of Posts and Telecommunications	1,2	X	X	Google based Query Expansion	Max. Entropy Classifier on sentence level judges the sentences as subjective and then another ME Classifier judges the doc as opinionated on behalf of Sentence classifier	X	Indri	Indri	Max. Entropy Classifier on sentence and doc level

Title of Paper	Tasks	External Collection	Lexical Resource	Query Exp.	ML Classifiers	Rel. FB	3rd Party SW	Retrieval	Opinion Find
Dublin City University Ireland	1,2,3	X	SWN	With Terrier DFR	Binary Classifier	X	Terrier	Terrier Okapi BM25	1) Surface Features 2)Syntactic Features 3)Lexicon features …. Used binary classifier for both opinion and polarity task and then SUM the scores of all using CombSum approach
IIT Kharagpur also Blog Distil	1,2,3	Pang and lee movie review data http://www.cs.cornell.edu/Peo ple/pabo/moviereview-data/ + Movie Database IMDB archive of rec.arts.movies.reviews + Sentence dataset: Objective sentences were extracted from Internet Movie Databases plot summaries while the subjective sentences are from Rotten Tomatoes review snippets + Set of opinion sentences from (www.minekey.com + Parallel sentences from QRELS of 2006–07	X	X	Nave Bayes, Maximum Entropy, Support Vector Machine like Pang and Lee	X	Lucene	Lucene+TFIDF+two indices i.e. one for sentence level and other Doc level	Machine Learning with Features
Indiana University	1,2,3	Wikipedia, movie reviews and plot summaries http://www.cs.cornell.edu/peo ple/pabo/movie-review-data/ and http://www.imdb.com/Section s/Plots/	Wilson Subj. List	Web based Query Expansion using Wikipedia and Google	X	X	X	Exact Match+Proximity Match+Noun Phrase Match+Non-Rel Match. Normalized by doc length	4 modules: 1) High Freq Module 2) Low Freq Module 3)IIU Module 4)Opinion Acronym module 5) Wilson's Lexicon Module. For polarity, they also used Valence Shifters i.e. not, never to change context
Korea University	1,2,3	X	GI, WordNet, Wilson= Final List	Pseudo RF	SVM		SVM light used for both Opinion and polarity task	Lang. Model with KL Div	Hybrid approach= ML+Lexicon. First by Lexicon and then combined using SVM

Title of Paper	Tasks	External Collection	Lexical Resource	Query Exp.	ML Classifiers	Rel. FB	3rd Party SW	Retrieval	Opinion Find
Univ. Of Lugano	1,2	X	X	X	SVM	X	Terrier	Terrier with DFR_BM25 and further used SVM-map rank learner to improve it	SVM to sum score of terms for doc
Pohang University korea_Also Blog Distillation	1,2,3	X	Different types of lexicon used for defining subj of a word: SWN, Corpus based lexicon,Query specific opinion lexicon, Passage Context	opinionated word (POW), representative of all opinionated terms added to original query	X	Passage based Pseudo RF	X	Passage Based LM to use score of best passage	add opinion score of all terms in the doc. Same tech for polarity with difference that terms are considered +ve o negative instead of subj or obj
The University of Texas at Dallas	1,2,3	movie review from pang and lee,Customer review from Hu and liu and 256 hotel reviews to generate sentiment terms using MI with seed words	X	For the title field only retrieval, a Google-set based query expansion method was used. We used the built-in pseudo feedback model in the	Not mentioned	X	Lemur IR toolkit	Lemur IR toolkit using TFIDF	split a blog into sentences based on sentence boundary detection2, and used Lemur to retrieve top 5 relevant sentences
Tsinghua University china	1,2,3	X	HowNet in Chinese and then translated to English	X	X	X	Tminer	TMiner search engine, from IR group of Tsinghua University using BM25	Simple. Read Paper

Title of Paper	Tasks	External Collection	Lexical Resource	Query Exp.	ML Classifiers	Rel. FB	3rd Party SW	Retrieval	Opinion Find
University of Glasgow	1,2,3	Aquaint2 collection for query expansion of a baseline	Sentimental words of Opinion Finder	through Aquaint2 collection and from target collection automatically	X	X	Terrier	Terrier with DFR	Sentence based using opinion Finder. Better read paper for opinion and polarity
University of Illinois at Chicago	1,2,3	Rateitall.com	X	Wikipedia and Google	SVM	X	X	concept identification, query expansion, concept based retrieval and document filter	SVM using NEAR operator+ Polarity: Sentence level PD+Doc level PD
Univ. of Neuchatel Switzerland	1,2,3	X	X	Blind RF and Wikipedia	Logistic Regression Classifier	Blind RF	X	Okapi BM25 and DFR	Characteristic Vocabulary (In Muller's approach the basic idea is to use Z-score (or standard score) to determine which terms can properly
University of Tor Vergata, Rome, italy	1,2,3	X	3 adhoc weighted lexicons: one for opinion and 2 for polarity	X	X	X	Terrier	DFR and PL2 using Terrier	DFR and PL2
University of Waterloo canada	1,2,3	X	lexicon of subjective words and phrases, gathered from a variety of sources, such as FrameNet , Levin's verb classes , Hatzivassiloglou and McKeown's list of subjective adjectives	X	X	X	Wumpus search engine	Wumpus search engine with BM25	Kullback-Leibler divergence (KLD) [4] to weight subjective words, and factors these weights into the document score.
York University Canada	1,2	X	Pre-determined sentiment word list	X	X	X	Compass IR system	Compass IR system with BM25	feautures: Sentimental term freq and Sentimental term

Features used for Combined Approach for Opinion Finding (Chapter 8)

Table B.1: *Features based on Simple Heuristics*

Emotivity	
Emot 1	Emotivity (1) of the Document
Emot 2	Emotivity (2) of the Document
Reflexivity	
TotRef	Total number of words from list R appearing in the Document
Refl	Average of words from list R appearing in the Document
Addressibility	
TotAdd	Total number of words from list A appearing in the Document
Addr	Average of words from list A appearing in the Document
Common Phrases	
TotCp	Total number of opinion phrases appearing in the Document
AvgCp	Average number of opinion phrases appearing in the Document
Subjectivity	
Subjectivity (1)	
VSubj1	Subjectivity (1) of the Verbs of the Document
AdvSubj1	Subjectivity (1) of the Adverbs of the Document
AdjSubj1	Subjectivity (1) of the Adjectives of the Document
Subjectivity (2)	
VSubj2	Subjectivity (2) of the Verbs of the Document
AdvSubj2	Subjectivity (2) of the Adverbs of the Document
AdjSubj2	Subjectivity (2) of the Adjectives of the Document
Subjectivity (3)	
VSubj3	Subjectivity (3) of the Verbs of the Document
AdvSubj3	Subjectivity (3) of the Adverbs of the Document
AdjSubj3	Subjectivity (3) of the Adjectives of the Document
Subjectivity (4)	
VSubj4	Subjectivity (4) of the Verbs of the Document
AdvSubj4	Subjectivity (4) of the Adverbs of the Document
AdjSubj4	Subjectivity (4) of the Adjectives of the Document

Table B.2: *POS-based Features*

Adjectives	
JJ	Number of Adjectives in a document marked as "JJ" by POS Tagger
JJR	Number of Adjectives in a document marked as "JJR" by POS Tagger
JJS	Number of Adjectives in a document marked as "JJS" by POS Tagger
Adverbs	
RB	Number of Adverbs in a document marked as "RB" by POS Tagger
RBR	Number of Adverbs in a document marked as "RBR" by POS Tagger
RBS	Number of Adverbs in a document marked as "RBS" by POS Tagger
Verbs	
VB	Number of verbs in a document marked as "VB" by POS Tagger
VBD	Number of verbs in a document marked as "VBD" by POS Tagger
VBZ	Number of verbs in a document marked as "VBZ" by POS Tagger
VBG	Number of verbs in a document marked as "VBG" by POS Tagger
VBN	Number of verbs in a document marked as "VBN" by POS Tagger
VBP	Number of verbs in a document marked as "VBP" by POS Tagger
NOUNS	
NN	Number of nouns in a document marked as "NN" by POS Tagger
NNP	Number of nouns in a document marked as "NNP" by POS Tagger
NNS	Number of nouns in a document marked as "NNS" by POS Tagger
NNPS	Number of nouns in a document marked as "NNPS" by POS Tagger

Table B.3: *Relevancy Based Features*

Relevancy Based Features	
Rank	Relevance rank of a Document as given in baseline
Score	Relevance Score of a Document as in Baseline

Table B.4: *Miscellaneous Features*

Totql Number of Polar Words	
TotPos	Tota Number of Positive Words in the Document
TotNeg	Tota Number of Negative Words in the Document
TotNeu	Tota Number of Neutral Words in the Document
Total Number of Words	
TPOS	Total Number of Words in POS Tagged Document
TSim	Total Number of Words in Original Document

Bibliography

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, O. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMASS at TREC 2004: Novelty and hard. In *Proceedings of TREC-13*. National Institute of Standards and Technology (NIST), 2004.

[2] Jennifer Foster Deirdre Hogan Adam Bermingham, Alan Smeaton. DCU at the TREC 2008 blog track. In *Proceedings of the Text REtrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[3] Nitin Agarwal and Huan Liu. Blogosphere: research issues, tools, and applications. *SIGKDD Explor. Newsl.*, 10(1):18–31, 2008.

[4] Central Intelligence Agency. The world factbook 2007. In *https://www.cia.gov/library/publications/the-world-factbook/*, 2007.

[5] Shaishav Agrawal and Tanveer j Siddiqui. Using syntactic and contextual information for sentiment polarity analysis. In *ICIS '09: Proceedings of the 2nd International Conference on Interaction Sciences*, pages 620–623, New York, NY, USA, 2009. ACM.

[6] Panagiota Alevizou. Digital generations: Children, young people, and the new media. In *Education, Communication & Information, Volume 5, Issue 1*, pages 83–90. Routledge, 2006.

[7] James Allan. Hard track overview in TREC-2004 (notebook) high accuracy retrieval from documents. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-12)*, pages 24–37. National Institute of Standards and Technology (NIST), 2003.

[8] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. Clustering and exploring search results using timeline constructions. In *Proceeding of the 18th ACM conference on Information and knowledge management (CIKM-09)*, pages 97–106, New York, NY, USA, 2009. ACM.

[9] Omar Alonso, Michael Gertz, and Ricardo A. Baeza-Yates. On the value of temporal information in information retrieval. In *Proceedings of SIGIR Forum*, volume 41, pages 35–41, 2007.

[10] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[11] Alina Andreevskaia and Sabine Bergler. Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proceedings of the 11rd Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 209–216, 2006.

[12] Alina Andreevskaia and Sabine Bergler. Semantic tag extraction from wordnet glosses. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-2006)*, 2006.

[13] National School Boards Association. Creating and connecting: Research and guidelines on online social and educational networking. In *www.nsba.org/site/docs/41400/41340.pdf*, 2007.

[14] J. Atserias, H. Zaragoza, M. Ciaramita, and G. Attardi. Semantically annotated snapshot of the English Wikipedia. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

[15] Giuseppe Attardi and Maria Simi. Blog mining through opinionated words. In *Proceedings of TREC 2006, the Fifteenth Text Retrieval Conference*, Gaithersburg, US, 2006. National Institute of Standards and Technlogy (NIST).

[16] Gamon Aue, Anthony and Michael. Customizing sentiment classifiers to new domains: a case study. In *RANLP-05: Proceedings of International Conference on Recent Advances in Natural Language Processing*, 2005.

[17] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *In 6th Intl Semantic Web Conference, Busan, Korea*, pages 11–15. Springer, 2007.

[18] W. Weerkamp B. Ernsting and M. de Rijke. Language modeling approaches to blogpost and feed finding. In *Proceedings of the Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[19] Peter Bailey, Arjen P. De Vries, Nick Craswell, and Ian Soboroff. Overview of the trec-2007 enterprise track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC-14)*. National Institute of Standards and Technology (NIST), 2007.

[20] K. Balog, J. He, M. Bron, M. de Rijke, and W. Weerkamp. The university of amsterdam (isla) at inex 2009. In *Proceedings of INEX-2009*, December 2009.

[21] K. Balog, P. Serdyukov, and A. de Vries. Overview of the TREC 2010 Entity Track. In *Proceedings of Text REtrieval Conference (TREC-2010)*. National Institute of Standards and Technology (NIST), 2010.

[22] Krisztian Balog. People search in the enterprise. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 916–916, 2007.

[23] Krisztian Balog, Marc Bron, and Maarten De Rijke. Category-based query modeling for entity search. In *In 32nd European Conference on Information Retrieval (ECIR 2010)*, pages 319–331, 2010.

[24] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145, London, UK, 2002. Springer-Verlag.

[25] Nilesh Bansal and Nick Koudas. Blogscope: spatio-temporal analysis of the blogosphere. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1269–1270, New York, NY, USA, 2007. ACM.

[26] J.R. Baron, D.D. Lewis, and D.W. Oard. TREC-2006 legal track overview. In *The Fifteenth Text REtrieval Conference (TREC-2006) Proceedings*. Citeseer, National Institute of Standards and Technology (NIST), 2006.

[27] Marco Baroni and Stefano Vegnaduzzo. Identifying subjective adjectives through web-based mutual information. In *Proceedings of the 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing) KONVENS04*, pages 613–619, 2004.

[28] Holger Bast, Alexandru Chitea, Fabian Suchanek, and Ingmar Weber. Ester: efficient search on text, entities, and relations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 671–678, 2007.

[29] Do T. Payne A. Jones S. Beaulieu, M. The enquire okapi project. *British Library Research and Innovation Report*, 1997.

[30] Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan. The sentimental factor: improving review classification via human-provided information. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[31] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 203–206, 2007. Short paper.

[32] Michael Bendersky and Oren Kurland. Utilizing passage-based language models for document retrieval. In *ECIR'08: Proceedings of the IR research, 30th European conference on Advances in information retrieval*, pages 162–174, Berlin, Heidelberg, 2008. Springer-Verlag.

[33] Michael Bendersky and Oren Kurland. Utilizing passage-based language models for ad hoc document retrieval. *Journal of Information Retrieval*, 13:157–187, 2010.

[34] K. Berberich, S. Bedathur, T. Neumann, and G. Weikum. A time machine for text search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 519–526. ACM New York, NY, USA, 2007.

[35] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR-1999, pages 222–229, 1999.

[36] Amanda Spink Bernard, Bernard J. Jansen, and H. Cenk Ozmultu. Use of query reformulation and relevance feedback by excite users. *Internet Research*, 10:317–328, 2000.

[37] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In James G. Shanahan, Janyce Wiebe, and Yan Qu, editors, *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pages 22–24, Stanford, US, 2004.

[38] Yang Liu Bin Li, Feifan Liu. UTDallas at TREC 2008 blog track. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[39] Jane Black. The perils and promise of online schmoozing. In *http://www.businessweek.com/technology/content/feb2004/tc20040220_ 3260_ tc073.htm*.

[40] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

[41] Kenneth Bloom, Sterling Stein, and Shlomo Argamon. Appraisal extraction for news opinion analysis at ntcir-6. 2007.

[42] Victoria Bobicev, Victoria Maxim, Tatiana Prodan, Natalia Burciu, and Victoria Anghelus. Emotions in words: Developing a multilingual wordnet-affect. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 375–384. Springer Berlin / Heidelberg.

[43] Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval*, 12:526–558, 2009. 10.1007/s10791-008-9070-z.

[44] Denise Sevick Bortree. Presentation of self on the web: an ethnographic study of teenage girls' weblogs. 5:25 – 39, March 2005.

[45] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[46] Marc Bron, Krisztian Balog, and Maarten de Rijke. Ranking related entities: components and analyses. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1079–1088, New York, NY, USA, 2010. ACM.

[47] Penelope Brown and Stephen C. Levinson. *Politeness : Some Universals in Language Usage (Studies in Interactional Sociolinguistics).* Cambridge University Press, February.

[48] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16:79–85, June 1990.

[49] Rebecca F. Bruce and Janyce M. Wiebe. Recognizing subjectivity: A case study in manual tagging. *Journal of Natural Language Engineering*, 5(2):187–205, 1999.

[50] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. In *Proceedings of the Text Retrieval Conference (TREC-03).*

[51] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *SIGIR*, pages 33–40, 2000.

[52] I. Soboroff C. Macdonald, I. Ounis. Overview of the TREC blog track 2007. In *The 16th Text REtrieval Conference, (TREC 2007) Proceedings*, 2007.

[53] I. Soboroff C. Macdonald, I. Ounis. Overview of the TREC blog track 2008. In *The 17th Text REtrieval Conference, (TREC 2008) Proceedings*, 2008.

[54] Fidel Cacheda, Vassilis Plachouras, and Iadh Ounis. A case study of distributed information retrieval architectures to index one terabyte of text. *Journal of Information Process Management*, 41(5):1141–1161, 2005.

[55] D. Cai, S. Yu, J.R. Wen, and W.Y. Ma. Block-based web search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 456–463, New York, NY, USA, 2004. ACM.

[56] James P. Callan. Passage-level evidence in document retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[57] Christopher S. Campbell, Paul P. Maglio, Alex Cozzi, and Byron Dom. Expertise identification using email communications. In *Proceedings of the twelfth*

international conference on Information and knowledge management, CIKM '03, pages 528–531, New York, NY, USA, 2003. ACM.

[58] François-Régis Chaumartin. Upar7: a knowledge-based system for headline sentiment tagging. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 422–425, Morristown, NJ, USA, 2007. Association for Computational Linguistics.

[59] Tao Cheng and Kevin Chen-Chuan Chang. Entity Search Engine: Towards Agile Best-Effort Information Integration over the Web. In *CIDR 2007, Third Biennial Conference on Innovative Data Systems Research, CA, USA, January 7-10, 2007, Online Proceedings*, pages 108–113. www.crdrdb.org, 2007.

[60] Tao Cheng, Xifeng Yan, and Kevin Chen-Chuan Chang. EntityRank: Searching Entities Directly and Holistically. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 387–398. VLDB Endowment, 2007.

[61] Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. Extracting semantics relationships between wikipedia categories. In *Proceedings of the First Workshop on Semantic Wikis – From Wiki To Semantics*, Workshop on Semantic Wikis, 2006.

[62] P. Chesley, B. Vincent, L. Xu, and R. Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pages 27–29, 2006.

[63] Yun Chi, Belle L. Tseng, and Junichi Tatemura. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 68–77, New York, NY, USA, 2006. ACM.

[64] Yun Chi, Belle L. Tseng, and Junichi Tatemura. Eigen-trend: trend analysis in the blogosphere based on singular value decompositions. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 68–77, New York, NY, USA, 2006. ACM.

[65] Yun Chi, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, and Belle L. Tseng. Structural and temporal analysis of the blogosphere through community factorization. In *KDD '07: Proceedings of the 13th ACM SIGKDD international*

conference on Knowledge discovery and data mining, pages 163–172, New York, NY, USA, 2007. ACM.

[66] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition with a maximum entropy approach. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 160–163, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[67] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, 1990.

[68] Jacques Savoy Claire Fautsch. UniNE at TREC-2008: Fact and opinion retrieval in the blogsphere. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[69] Cyril W Cleverdon. Aslib cranfield research project: report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report, Staff publications - Cranfield Library, 1962.

[70] Cyril W Cleverdon. The cranfield tests on index language devices. pages 47–59, 1997.

[71] CLICKZ. Clickz stats: Web worldwide. In *http://www.clickz.com/*, 2007.

[72] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[73] William W. Cohen and Sunita Sarawagi. Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 89–98, New York, NY, USA, 2004. ACM.

[74] Jack G. Conrad, Jochen L. Leidner, Frank Schilder, and Ravi Kondadadi. Query-based opinion summarization for legal blog entries. In *ICAIL '09: Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 167–176, New York, NY, USA, 2009. ACM.

[75] Jack G. Conrad and Mary Hunter Utt. A system for discovering relationships by feature extraction from text databases. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 260–270, 1994.

[76] Andres Corrada-Emmanuel, W. Bruce Croft, and Vanessa Murdock. Answer passage retrieval for question answering. Technical report, University of Massachusetts, 2003.

[77] Nick Craswell, Gianluca Demartini, Julien Gaugaz, and Tereza Iofciu. *L3S at INEX 2008: Retrieving Entities Using Structured Information*, pages 253–263. Springer-Verlag, Berlin, Heidelberg, 2009.

[78] Nick Craswell, David Hawking, Anne-Marie Vercoustre, and Peter Wilkins. P@noptic expert: Searching for experts not just for documents. In *In Ausweb*, pages 21–25, 2001.

[79] Nick Craswell, Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Relevance weighting for query independent evidence. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, New York, NY, USA, 2005. ACM.

[80] Bruce W. Croft and John Lafferty. *Language Modeling for Information Retrieval (The Information Retrieval Series)*. Springer, December 1999.

[81] W. B. Croft and D. J. Harper. *Using probabilistic models of document retrieval without relevance information*, pages 339–344. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[82] Sanjiv R. Das and Mike Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Manage. Sci.*, 53(9):1375–1388, 2007.

[83] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW-03, 12th International Conference on the World Wide Web*, pages 519–528, Budapest, HU, 2003. ACM Press.

[84] Yohei Seki Le Sun Hsin-Hsi Chen Noriko Kando David Kirk Evans, Lun-Wei Ku. Opinion analysis across languages: An overview of and observations from the NTCIR-6 opinion analysis pilot task. In *Proceedings of NTCIR-6 Workshop Meeting*, 2007.

[85] Gianluca Demartini, Claudiu Firan, Tereza Iofciu, Ralf Krestel, and Wolfgang Nejdl. Why finding entities in Wikipedia is difficult, sometimes. *Information Retrieval*, 13:534–567, October 2010.

[86] Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, and Wolfgang Nejdl. Semantically enhanced entity ranking. In *Proceedings of the 9th international conference on Web Information Systems Engineering*, WISE '08, pages 176–188, Berlin, Heidelberg, 2008. Springer-Verlag.

[87] Gianluca Demartini, Julien Gaugaz, and Wolfgang Nejdl. A vector space model for ranking entities and its application to expert search. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 189–201, Berlin, Heidelberg, 2009. Springer-Verlag.

[88] Gianluca Demartini, Tereza Iofciu, and Arjen de Vries. Overview of the INEX 2009 entity ranking track. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *Focused Retrieval and Evaluation*, volume 6203 of *Lecture Notes in Computer Science*, pages 254–264. Springer Berlin / Heidelberg.

[89] Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, and Hugo Zaragoza. Entity Summarization over time in News Articles. In *ACM SIGIR Special Interest Group on Information Retrieval (SIGIR'2010), Geneva, Switzerland*, pages 795–796. ACM, July 2010.

[90] Gianluca Demartini, Malik Muhammad Saad Missen, Roi Blanco, and Hugo Zaragoza. TAER: Time Aware Entity Retrieval. In *Conference on Information and Knowledge Management (CIKM), Toronto, Canada*. ACM, 2010.

[91] Gianluca Demartini, Arjen P. Vries, Tereza Iofciu, and Jianhan Zhu. Overview of the INEX 2008 entity ranking track. pages 243–252, 2009.

[92] K Denecke. Using Sentiwordnet for multilingual sentiment analysis. In *Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on Data Engineering*, pages 507 – 512. IEEE, 2008.

[93] Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML corpus. *SIGIR Forum*, 40(1):64–69, 2006.

[94] Fernando Diaz. Integration of news content into web results. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 182–191, New York, NY, USA, 2009. ACM.

[95] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, R. Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A. Tomlin, and Jason Y. Zien. Semtag and seeker: bootstrapping the semantic web via

automated semantic annotation. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 178–186, New York, NY, USA, 2003. ACM.

[96] Xiaowen Ding, Bing Liu, and Yu Philip S. A holistic lexicon-based approach to opinion mining. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 231–240, New York, NY, USA, 2008. ACM.

[97] Paul Ekman. Facial expression and emotion. *American Psychologist*, 48:384–392, 1993.

[98] Charlotta Engström. Topic dependence in sentiment classification. Master's thesis, University of Cambridge, July 2004.

[99] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of Fourteenth Conference on Information and Knowledge Management (CIKM-05)*, pages 617–624, 2005.

[100] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06)*, pages 417–422, 2006.

[101] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: an experimental study. *Artif. Intell.*, 165(1):91–134, 2005.

[102] Henry Farrell and Daniel Drezner. The power and politics of blogs. *Public Choice*, 134(1):15–30, January 2008.

[103] Nicola Ferro. CLEF, CLEF 2010, and Promises: Perspectives for the cross-language evaluation forum. In *Proceedings of NTCIR-8 Workshop Meeting*. ACM, 2010.

[104] Radu Florian. Named entity recognition as a house of cards: classifier stacking. In *proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[105] Tadanobu Furukawa, Yutaka Matsuo, Ikki Ohmukai, Koki Uchiyama, and Mitsuru Ishizuka. Social networks and reading behavior in the blogosphere abstract. In *Int'l Conf on Weblogs and Social Media*, San Jose, California,USA, 2007.

[106] Michael Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 841–847, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[107] Kavita Ganesan and ChengXiang Zhai. Opinion-based entity ranking. *Journal of Information Retrieval*.

[108] G. Gaughan and A.F. Smeaton. Finding new news: Novelty detection in broadcast news. *Lecture notes in computer science*, 3689:583, 2005.

[109] Shima Gerani, Mark J. Carman, and Fabio Crestani. Investigating learning approaches for blog post opinion retrieval. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 313–324, Berlin, Heidelberg, 2009. Springer-Verlag.

[110] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288, 2002.

[111] Dan Gillmor. *We the Media: Grassroots Journalism by the People, for the People*. O'Reilly Media, January.

[112] Ayse Goker, John Davies, and Margaret Graham. *Information Retrieval: Searching in the 21st Century*. John Wiley & Sons, 2007.

[113] J. Golbeck and J. Hendler. FilmTrust: movie recommendations using trust in web-based social networks. In *3rd IEEE Consumer Communications and Networking Conference (CCNC-06)*, pages 282–286.

[114] Jennifer Golbeck. Generating predictive movie recommendations from trust in social networks. In *The 4th International Conference on Trust Management*, pages 1–16, May 2006.

[115] Jennifer Golbeck and James Hendler. Inferring binary trust relationships in web-based social networks. *ACM Trans. Internet Technol.*, 6(4):497–529, 2006.

[116] Jennifer Ann Golbeck. *Computing and applying trust in web-based social networks*. PhD thesis, College Park, MD, USA, 2005.

[117] Gregory Grefenstette, Yan Qu, David Evans, and James Shanahan. Validating the coverage of lexical resources for affect analysis and automatically classifying new words along semantic axes. In James Shanahan, Yan Qu, and Janyce

Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*, volume 20 of *The Information Retrieval Series*, pages 93–107. Springer Netherlands, 2006.

[118] Gregory Grefenstette, Yan Qu, James G. Shanahan, and David A. Evans. Coupling niche browsers and affect analysis for an opinion mining. In *Proceedings of RIAO*, pages 186–194, 2004.

[119] Ralph Gross and Alessandro Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society (WPES-05)*, pages 71–80. ACM.

[120] Hemant Joshi GuangXu Zhou and Coskun Bayrak. Topic categorization for relevancy and opinion detection. In *Proceedings of the Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[121] R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 403–412, 2004.

[122] Donna Harman. Overview of the trec 2002 novelty track. In *TREC*, 2002.

[123] S.P. Harter. An algorithms for probabilistic indexing. *ournal of the American Society for Information Science*, 26(4):280–289, 1975.

[124] Jean Hartley. Employee surveys - strategic aid or hand-grenade for organisational and cultural change? *International Journal of Public Sector Management*.

[125] Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min yen Kan, and Kathleen R. McKeown. Simfinder: A flexible clustering tool for summarization. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 41–49, 2001.

[126] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[127] Vasileios Hatzivassiloglou and Janyce M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics*, pages 299–305, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[128] S. I. Hayakawa. *Choose the Right Word.* HarperCollins Publishers, 1994.

[129] Ben He, Craig Macdonald, Jiyin He, and Iadh Ounis. An effective statistical approach to blog post opinion retrieval. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1063–1072, New York, NY, USA, 2008. ACM.

[130] Ben He, Craig Macdonald, and Iadh Ounis. Ranking opinionated blog posts using opinionfinder. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 727–728, New York, NY, USA, 2008. ACM.

[131] Ben He, Craig Macdonald, Iadh Ounis, Jie Peng, and Rodrygo L.T. Santos. University of glasgow at TREC-2007: Experiments in blog and enterprise tracks with terrier. In *Proceedings of the 16th Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[132] Ben He, Craig Macdonald, Iadh Ounis, Jie Peng, and Rodrygo L.T. Santos. University of glasgow at TREC-2008: Experiments in blog and enterprise tracks with terrier. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[133] Marti A. Hearst, Matthew Hurst, and Susan T. Dumais. What should blog search look like? In *SSM '08: Proceeding of the 2008 ACM workshop on Search in social media*, pages 95–98, New York, NY, USA, 2008. ACM.

[134] Coskun Bayrak Hemant Joshi and Xiaowei Xu. UALR at TREC blog track. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[135] William Hersh. *Information Retrieval: A Health and Biomedical Perspective.* Health Informatics. Springer, third edition.

[136] Feng Hou and Guo-Hui Li. Mining chinese comparative sentences by semantic role labeling. In *Proceedings of International Conference on Machine Learning and Cybernetics, 2008*, pages 2563 – 2568, 2008.

[137] Keyun Hu, Yuchang Lu, Lizhu Zhou, and Chunyi Shi. Integrating classification and association rule mining: A concept lattice framework. In *Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, RSFDGrC '99, pages 443–447, London, UK, 1999. Springer-Verlag.

[138] M. Hu and B. Liu. Opinion feature extraction using class sequential rules. In *Proceedings of the Spring Symposia on Computational Approaches to Analyzing Weblogs*, 2006.

[139] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of KDD '04, the ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, Seattle, US, 2004. ACM Press.

[140] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*, pages 755–760. AAAI Press, 2004.

[141] David A. Huffaker and Sandra L. Calvert. Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication*, (2).

[142] Peter Hui and Michelle Gregory. Quantifying sentiment and influence in blogspaces. In *1st Workshop on Social Media Analytics*, SOMA-10, Washington, DC, USA, 2010. ACM.

[143] Lei Du Si Li Huiji Gao Weiran Xu Jun Guo Hui He, Bo Chen. PRIS in TREC 2008 Blog Track. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[144] Jamie Callan Hui Yang, Luo Si. Knowledge transfer and opinion detection in the trec2006 blog track. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[145] Robert A. Hummel and Steven W. Zucker. *On the foundations of relaxation labeling processes*, pages 585–605. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[146] Munawar Hussain. Language modeling based passage retrieval for question answering systems. Master's thesis, Saarland University, 2004.

[147] Charles J. III Infante, Dominic A.; Wigley. Verbal aggressiveness: An interpersonal model and measure. *Communications Monographs*, 53:61–69, 1986.

[148] Kelly Y. Itakura and Charles L. A. Clarke. University of waterloo at inex 2009: Ad hoc, book, entity ranking, and link-the-wiki tracks. In *Proceedings of INEX-2009*, December 2009.

[149] Spink A. & Saracevic T. Jansen, B. J. The use of relevance feedback on the web: Implications for web ir system design. In *Proceedings of WebNet World Conference on the WWW and Internet 1999*, pages 550–555. AACE, 1999.

[150] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[151] A. Java, P. Kolari, T. Finin A. Joshi, J. Martineau, and J. Mayfield. The blogvox opinion retrieval system. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[152] Frederick Jelinek. *Statistical methods for speech recognition*. MIT Press, Cambridge, MA, USA, 1997.

[153] Lifeng Jia, Clement Yu, and Wei Zhang. UIC at TREC 2008 blog track. In *Proceedings of the 17th Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[154] Jiepu Jiang, Wei Lu, Xianqian Rong, and Yangyan Gao. *CSIR at INEX 2008 Entity-Ranking Task: Entity Retrieval and Entity Relation Search in Wikipedia*, pages 254–270. Springer-Verlag, Berlin, Heidelberg, 2009.

[155] Xiaoyan Zhu Jianshu Sun, Chong Long and Minlie Huang. Mining reviews for product comparison and recommendation. *Research journal on Computer science and computer engineering with applications*, (39):33–40, 2009.

[156] Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 244–251, New York, NY, USA, 2006. ACM.

[157] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, pages 143–151, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[158] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[159] M. Jones and I. Alony. Blogs - the new source of data analysis. *Journal of Issues in Informing Science and Information Technology*, 5:433–446.

[160] Andreas Juffinger and Elisabeth Lex. Crosslanguage blog mining and trend visualisation. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1149–1150, New York, NY, USA, 2009. ACM.

[161] Anubhav Kale, Amit Kar, Pranam Kolari, Akshay Java, Tim Finin, and Anupam Joshi. Modeling trust and influence in the blogosphere using link polarity. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[162] J. Kamps, M. Marx, R. Mokken, and M. de Rijke. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1115–1118.

[163] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web (WWW-03)*, pages 640–651, New York, NY, USA, 2003. ACM.

[164] Noriko Kando. Overview of the sixth NTCIR workshop. In *Proceedings of NTCIR-6 Workshop Meeting*, 2007.

[165] Noriko Kando. Overview of the seventh ntcir workshop. In *Proceedings of NTCIR-7 Workshop Meeting*. ACM, 2008.

[166] Gjergji Kasneci, Fabian M. Suchanek, Georgiana Ifrim, Maya Ramanath, and Gerhard Weikum. NAGA: Searching and ranking knowledge. *International Conference on Data Engineering (ICDE)*, pages 953–962, 2008.

[167] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, USA, 1997. ACM.

[168] Marcin Kaszkiel and Justin Zobel. Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology*, 52(4):344–364, 2001.

[169] Shohei Sato Kazuhiro Seki, Yoshihiro Kino and Kuniaki Uehara. Trec 2007 blog track experiments at kobe university. In *Proceedings of the Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[170] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

[171] Ahmed Khorsi. An overview of content-based spam filtering techniques. *Informatica Ljubljana*, 31:269–278, 2007.

[172] Alejandro Valerio Kiduk Yang, Ning Yu and Hui Zhang. WIDIT in TREC-2006 blog track. In *Proceedings of the 15th Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[173] Hui Zhang Kiduk Yang, Ning Yu. WIDIT in TREC-2007 blog track: Combining lexicon-based methods to detect opinionated blogs. In *Proceedings of the Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[174] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, pages 1367–1373, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[175] Judith Klavans and Min-Yen Kan. Role of verbs in document analysis. In *Proceedings of the 17th international conference on Computational linguistics*, pages 680–686, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[176] David Kline and Dan Burstein. *Blog!: How the Newest Media Revolution is Changing Politics, Business, and Culture*. CDS Books, September.

[177] Pavel Serdyukov Paul Thomas Thijs Westerveld Krisztian Balog, Arjen P. de Vries. Overview of the trec 2009 entity track, trec 2009. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2008. ACM.

[178] L. W. Ku, Y. T. Liang, and H. H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*.

[179] Lun-Wei Ku and Hsin-Hsi Chen. Mining opinions from the web: Beyond relevance retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 58(12):1838–1850, 2007.

[180] T. Kudo and Y. Matsumoto. A boosting algorithm for classification of semi-structured text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 301–308, Barcelona , Spain, 2004.

[181] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM.

[182] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Structure and evolution of blogspace. *Commun. ACM*, 47(12):35–39, 2004.

[183] J. Lafferty and C. Zhai. *Probabilistic Relevance Models Based on Document and Query Generation*. Kluwer International Series on Information Retrieval.

[184] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119, New York, NY, USA, 2001. ACM.

[185] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33.

[186] S. Le Cessie and JC Van Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):191–201, 1992.

[187] Adrienne Lehrer. *Semantic fields and lexical structure*. North-Holland, American Elsevier, Amsterdam, New York, 1974.

[188] Mohsen Lesani and Saeed Bagheri. Fuzzy trust inference in trust graphs and its application in semantic web social networks. In *Proceedings of World Automation Congress (WAC-06)*, pages 1–6, Budapest, 2006. IEEE.

[189] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, New York, NY, USA, 1986. ACM.

[190] Raph Levien and Alexander Aiken. Attack-resistant trust metrics for public key certification. In *Proceedings of the 7th conference on USENIX Security Symposium - Volume 7*, pages 18–18, Berkeley, CA, USA, 1998. USENIX Association.

[191] Xiaonan Li, Chengkai Li, and Cong Yu. Entity-relationship queries over wikipedia. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, SMUC '10, pages 21–28, New York, NY, USA, 2010. ACM.

[192] Xiaoyan Li and W. Bruce Croft. Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management (CIKM-03)*, pages 469–475, New York, NY, USA, 2003. ACM.

[193] Xiaoyan Li and W. Bruce Croft. Novelty detection based on sentence level patterns. In *CIKM*, pages 744–751, 2005.

[194] Xiaoyan Li and W. Bruce Croft. Improving novelty detection for general topics using sentence level information patterns. In *CIKM*, pages 238–247, 2006.

[195] Xiaoyan Li and W. Bruce Croft. An information-pattern-based approach to novelty detection. *Inf. Process. Manage.*, 44(3):1159–1188, 2008.

[196] Cao D. Tan S. Liu Y. Ding G. Liao, X. and Cheng X. Combining language model with sentiment analysis for opinion retrieval of blog-post. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[197] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA, 1998. Association for Computational Linguistics.

[198] Kevin Hsin-Yih Lin and Hsin-Hsi Chen. NTU at TREC 2007 blog track. In *Proceedings of the Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[199] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 109–116, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[200] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, and Belle L. Tseng. Blog community discovery and evolution based on mutual awareness expansion. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 48–56, Washington, DC, USA, 2007. IEEE Computer Society.

[201] Gumwon Hong Joo-Young Lee Linh Hoang, Seung-Wook Lee and Hae-Chang Rim. A hybrid method for opinion finding task. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[202] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 1st ed. 2007. corr. 2nd printing edition, January.

[203] Bing Liu. Opinion mining and summarization-sentiment analysis. In *Tutorial in the Proceedings of WWW'08*. ACM.

[204] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of WWW '05, the 14th international conference on World Wide Web*, pages 342–351, Chiba, JP, 2005. ACM Press.

[205] X. Liu and W. B. Croft. Passage retrieval based on language models. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382, New York, NY, USA, 2002. ACM.

[206] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[207] Arevalo M., Carreras X., Marquez L., Marti M. A., Padro L., and Simon M. J. A proposal for wide-coverage spanish named entity recognition. *Sociedad Espanola para el Procesamiento del Lenguaje Natural*, 28:63–80, 2002.

[208] K. Balog M. Bron and M. de Rijke. Related entity finding based on co-occurance. In *Proceedings of the Text Retrieval Conference (TREC-2009)*, USA, 2009. National Institute of Standards and Technology (NIST).

[209] C. Macdonald and I. Ounis. Voting techniques for expert search. *JKnowledge and Information Systems*, 16:259–280, 2008.

[210] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC blog track 2009. In *The Eighteenth Text REtrieval Conference, (TREC 2009) Proceedings*, 2009.

[211] Craig Macdonald, David Hannah, and Iadh Ounis. High quality expertise evidence for expert search. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 283–295, Berlin, Heidelberg, 2008. Springer-Verlag.

[212] Craig Macdonald and Iadh Ounis. The trec blog06 collection : Creating and analysing a blog test collection. In *DCS Technical Report TR-2006-224*, 2006.

[213] Bernardo Magnini and Gabriela Cavaglià. Integrating subject field codes into wordnet. In *Proceedings of International Conference on Language Resources & Evaluation*, pages 1413–1418, 2000.

[214] Stuart Watt David Harper Malcolm Clark, Ulises Cervino Beresi. Rgu at the trec blog track. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[215] Thomas Mandl. Recent developments in the evaluation of information retrieval systems: Moving toward diversity and practical applications. 2008.

[216] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[217] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*, chapter Scoring, Term Weighting and the Vector Space Model. Cambridge University Press, New York, NY, USA, 2008.

[218] Christopher D. Manning and Heinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[219] Gary Marchionini. Interfaces for end-user information seeking. *Journal of the American Society for Information Science*, 43:156–163, 1992.

[220] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7:216–244, July 1960.

[221] Braschler Martin and Peters Carol. Cross-language evaluation forum: Objectives, results, achievements. *Information retrieval*, 7:7–31, 2004.

[222] J. R. Martin and Peter R. R. white. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan, Basingstoke, 2005.

[223] Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In Tu Bao Ho, David Cheung, and Huan Liu, editors, *Advances in Knowledge Discovery and Data Mining*, volume 3518 of *Lecture Notes in Computer Science*, pages 301–311. Springer Berlin / Heidelberg.

[224] Manya Mayes. On text data quality. In *http://blogs.sas.com/text-mining/index.php?/archives/48-On-Text-Data-Quality.html*. The Text Frontier, 2009.

[225] Frank McSherry and Marc Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 414–421. Springer, 2008.

[226] Edgar Meij, Peter Mika, and Hugo Zaragoza. An evaluation of entity and frequency based query completion methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 678–679, 2009.

[227] Juan J. Merelo, Beatriz Prieto, and Fernando Tricas. Blogosphere community formation, structure and visualization abstract. In *The European Conference on Webblogs, BlogTalks*, Vienna, 2004.

[228] Claire Cain Miller. How Obamas internet campaign changed politics. In *http://bits.blogs.nytimes.com/2008/11/07/how-obamas-internet-campaign-changed-politics/*. Nytimes.com, 2008.

[229] David R. H. Miller, Tim Leek, and Richard M. Schwartz. A hidden markov model information retrieval system. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 214–221, New York, NY, USA, 1999. ACM.

[230] George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[231] Gilad Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[232] Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. In *In Third annual workshop on the Weblogging ecosystem*, 2006.

[233] Gilad Mishne and Maarten Rijke. A study of blog search. In *Proceedings of 28th European Conference on Information Retrieval (ECIR*, pages 289–301. Springer-Verlag, 2006.

[234] Malik Muhammad Saad Missen, Faiza Belbachir, Guillaume Cabanac, and Mohand Boughanem. Using a mixed approach for opinion detection in blogs. In *Unknown*, XXXXX, XXX 2010. XXXX.

[235] Malik Muhammad Saad Missen and Mohand Boughanem. Sentence-Level Opinion-Topic Association for Opinion Detection in Blogs. In *IEEE International Symposium on Mining and Web, Bradford, UK*, pages 733–737. IEEE Computer Society, 2009.

[236] Malik Muhammad Saad Missen and Mohand Boughanem. Using WordNet's Semantic Relations for Opinion Detection in Blogs. In Mohand Boughanem, editor, *European Conference on Information Retrieval - Poster Session (ECIR), Toulouse*, number 5478 in LNCS, pages 729–733. Springer-Verlag, 2009.

[237] Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. Challenges for Sentence Level Opinion Detection in Blogs. In *International Conference on Computer and Information Science (ICIS), Shanghai, China*, pages 347–351. IEEE Computer Society, juin 2009.

[238] Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. Comparing Semantic Associations in Sentences and Paragraphs for Opinion Detection in Blogs. In *ACM student workshop on Management of Emergent Digital EcoSystems (MEDES-SW), Lyon, France*, pages 483–488. ACM, octobre 2009.

[239] Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. Opinion Detection in Blogs: What is still Missing? In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Odense, Denmark*. IEEE Computer Society, August 2010.

[240] Malik Muhammad Saad Missen, Mohand Boughanem, and Guillaume Cabanac. Opinion Finding in Blogs: A Passage-Based Language Modeling Approach. In *International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO), Paris, France*, page (electronic medium). Centre de hautes etudes internationales d'Informatique Documentaire (C.I.D.), avril 2010.

[241] Elke Mittendorf and Peter Schäuble. Document and passage retrieval based on hidden markov models. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information*

retrieval, pages 318–327, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[242] Audris Mockus and James D. Herbsleb. Expertise browser: a quantitative approach to identifying expertise. In *Proceedings of the 24th International Conference on Software Engineering*, ICSE '02, pages 503–512, New York, NY, USA, 2002. ACM.

[243] Dan Moldovan and Adrian Novischi. Word sense disambiguation of wordnet glosses. *Journal of Computer Speech & Language*, 18(3):301 – 317, 2004.

[244] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 412–418, Barcelona, Spain, July. Association for Computational Linguistics.

[245] Vanessa Murdock and W. Bruce Croft. A translation model for sentence retrieval. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 684–691, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[246] Jin-Cheon Na, Haiyang Sui, Christopher Khoo, Syin Chan, and Yunyun Zhou. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In *Conference of the International Society for Knowledge Organization (ISKO)*, pages 49–54, 2004.

[247] Seung-Hoon Na, Yeha Lee, Sang-Hyob Nam, and Jong-Hyeok Lee. Improving opinion retrieval based on query-specific sentiment lexicon. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 734–738, Berlin, Heidelberg, 2009. Springer-Verlag.

[248] S.H. Na, I.S. Kang, Y.H. Lee, and J.H. Lee. Applying completely-arbitrary passage for pseudo-relevance feedback in language modeling approach. In *AIRS'08: Proceedings of the 4th Asia information retrieval conference on Information retrieval technology*, pages 626–631, Berlin, Heidelberg, 2008. Springer-Verlag.

[249] T. Nasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing. In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, New York, NY, USA, 2003. ACM.

[250] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: capturing favorability using natural language processing. In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77, New York, NY, USA, 2003. ACM.

[251] Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 611–618, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[252] Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of COLING/ACL 2006 Main Conference Poster Sessions*, pages 611–618, 2006.

[253] David M. Nichols. Implicit rating and filtering. In *Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering*, pages 31–36, 1998.

[254] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. volume 39, pages 103–134, Hingham, MA, USA, May 2000. Kluwer Academic Publishers.

[255] Stephanie Nilsson. The function of language to facilitate and maintain social networks in research weblogs. Master's thesis, Umea Universitet, Engelska lingvistik, 2004.

[256] D. W. Oard, B. Hedin, S. Tomlinson, and J. R. Baron. Overview of the trec 2008 legal track. In *TREC*, 2008.

[257] Douglas W. Oard and Jinmook Kim. Modeling information content using observable behavior, 2001.

[258] Office of Fair Trading. Findings from consumer surveys on internet shopping. In `http://www.oft.gov.uk/shared_oft/reports/Evaluating-OFTs-work/oft1079.pdf`. Office of Fair Trading, 2009.

[259] D. Osman, J. Yearwood, and P. Vamplew. Automated opinion detection: Implications of the level of agreement between human raters. *Journal of Information Processing and Management (IPM)*, 46:331–342, 2009.

[260] Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, and Ian Sobo-roff. Overview of the TREC blog track 2006. In *The 15th Text REtrieval Conference, (TREC 2006) Proceedings*, 2006.

[261] Sara Owsley, Sanjay Sood, and Kristian J. Hammond. Domain specific affective classification of documents. In *Proceedings of the AAAI-CAAW06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, pages 181–183, 2006.

[262] Bo Pang and Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Mor-ristown, NJ, USA, 2004. Association for Computational Linguistics.

[263] Bo Pang and Lillian Lee. Seeing stars: exploiting class relationships for senti-ment categorization with respect to rating scales. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, Morristown, NJ, USA, 2005. Association for Computational Linguis-tics.

[264] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Journal of Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[265] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language process-ing*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[266] Marius Pasca. Finding instance names and alternative glosses on the web: Wordnet reloaded. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 3406 of *Lecture Notes in Computer Science*, pages 280–292. Springer Berlin / Heidelberg.

[267] T. Pedersen and S. Patwardhan. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-2004)*, pages 1024–1025, San Jose, CA, 2004.

[268] Jovan Pehcevski, Anne-Marie Vercoustre, and James A. Thom. Exploiting Lo-cality of Wikipedia Links in Entity Ranking. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in*

Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings, volume 4956 of *Lecture Notes in Computer Science*, pages 258–269. Springer, 2008.

[269] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM-07)*, pages 731–740, New York, NY, USA, 2007. ACM.

[270] Desislava Petkova and W. Bruce Croft. Hierarchical language models for expert finding in enterprise corpora. In *Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '06, pages 599–608, 2006.

[271] Jakub Piskorski, Marcin Sydow, and Dawid Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, AIRWeb '08, pages 25–28, 2008.

[272] Manolis Platakis, Dimitrios Kotsakos, and Dimitrios Gunopulos. Searching for events in the blogosphere. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1225–1226, New York, NY, USA, 2009. ACM.

[273] Polanyi, Livia and Zaenen, Annie. Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications*, 20:1–10.

[274] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM.

[275] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[276] Borislav Popov, Atanas Kiryakov, Dimitar Manov, Angel Kirilov, and Ognyanoff M. Goranov. Towards Semantic Web Information Extraction. In *In 2nd International Semantic Web Conference: Workshop on Human Language Technology for the Semantic Web and Web Services*, volume 20, 2003.

[277] M. F. Porter. *An algorithm for suffix stripping*, pages 313–316. Morgan Kauf-
mann Publishers Inc., San Francisco, CA, USA, 1980.

[278] M. Pucher. Performance evaluation of wordnet-based semantic relatedness mea-
sures for word prediction in conversational speech. In *Proceedings of Sixth Inter-
national Workshop on Computational Semantics*, Tilburg, Netherlands, 2004.

[279] Lide Wu Xuanjing Huang Qi Zhang, Bingqing Wang. FDU at TREC-2007:
Opinion retrieval of blog track. In *Proceedings of the Text Retrieval Confer-
ence (TREC-2007)*, USA, 2007. National Institute of Standards and Technology
(NIST).

[280] W Quick. Dailypundit.com. In `www.iw3p.com/DailyPundit/2001_12_`
`30_dailypundit_archive.php#8315120`. www.iw3p.com, 2001.

[281] I. Ounis J. Peng R. McCreadie, C. Macdonald and R. L. T. Santos. University
of glasgow at trec 2009: experiments with terrier. In *Proceedings of the Text
Retrieval Conference (TREC-2009)*, USA, 2009. National Institute of Standards
and Technology (NIST).

[282] Hema Raghavan, James Allan, and Andrew Mccallum. An exploration of entity
models, collective classification and relation description. In *Proceedings of KDD
Workshop on Link Analysis and Group Detection*, pages 33–3, 2004.

[283] Robert Lothian Rahman Mukras, Nirmalie Wiratunga. The robert gordon uni-
versity at the opinion retrieval task of the 2007 trec blog track. In *Proceedings
of the 16th Text Retrieval Conference (TREC-2007)*, USA, 2007. National In-
stitute of Standards and Technology (NIST).

[284] Jonathon Read. Using emoticons to reduce dependency in machine learning
techniques for sentiment classification. In *ACLstudent '05: Proceedings of the
ACL Student Research Workshop*, pages 43–48, Morristown, NJ, USA, 2005.
Association for Computational Linguistics.

[285] Olivier ReichenStein. The future of news - how to survive the new media shift.
The American Political Science Review.

[286] Akamai Quarterly Report. State of the internet: Q-3 2009 report. In
http://www.akamai.com/stateoftheinternet/. AKAMAI, 2009.

[287] Arjen de Vries Jaap Kamps Rianne Kaptein, Pavel Serdyukov. Entity ranking
using wikipedia as a pivot. In *Proceedings of international Conference on Infor-*

mation and Knowledge Management (CIKM-2010), Toronto, Ontario, Canada, 2010. ACM.

[288] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 1979.

[289] Ellen Riloff, Janyce Wiebe, and William Phillips. Exploiting subjectivity classification to improve information extraction. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 3*, pages 1106–1111. AAAI Press, 2005.

[290] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 25–32, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[291] Madhu RM, Srikant R, Venkatesh Karthik S, Meenakshi Sundaram Murugeshan, and Saswati Mukherjee. A recursive approach to entity ranking and list completion using entity determining terms , qualifiers and prominent n-grams. In *Proceedings of INEX-2009*, December 2009.

[292] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, (4), 2009.

[293] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. Probabilistic models of indexing and searching. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, SIGIR '80, pages 35–56, Kent, UK, UK, 1981. Butterworth & Co.

[294] Stephen E. Robertson and Karen Sparck Jones. *Relevance weighting of search terms*, pages 143–160. Taylor Graham Publishing, London, UK, UK, 1988.

[295] Sudeshna Sarkar Robin Anil. IIT kharagpur at TREC-2008 blog track. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[296] J. J Rocchio. *The SMART Retrieval System - Experiments in Automatic Document Processing*, chapter Relevance feedback in information retrieval: The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[297] Henning Rode, Pavel Serdyukov, Djoerd Hiemstra, and Hugo Zaragoza. Entity ranking on graphs: Studies on expert finding. Technical report, University of Twente, Enschede, 2007.

[298] Everett M. Rogers and Everett Rogers. *Diffusion of Innovations, 5th Edition.* Free Press, original edition, August.

[299] James A. Russell. Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, 45(6):1281 – 1288, 1983.

[300] Robertson S. and Walker S. Some simple effective approximations to the 2 poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, New York, NY, USA, 1994. ACM/Springer Verlag.

[301] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, November 1975.

[302] Gerard. Salton. *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968.

[303] Gerard Salton, J. Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58, New York, NY, USA, 1993. ACM.

[304] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York u.a., 1983.

[305] Franco Salvetti, Stephen Lewis, and Christoph Reichenbach. Automatic opinion polarity classification of movie reviews. *Colorado Research in Linguistics*, 17(1), 2004.

[306] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[307] Rodrygo L. T. Santos, Ben He, Craig Macdonald, and Iadh Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 325–336, Berlin, Heidelberg, 2009. Springer-Verlag.

[308] Rodrygo L. T. Santos, Ben He, Craig Macdonald, and Iadh Ounis. Integrating proximity to subjective sentences for blog opinion retrieval. In *ECIR '09: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 325–336, Berlin, Heidelberg, 2009. Springer-Verlag.

[309] Robert E. Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. *Journal of Machine Learning*, (2/3):135–168.

[310] Jan Schmidt. Blogging practices: An analytical framework. *Journal of Computer-Mediated Communication*, (4):1409–1427.

[311] Doc Searls and David Sifry. Building with blogs. In *http://www.linuxjournal.com/article/6497*. Linux Journal, 2003.

[312] And Fabrizio Sebastiani, Andrea Esuli, and Fabrizio Sebastiani. Determining term subjectivity and term orientation for opinion mining andrea esuli. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL06)*, 2006.

[313] Irene Sege. Where everybody knows your name. In `http://www.boston.com/news/globe/living/articles/2005/04/27/where_everybody_knows_your_name/`. Boston.com, 2005.

[314] Kazuhiro Seki and Kuniaki Uehara. Adaptive subjective triggers for opinionated document retrieval. In *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 25–33, New York, NY, USA, 2009. ACM.

[315] Wei Lu SHaozhen Zhao?Zhicheng Luo. WHU at TREC blog track 2007. In *Proceedings of the 16th Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[316] Mark Carman Robert Gwadera Davide Taibi Shima Gerani, Mostafa Keikha and Fabio Crestani. University of lugano at TREC-2008 blog track. In *Proceedings of the 17th Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[317] B. Shin. An exploratory investigation of system success factors in data warehousing. *Journal of the Association for Information Systems (JAIS)*, 4:141–170, 2003.

[318] M. de Rijke Sisay Fissaha Adafre and E. Tjong Kim Sang. Entity retrieval. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-07)*, 2007.

[319] Ian Soboro, , Ian Soboro, and Donna Harman. Overview of the trec 2003 novelty track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003*, 2003.

[320] Ian Soboroff. Overview of the trec 2004 novelty track. In Ellen M. Voorhees and Lori P. Buckland, editors, *TREC*, volume Special Publication 500-261. National Institute of Standards and Technology (NIST), 2004.

[321] Ian Soboroff and Nick Craswell. Overview of the TREC 2006 enterprise track. In *Proceedings of TREC-06*. National Institute of Standards and Technology (NIST), 2006.

[322] Ian Soboroff and Donna Harman. Overview of the trec 2003 novelty track. In *TREC*, pages 38–53, 2003.

[323] Ian Soboroff and Donna Harman. Novelty detection: the trec experience. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[324] Fei Song and W. Bruce Croft. A general language model for information retrieval. In *CIKM '99: Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321, New York, NY, USA, 1999. ACM.

[325] Xiaodan Song, Yun Chi, Koji Hino, and Belle Tseng. Identifying opinion leaders in the blogosphere. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM-07, pages 971–974, 2007.

[326] Daming Shi Hongfei Lin Zhihao Yang Song Rui, Tang Qin. DUTIR at TREC-2007 blog track. In *Proceedings of the 16th Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[327] K. Sparck Jones and C. J. Van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32:59–75, 1976.

[328] Ellen Spertus. Smokey: Automatic recognition of hostile messages. In *Proceedings of IAAI-97, the 9th Conference on Innovative Application o f Artificial Intelligence*, pages 1058–1065, Providence, US, 1997.

[329] Susan Jones Micheline Hancock-Beaulieu Stephen E. Robertson, Steve Walker and Mike Gatford. Okapi at TREC-3. In *TREC*. National Institute of Standards and Technlogy (NIST), 1994.

[330] Michael Stephens. The ongoing Web revolution. *Library Technology Reports*, 43(5):10+, 2007.

[331] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

[332] Philip J. Stone and Earl B. Hunt. A computer approach to content analysis: studies using the general inquirer system. In *AFIPS '63 (Spring): Proceedings of the May 21-23, 1963, spring joint computer conference*, pages 241–256, New York, NY, USA, 1963. ACM.

[333] Stephanie Strassel, Alexis Mitchell, and Shudong Huang. Multilingual resources for entity extraction. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition - Volume 15*, MultiNER '03, pages 49–56, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[334] Pero Subasic and Alison Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transaction on Fuzzy Systems*, 9(4):483–496, 2001.

[335] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.

[336] Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. Application of semi-supervised learning to evaluative expression classification. In *Proceedings of CICLing-06, the 7th international conference on Computational Linguistics and Intelligent Text Processing*, pages 502–513, Mexico City, MX.

[337] Huifeng Tang, Songbo Tan, and Xueqi Cheng. A survey on sentiment detection of reviews. *Expert Systems with Application*, 36(7):10760–10773, 2009.

[338] Junichi Tatemura. Virtual reviewers for collaborative exploration of movie reviews. In *IUI '00: Proceedings of the 5th international conference on Intelligent user interfaces*, pages 272–275, New York, NY, USA, 2000. ACM.

[339] Technorati. The state of the blogosphere. In *technorati.com/state-of-the-blogosphere/*, 2006.

[340] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47, New York, NY, USA, 2003. ACM.

[341] J. A. Thom, J. Pehcevski, and A. M. Vercoustre. Use of Wikipedia Categories in Entity Ranking. *Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia*, 2007.

[342] Ryoko Tokuhisa and Ryuta Terashima. Relationship between utterances and "enthusiasm" in non-task-oriented conversational dialogue. In *SigDIAL '06: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 161–167, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[343] R Tong. An operational system for detecting and tracking opinions in on-line discussion. In *SIGIR Workshop on Operational Text Classification (SIGIR-2001)*, pages 1–6, 2001.

[344] Theodora Tsikrika, Pavel Serdyukov, Henning Rode, Thijs Westerveld, Robin Aly, Djoerd Hiemstra, and Arjen P. Vries. Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking Using PF/Tijah. In *Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007 Dagstuhl Castle, Germany, December 17-19, 2007. Selected Papers*, pages 306–320, Berlin, Heidelberg, 2008. Springer-Verlag.

[345] R. Tumarkin and R. Whitelaw. News or noise? internet postings and stock prices. *Financial Analysts Journal*, (3):41–51, May.

[346] Peter Turney, Council Canada, and Michael Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical report, Institute for Information Technology, Technical Report ERB-1094, 2002.

[347] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[348] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions for Information Systems*, 21(4):315–346, 2003.

[349] Ro Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086, 2004.

[350] David Vallet and Hugo Zaragoza. Inferring the most important types of a query: a semantic approach. In *Proceedings of the 31st annual international*

ACM SIGIR conference on Research and development in information retrieval,
SIGIR-08, pages 857–858, 2008.

[351] Olga Vechtomova. Using subjective adjectives in opinion retrieval from blogs.
In *Proceedings of the Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[352] Paola Velardi, Paolo Fabriani, and Michele Missikoff. Using text processing
techniques to automatically enrich a domain ontology. In *Proceedings of the
international conference on Formal Ontology in Information Systems - Volume
2001*, FOIS '01, pages 270–284, New York, NY, USA, 2001. ACM.

[353] A.M. Vercoustre, J. A. Thom, and J. Pehcevski. Entity ranking in wikipedia.
In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*,
pages 1101–1106, New York, NY, USA, 2008. ACM.

[354] Anne-Marie Vercoustre, Jovan Pehcevski, and Vladimir Naumovski. Topic Difficulty Prediction in Entity Ranking. In *Advances in Focused Retrieval: 7th
International Workshop of the Initiative for the Evaluation of XML Retrieval,
INEX 2008, Dagstuhl Castle, Germany, December 15-18, 2008. Revised and
Selected Papers*, pages 280–291, Berlin, Heidelberg, 2009. Springer-Verlag.

[355] Anne-Marie Vercoustre, Jovan Pehcevski, and James A. Thom. Focused access
to XML documents. chapter Using Wikipedia Categories and Links in Entity
Ranking, pages 321–335. Springer-Verlag, Berlin, Heidelberg, 2008.

[356] Anne-Marie Vercoustre, James A. Thom, and Jovan Pehcevski. Entity ranking
in wikipedia. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied
computing*, pages 1101–1106, New York, NY, USA, 2008. ACM.

[357] Steven H. Chaffee Vincent Price. *Public Opinion*. SAGE Publications, June.

[358] E. Voorhees. Common evaluation measures. In *Proceedings of the 16th Text
Retrieval Conference (TREC-2007)*, USA, 2007. NIST TREC.

[359] E. Voorhees. Common evaluation measures. In *Proceedings of the 16th Text
REtrieval Conference (TREC-2007)*, volume 500-274. NIST Special Publication, 2008.

[360] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation
in Information Retrieval*. Digital Libraries and Electronic Publishing. MIT
Press, September.

[361] Arjen P. Vries, Anne-Marie Vercoustre, James A. Thom, Nick Craswell, and Mounia Lalmas. Overview of the INEX 2007 entity ranking track. pages 245–251, 2008.

[362] Annika Waern. Rosalind picard: Affective computing. *User Modeling and User-Adapted Interaction*, 12:85–89.

[363] Jianqiang Wang, Ying Sun, Omar Mukhtar, and Rohini Srihari. TREC 2008 at the university at buffalo: Legal and blog track. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[364] Ian Ward and James Cahill. Old and new media: Blogs in the third age of political communication. *Australian Journal of Communication*, 34(3):1–21, 2007.

[365] W. Weerkamp and M de Rijke. External query expansion in the blogosphere. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[366] Wouter Weerkamp, Jiyin He, Krisztian Balog, and Edgar Meij. *The University of Amsterdam (ILPS) at INEX 2008*, pages 276–286. Springer-Verlag, Berlin, Heidelberg, 2009.

[367] Clement Yu Wei Zhang. Uic at trec 2006 blog track. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[368] Clement Yu Wei Zhang. UIC at TREC 2007 blog track. In *Proceedings of the 16th Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[369] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using Appraisal Taxonomies for Sentiment Analysis. In *Proceedings of MCLC-05, the 2nd Midwest Computational Linguistic Colloquium*, Columbus, US, 2005.

[370] Janyce Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740. AAAI Press, 2000.

[371] Janyce Wiebe and Claire Cardie. *Language Resources and Evaluation (formerly Computers and the Humanities*, volume 39, chapter Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, pages 165–210. 2005.

[372] Janyce Wiebe, Theresa Wilson, and Matthew Bell. Identifying collocations for recognizing opinions. In *Proceedings of ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, pages 24–31, 2001.

[373] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Comput. Linguist.*, 30(3):277–308, 2004.

[374] Janyce M. Wiebe. Identifying subjective characters in narrative. In *Proceedings of the 13th conference on Computational linguistics*, pages 401–406, Morristown, NJ, USA, 1990. Association for Computational Linguistics.

[375] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[376] Yorick Wilks and Janusz Bien. Beliefs, points of view and multiple environments. In *Proceedings of the international NATO symposium on Artificial and human intelligence*, pages 147–171, New York, NY, USA, 1984. Elsevier North-Holland, Inc.

[377] Gbolahan K. Williams and Sarabjot Singh Anand. Predicting the polarity strength of adjectives using wordnet. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM-09)*, 2009.

[378] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[379] Theresa Wilson, David R. Pierce, and Janyce Wiebe. Identifying opinionated sentences. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 33–34, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[380] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[381] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Just how mad are you? finding strong and weak opinion clauses. In *AAAI'04: Proceedings of the 19th national conference on Artifical intelligence*, pages 761–767. AAAI Press, 2004.

[382] Theresa Ann Wilson. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states.* PhD thesis, Pittsburgh, PA, USA, 2008.

[383] D Winer. What makes a weblog a weblog? *Published by: Weblogs at Harvard Law.*

[384] E Airoldi X Bai, R Padman. Sentiment Extraction from Unstructured Text using Tabu Search-Enhanced Markov Blanket. In *Workshop on Mining the Semantic Web 10th ACM SIGKDD Conference (2004)*, pages 24–35, 2007.

[385] Yu Wang Wei Liu Songbo Tan Hongbo Xu Xiangwen Liao, Donglin Cao and Xueqi Cheng. Experiments in trec 2007 blog opinion task at cas-ict. In *Proceedings of the Text Retrieval Conference (TREC-2007)*, USA, 2007. National Institute of Standards and Technology (NIST).

[386] Z. Yu Y. Xian Y. Fang, L. Si and Y. Xu. Entity retrieval with hierarchical relevance model. In *Proceedings of the Text Retrieval Conference (TREC-2009)*, USA, 2009. National Institute of Standards and Technology (NIST).

[387] Kiduk Yang. WIDIT in TREC-2008 blog track: Leveraging multiple sources of opinion evidence. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[388] Jungi Kim Sang-Hyob Nam Hun-young Jung Jong-Hyeok Lee Yeha Lee, Seung-Hoon Na. Kle at trec 2008 blog track: Blog post and feed retrieval. In *Proceedings of the Text Retrieval Conference (TREC-2008)*, USA, 2008. National Institute of Standards and Technology (NIST).

[389] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the Third IEEE International*

Conference on Data Mining (ICDM-03), pages 427–434, Washington, DC, USA, 2003. IEEE Computer Society.

[390] Zhengtao Yu Yantuan Xian Yi Fang, Luo Si and Yangbo Xu. Entity retrieval with hierarchical relevance model. In *The Eighteenth Text REtrieval Conference, (TREC 2009) Proceedings*, 2009.

[391] Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111, New York, NY, USA, 2006. ACM.

[392] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, New York, NY, USA, 2008. ACM.

[393] Dawit Yimam-Seid and K. O. B. S. Alfred Kobsa. Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. *Journal of Organizational Computing and Electronic Commerce*, 13:1–24, 2003.

[394] Le Sun Hsin-Hsi Chen Yohei Seki, Lun-Wei Ku and Noriko Kando. Overview of multilingual opinion analysis task at NTCIR-8. In *Proceedings of NTCIR-8 Workshop Meeting*, 2010.

[395] Lun-Wei Ku-Le Sun Hsin-Hsi Chen Yohei Seki, David Kirk Evans and Noriko Kando. Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of NTCIR-7 Workshop Meeting*, 2008.

[396] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP-03*, pages 129–136, 2003.

[397] Hugo Zaragoza, Henning Rode, Peter Mika, Jordi Atserias, Massimiliano Ciaramita, and Giuseppe Attardi. Ranking very many typed entities on wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 1015–1018, 2007.

[398] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th*

annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR-01, pages 334–342, 2001.

[399] V.G.Vinod Vydiswaran; Kavita Ganesan; Yuanhua Lv; ChengXiang Zhai. Finding related entities by retrieving relations: UIUC at TREC 2009 entity track. In *Proceedings of Text Retrieval Conference (TREC) (2009)*. National Institute of Standards and Technology (NIST), 2009.

[400] Ethan Zhang and Yi Zhang. Ucsc on trec 2006 blog opinion mining. In *Proceedings of the Text Retrieval Conference (TREC-2006)*, USA, 2006. National Institute of Standards and Technology (NIST).

[401] Tong Zhang and David Johnson. A robust risk minimization based named entity recognition system. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 204–207, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[402] Wei Zhang, Lifeng Jia, Clement Yu, and Weiyi Meng. Improve the effectiveness of the opinion retrieval and opinion polarity classification. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1415–1416, New York, NY, USA, 2008. ACM.

[403] Yi Zhang and Flora S. Tsai. Combining named entities and tags for novel sentence detection. In *ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 30–34, New York, NY, USA, 2009. ACM.

[404] Ziqiong Zhang, Qiang Ye, Rob Law, and Yijun Li. Automatic detection of subjective sentences based on chinese subjective patterns. In Yong Shi, Shouyang Wang, Yi Peng, Jianping Li, and Yong Zeng, editors, *Cutting-Edge Research Topics on Multiple Criteria Decision Making*, volume 35 of *Communications in Computer and Information Science*, pages 29–36. Springer Berlin Heidelberg.

[405] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 473–480, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[406] L. Zhou, D. P. Twitchell, T. Qin, J. K. Burgoon, and J. F. Nunamaker. An exploratory study into deception detection in text-based computer-mediated communication. *Hawaii International Conference on System Sciences*, 1:44b, 2003.

[407] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. Movie review mining and summarization. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 43–50, New York, NY, USA, 2006. ACM.

[408] Cai-Nicolas Ziegler and Georg Lausen. Spreading activation models for trust propagation. In *Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04)*, EEE '04, pages 83–97, 2004.