Fine-Grained Opinion Mining in Resource-Scarce Languages with Sentence-level Subjectivity Analysis

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science (by Research)
in
Computer Science and Engineering

by

M. Aditya
201107630
aditya.m@research.iiit.ac.in

Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2013

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **"Fine-Grained Opinion Mining in Resource-Scarce Languages with Sentence-level Subjectivity Analysis"** by *M. Aditya*, for the award of the Degree of Master of Science (by Research) is a record of bona-fide research work carried out by him under my supervision and guidance. The contents of this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Date

Advisor: Prof. Vasudeva Varma

To My Parents, Sister and Friends

Acknowledgments

I would like to take this opportunity to acknowledge all those people who have helped me during my sprint at Masters by Research.

First and foremost, I would like to dedicate this thesis to my advisor Prof. Vasudeva Varma, without whom i would not be able to accomplish my research aspirations. Thanks to the opportunity he provided, i was able to concentrate on research under his supervision after leaving my job without any financial constraints. His devoted effort in providing me everything that is requested show his commitment towards students. I was able to learn many things during the time i spent with him either during meetings, courses or informal conversations. His vision to achieve perfection has inculcated passion in me to do something brilliant. I am very glad that i got an opportunity to work with a person who is not only professionally acclaimed but also down to earth.

I would also like to thank my lab-mates Srikanth and Nikhil for spending most of the time with me on many discussions in the lab. I am very happy to be associated with Srikanth on many journeys for CLIA meetings. His insights into technology has given me opportunity to learn more about systems. His helping nature and winning attitude always inspired me. Nikhil helped me to improve personally. He also helped me to accomplish many things, without him i would have delayed them forever. I would also like to thank Bhupal, Mahender, Arpit, Ajay, Jayanth, Priya, Akshat, Piyush, Ankit and Krish for all the useful discussions and for the friendly environment in the lab. I thank Microsoft Research for providing travel grant to attend PACLIC with my friends Nitesh and Praveen in Indonesia.

I would like to offer gratitude to my B.Tech friends from IIIT Hyderabad. Though they graduated long back from the institute. They used to spend time in campus to motivate me for good or bad reasons. My best critic is Gayam Raja, without whom i would have never pushed myself very hard. His criticism has helped to push my limits to achieve things. Lova always been a helping hand and was there for me in any situation. Discussions with Bhanukiran helped me to think beyond science and most of the time our discussions ended on a philosophical note. Time spent with Bharath is like a roller coaster ride having many activities to accomplish. Karthik has always supported my thoughts and advised me on many personal and professional issues.

Last but not least, i would like to thank my parents and sister for belief in my actions and steps i took to achieve them.

Abstract

Identifying subjective text and extracting opinion information about different entities from News, Blogs and other forms of user generated content like social media has a lot of applications. We wish to address this challenging task of mining opinions about different entities like organizations, people and places from different languages. But, it is known that resources and tools for mining opinions are scarce in many languages.

In our research, we try to address resource scarcity problem by designing approaches which leverages resources from rich resource language like English for opinion mining. Named Entities, adjectives and verbs treated as possible opinion words are extracted from English in a comparable corpora. Possible opinions on targets (mainly people) represented by adjectives, verbs are translated and transliterated to resource-scarce languages. Initially, these transformed words are used to identify subjective sentences in resource-scarce languages. Though this approach is unsupervised, subjective and objective text classification at sentence-level can be improved using a supervised method. Accuracy of sentence-level subjectivity detection is further improved using language independent feature weighting and selection methods which are consistent across different languages. Once the subjective sentences are identified, they are used to create a subjective language model (SLM). While transliterated and translated words are used to form structured opinion queries (OQs) using inference network (IN) for retrieval to confirm the opinions about opinion targets in the documents.

Experiments are performed at sentence-level and fine-grain level using Hindi comparable corpora to evaluate the effectiveness of approaches in finding opinions. Even though experiments are performed on Hindi, approach can be easily extended to other resource-scarce languages. Results have shown that for fine-grained opinion mining, OQs and SLM with IN framework are effective for opinion mining tasks in minimal resource languages when compared to other retrieval approaches. While Entropy based category coverage difference criterion (ECCD) feature selection method with our proposed feature weighting methods along with Naive Bayes Multinomial classifer is consistent for sentence-level subjective classification in Hindi, Telugu and other languages including English and European languages.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Information growth on the web can be attributed to the contribution done by different individuals from diverse backgrounds. According to International Telecommunication Union(ITU)[1] there are around 1066 and 461 million Internet users alone in Asia Pacific and Europe respectively. Due to this humongous Internet penetration and the advancements made in Web 2.0 technology to support content in multiple languages. Most of these Internet users post content in their native languages on various sources on the web, making the web diverse with languages other than English.

Considering the case of India, main sources of Indian language content on the web are newspaper websites and blogs. Dainik Jagran[2] and Eeenadu[3] are one of many newspaper agencies which post news in Hindi and Telugu respectively. Indian language blogs have also seen growth due to popularity of phonetic transliteration based typing tools for Unicode Indic typing and are covering areas like entertainment, current affairs and politics. OneBlogs[4] is one such service which collocates Indian language blogs posted in different categories. Similarly, European Union has around 23 official languages[5] which has its presence on the web in form of blogs, newspapers, social media and review sites[6].

But, most of the information posted on web in multiple languages in Blogs, News and other forms of user generated content is not always factual. It is mixed with opinions, sentiments and beliefs of the user. For Example, News articles are written in different languages containing different entities like places, people or organizations. Factual information provided about these entities (mainly people) is sometimes added with support or expressions of the writer. This induces opinion about people in the article indirectly. These opinions can be present in various granularities such as a word, a sentence or a document. Similar observations can be made for Blogs, Social Media content and other discussion forums.

Identifying user opinions and subjectivity in the text has lot of applications. Mining opinions about people in news articles [38] will help us to distinguish the factual information and writers view on them.

[1] http://www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html
[2] http://www.jagran.com/
[3] http://www.eenadu.net/
[4] http://www.oneblogs.in/
[5] http://ec.europa.eu/languages/languages-of-europe/eu-languages_en.htm
[6] http://www.ciao.es/

Also, it will help us understand the bias of news agencies or writers towards them. Comparison of opinions about people mentioned in different news agencies articles can be done to see the authenticity of information. It also help us in tracking opinions about people over different time-lines. Some of the other useful natural language processing applications are mining opinions from product reviews [19], summarizing different opinions [54] in blogs, question answering [4] etc.

Initially, we try to understand how a subjective text is represented in a document and further leverage on it to mine opinions at different granularity.

1.1 Subjectivity Analysis

Subjectivity analysis identifies a linguistic expression which has somebody's opinions, sentiments, emotions, evaluations, beliefs or speculations known as private states [45]. It is a state that is not open to objective observation or verification. Subjectivity analysis classifies content into subjective or objective text. Subjectivity is observed at many levels in blogs, editorials, reviews (of products, movies, books, etc.) and newspaper articles. Subjectivity analysis can be either done at document-level or sentence-level.

1.1.1 Document-level

We understood that subjective text contain private states of the users, while objective text generally report facts. Documents containing user views like "Movie or Book" reviews are typically considered to have sentiment or beliefs at document-level and are classified for subjectivity at document-level. But, when we consider something like "Product or Hotel" reviews they provide user opinions on specific features of an object. So, focus may not be required on the entire review, but on few subjective sentences. Similar observations are required on blogs and news articles which has opinions expressed on different entities. Thus subjectivity analysis at document-level provide overall insight on the document but miss out some fine-grain observations.

1.1.2 Sentence-level

Document-level subjectivity classification is too coarse for most applications. Subjectivity analysis at sentence-level is less attempted due to common assumption that documents such as reviews needs to be distinguished as "subjective texts" and documents such as newspaper articles are distinguished as "objective texts". But in reality, most of the documents contain a mixture of both subjective and objective sentences. Sentence-level subjectivity analysis aims to identify subjective text at sentence-level. Finding subjective and objective sentences in news collections, blogs, product reviews, etc helps in various natural language processing applications.

1.2 Opinion Mining

Subjectivity classification at both document and sentence levels are useful, but they do not find what the opinion holder liked and disliked about a target. We need approaches which mine opinions at feature level. Opinion mining can be broadly divided into different areas like opinion retrieval, opinion question answering, opinion summarization etc. Major components of opinion mining are Opinion Holder [OH], Opinion Target [OT] and Opinion [OP]. Opinion Holder holds a opinion on a target.

1.2.1 Fine-Grain level

It is generally assumed that each document focuses on a single object and contains opinion from a single opinion holder. But it is not true in many cases. We need more fine-grain analysis to detect opinions. For Example, identifying different features of a product and opinions on them in product reviews, opinions expressed by different entities in newspaper articles and blogs, sentiment on different organizations and people in social media data etc. Although each granularity in a document is important, focus is on word-level opinion detection.

1.3 What are Resource-Scarce Languages

Even though web constitutes different language pages other than English. There are very few linguistic resources developed to process the content in these languages with high accuracy. For Example, obtaining transliterations from English to other languages, part of speech information tagging, named entity recognition are few such applications. Due to dearth of these resources, an immediate need is observed to port lot of applications which works well for English to these languages. But porting tools from resource rich languages like English is a non-trivial task due to numerous challenges. These issues creates scarcity of resources in these languages making them resource-scarce languages. South Asian and some European languages fall into category of resource-scarce languages.

1.4 Motivation

Internet penetration around the world has increased the content posted on the web. Especially, user generated information posted in blogs, news, social media content, etc in multiple languages. Mining this information can provide lot of insights about user sentiment or beliefs about products, services and people. But, mining this information from various languages need several tools and linguistic resources. English language has enjoyed rich resources due to the availability of funding and human resources. Several tools and linguistic resources are developed for English to perform natural language processing tasks like subjectivity analysis and opinion mining. But, resource-scarce languages lack state-of-the-art

resources and tools which are required for subjectivity and opinion mining making our research process slow.

This motivated us to design approaches that are less dependent on language specific resources and can be consistently used across resource-scarce languages.

1.5 Challenges

Languages other than English are of interest due to larger non-English speaking community. Understanding natural language written in multiple languages is a challenging task. Natural language applications designed for different languages are still evolving. In this context, opinion and subjectivity analysis in languages other than English also needs to overcome several challenges. These challenges can be divided into two categories.

1.5.1 Resources specific Challenges

- Opinion and subjectivity detection have lot of dependency on the linguistic resources. Substantial human efforts are required to create resources for many languages, as only few specific linguistic resources are available in particular languages.

- Scarcity of high precision state-of-the-art natural language processing tools like named entity recognizers, part of speech (POS) taggers, dependency parsers, semantic role labeling tools etc.

- Supervised techniques [33] which identify subjectivity at sentence-level require lot of labeled data.

- Investment or funding from industry required to create opportunities and to employ more human resources.

- Improve minuscule revenues by leveraging usage of resources by generating different opportunities.

1.5.2 Language specific Challenges

- Languages have different dialects. Generally, user generated information like blogs, reviews etc in different languages is embedded with contextual information. Identifying opinion or subjectivity information based on context information is non-trivial.

- Agglutinative languages create problems in decoding the word-level specific information. Opinion mining aims to understand opinions at word-level.

- Content in blogs, reviews etc on the web in european and indian languages are created using machine translation from English like Google Transliterate [7] etc. It adds lot of wrong variations of the words creating problems for subjectivity and opinion analysis.

- Informal and vague writing by the user can miss valuable information used for subjectivity and opinion detection.

1.6 Approaches

Sentence-level subjectivity analysis and fine-grain opinion mining was performed earlier on different languages using different methods. These approaches can be broadly divided into unsupervised and supervised methods.

1.6.1 Sentence-level Subjectivity Analysis

Subjectivity analysis was performed at document-level and sentence-level. Mainly, Wiebe [44] extracted subjective information at sentence-level for English. They concentrated on learning subjective and objective expressions in English at sentence-level using boot-strapping algorithms. Even though this approach is unsupervised it lacks scalability and is language dependent. Recently, focus shifted from English to multilingual space [5]. Banea [27] worked on sentence-level subjectivity analysis using machine translation approaches by leveraging resources and tools available for English. Another approach by Banea [26] used multilingual space and meta classifiers to build high precision classifiers for subjectivity classification.

However, aforementioned work [27] concentrated more on language specific attributes due to variation in expression of subjectivity in different languages. This creates a problem of portability of approaches to different languages. Other approach [26] which achieved language independence created large feature vectors for subjectivity classification. Different languages parallel sentences are taken into consideration to build high-precision classifier for each language. This approach not only increases the complexity and time for classification but also completely dependent on parallel corpus to get good accuracies.

An approach which tried to minimize the supervision has proposed a weakly supervised method [15] for sentence-level subjectivity detection using subjLDA. It reduced the usage of training data, but available only for English. There are several other approaches which was experimented on languages like Japanese [22], Chinese [53], Romanian [27, 5] are performed at document-level and are not language independent.

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

1.6.2 Fine-Grained Opinion Mining

Mining opinions from documents has been approached from several scenarios. Opinion retrieval is one such method, where a opinionated document is retrieved based on information need supplied by a query. Similarly, opinion summarization proposed in TAC 2008[8] aim to summarize answers based on opinions to the complex questions. Another task proposed in TAC 2008 aims to return different aspects of opinion (holder, target, support) with a particular polarity in response to opinion question. In our study, we aim to solve this problem by devising approaches to extract opinion targets and opinions expressed on them. Some of the similar approaches [16] attempted earlier used parsing and Maximum Entropy ranker methods based on parse features. However, this approach cannot be easily portable to resource-scarce languages due to low quality parsers. Another approach [17] used opinion word labels which are collected manually to find opinion-related frames (FrameNet) that are then semantic role labeled to identify fillers for the frames. However, we cannot apply this approach due to non-availability of semantic role labeling tools in resource-scarce languages.

1.7 Applications

In the pre-web era opinions were collected from friends, relatives or acquaintances for buying products, reviews on movies or books etc. Explosion of web 2.0 has made the process of gathering opinions scalable. Lot of consumers rely on Internet to make informed decisions. Some of the applications of opinion and subjectivity analysis which are thought to benefit businesses and individuals are listed below.

1.7.1 Businesses and Organizations

Businesses and organizations spend lot of time and money to know their consumers. They are interested in knowing the opinions on their and competitor products and services. To achieve it they conduct surveys and do market research. But each of these approaches are time consuming and explicitly require consumers participation. It is very helpful for a company to automatically extract consumer intents from public forums like Blogs, review sites etc which enable them to spend less expenditure on their market research activities. They can also find comparative opinions in these forums for related products or services. It always help them to get indirect feedback.

1.7.2 Individuals

Individuals are interested in finding peers opinion when purchasing a product or subscribing to a service. But it is very hard to go through every review. They will be benefited from some opinion

[8] http://www.nist.gov/tac/2008/cfp.html

summarization system which automatically extracts product features and summarize based on its effectiveness and drawbacks. Blogs and News articles report many topics added with opinion of author written in different languages. Tracking topics like politics, sports become easier in multiple languages, if an opinion retrieval show the change in opinion on political leaders, political parties and sports person etc.

1.7.3 Advertisements Placement

Internet has become an important market place for the business to sell their products and services. Most of the time this is done by placing advertisements on blogs, social media, search etc. There has been research done to place relevant advertisements for the static content posted on the page. But sometimes user-generated content like social media see dynamic changes in content added with metadata. So adding the user intent or opinion to advertisement placement will make the advertisement relevant to the user. For example, an advertisement on relevant product when one praises a product, placing an advertisement from a competitor if one criticizes a product.

1.7.4 Opinion search

General search for opinions in News and Blogs can be done using a opinion search or retrieval system. It helps to search for opinionated topics. When people publish blogs or write News articles, opinions forms very important feature. An opinion retrieval system can provide search of opinionated blog or News documents expressing opinions about a query. Subjectivity analysis helps in finding the subjective clues by decomposing documents into sentences. A relevant document satisfies criteria of finding relevance to the query, and contains opinions or comments about a query.

1.8 Problem Statement

Mining subjectivity and opinion information from different languages is a challenging task to achieve due to high dependency of language specific resources and tools. Given any resource-scarce language, our aim of research is to devise approaches for subjectivity and opinion analysis that uses minimal language specific resources and are scalable to new languages with minimum effort. In-order to achieve it, fine-grain opinion mining solutions should be designed by leveraging rich language resources for doing natural language processing tasks and later projected to resource-scarce languages. Similarly, subproblem of sentence-level subjectivity detection for different languages is to be achieved by eliminating the dependency on language specific tools like POS taggers, Named Entity recognizers etc. Approaches needs to be designed by leveraging the existing machine learned models that are consistent across languages for subjectivity detection.

1.9 Contribution

Aim of the research is to find opinions mentioned in the resource-scarce languages. We have earlier discussed about the different challenges in the Section 1.5 to achieve it. Our goal was to come up with a framework and solutions to surpass these challenges and issues. Main contributions in doing so are listed below.

- A framework is designed to extract opinions about entities mentioned in the resource-scarce language articles. Also, framework was extensively evaluated with different metrics using the comparable News corpora as dataset.
- Subproblem of opinion mining, sentence-level subjectivity detection is achieved in multiple languages with feature selection and our proposed feature weighing methods.
- A re-usable dataset is created in Hindi and Telugu to facilitate further research in sentence-level subjectivity and sentiment analysis.

1.10 Thesis Organization

Thesis is organized to get the in-depth view of the work carried out in finding subjectivity and opinions in resource-scarce languages.

- In Chapter 2, we discuss about the process carried out in preparing the datasets in Indian languages for subjectivity and opinion analysis. We show it in the form of two subsections one for Hindi and Telugu language.
- In Chapter 3, we discuss the approach used to perform opinion mining in resource-scarce languages with minimal resources. Initially, we give background about this work by presenting a list of previous and related works in this area. It is then followed by our approach, where emphasis is given on extracting subjective sentences and creating subjective language models for opinion mining. To prove the effectiveness of our approach, we present experimental results by comparing the approach with various methods. Finally, it is concluded with discussion on the results.
- In Chapter 4, we revisit the sentence-level subjective detection used for opinion mining in Chapter 3 by finding better and language independent methods. Initially, we give background about this work by presenting a list of previous and related works in this area. It is followed by mention of different feature extraction, weighing and selection methods which are consistent across different languages. To find the best combination of feature methods, experimental results are given using different learning algorithms. Finally, it is concluded with discussion on the results.
- Chapter 5 and Chapter 6 give the conclusion of the work and future research problems respectively.

Chapter 2

Resource Generation

Availability and access to text data plays an important role for performing any natural language processing tasks. In our research, we aim to design approaches for subjectivity and opinion analysis in resource-scarce languages which have dearth of datasets. To facilitate our research, initially time is spent in creating gold standard datasets in south Asian languages like Hindi and Telugu to perform subjectivity and opinion analysis. Sections below provide details about approaches used to create these datasets.

2.1 Hindi Dataset

Hindi is the major language spoken in India and has around 180 million[1] speakers worldwide. It has larger presence in the form of print media and has been growing its presence on the web. Hindi Content on the web can be found in the form of blogs, news articles etc. Previous studies [45] on English states that 44% of sentences in a news collection are subjective. Our motivation was to leverage the idea for Hindi News articles and perform sentence-level subjectivity analysis and fine-grained opinion mining.

A dataset is created using 5 different topics randomly selected from FIRE 2010[2] collection. It is a comparable corpora consisting of news articles in English and Hindi languages covering different areas like sports, politics, business, entertainment etc. The relevance judgments are provided for the topics. Relevance judgments of Hindi and English are used to select the relevant documents for a topic.

Table 2.1 show the topics used to create the dataset. For each topic, Hindi documents are used to create a gold standard dataset[3] of subjective, objective sentences and tuples of opinion targets and opinions. Subjective sentences constituted positive and negative sentiment labels. Figure 2.1 shows a sample news article in Hindi obtained from the FIRE 2010 corpus that is relevant to a topic. While, relevant and comparable documents in English are used for extraction of NE's, adjectives and verbs.

[1] http://www.ethnologue.com/show_language.asp?code=hin
[2] http://www.isical.ac.in/ fire/2010/index.html
[3] http://search.iiit.ac.in/uploads/README_0.

<DOC>
<DOCNO>fullnews_id_3615631_date_5_10_2005.utf8</DOCNO>
<TEXT>
संघ परिवार में बगावत
आडवाणी की मौजूदगी से खफा विहिप ने बैठक का बहिष्कार किया
हरिद्वार।
हिंदुत्व के मुद्दे पर भाजपा को शक की निगाह से देख रही विहिप आज खुलकर सामने आ गई और उसने यहां राष्ट्रीय स्वयं सेवक संघ की राष्ट्रीय कार्यकारिणी की बैठक का बहिष्कार कर दिया। बैठक में भाजपा अध्यक्ष लालकृष्ण आडवाणी की मौजूदगी से नाराज विहिप नेता अशोक सिंघल, प्रवीण तोगड़िया और आचार्य गिरिराज किशोर दिनभर चली मान-मनौव्वल के बावजूद दूधाधारी आश्रम के कमरों में दरवाजे बंद कर बैठे रहे।भाजपा की राजनीति को हिंदुत्व और राम विरोधी करार देते हुए विहिप ने दूधाधारी मंदिर में हिंदू जागरण अनुष्ठान भी शुरू कर दिया है। संघ के इतिहास में पहली बार हुए इस तरह के बहिष्कार से बुरी तरह आहत राष्ट्रीय कार्यकारिणी शुक्रवार को कोई भी प्रस्ताव बैठक में नहीं रख पाई। संघ के कार्यकर्ता दिनभर बंद हाल में उपस्थित आडवाणी तथा वेंकेया पर जमकर बरसते रहे।
विहिप नेताओं का मानना है कि भाजपा मंदिर मुद्दे से राजनीतिक लाभ उठाने के बाद इससे दूर भाग रही है। वह आडवाणी की भाजपा अध्यक्ष के रूप में नियुक्ति से भी खफा हैं। सूत्रों के मुताबिक बृहस्पतिवार से हरिद्वार में डेरा डाले विहिप नेता इस बात से भी नाराज हैं कि दो दिवसीय बैठक के एजेंडे में राम मंदिर का मुद्दा शामिल नहीं किया। उनका मानना है कि ऐसा इसलिए किया गया है ताकि भाजपा अध्यक्ष आडवाणी को किसी तरह की असुविधा न हो।
विहिप नेताओं ने कल रज्जू भैया परिसर स्थित बैठक स्थल जाकर अनौपचारिक सत्रों में भाग लिया था। लेकिन आज सवेरे नौ बजे जब संघ राष्ट्रीय कार्यकारिणी की बैठक प्रारंभ हुई तो विहिप नेताओं ने यह कहते हुए बैठक में भाग लेने से मना कर दिया कि इसमें हिंदुत्व व राम मंदिर के मुखर विरोधी लालकृष्ण आडवाणी व वेंकेया नायडू भाग ले रहे हैं। भाजपा के अनेक नेताओं को दूधाधारी आश्रम भेजा गया, लेकिन सिंघल, तोगड़िया तथा गिरिराज किशोर नहीं माने। आखिर में सरसंघ चालक केसी सुदर्शन भी वहां गए लेकिन विहिप नेता टस से मस नहीं हुए।
संघ की बैठक में विहिप के बहिष्कार की खबर जैसे ही रज्जू भैया परिसर पहुंची, मानो कोहराम मच गया। बैठक में उपस्थित देश भर से आए ३०० प्रतिनिधियों ने सरसंघ चालक केसी सुदर्शन की उपस्थिति में आडवाणी पर प्रहार शुरू कर दिए। आडवाणी के यह कहने पर कि जब केंद्र में भाजपा सत्तासीन थी, तब वे पार्टी अध्यक्ष नहीं थे, सारा गुस्सा वेंकेया पर फूट पड़ा।

</TEXT>
</DOC>

Figure 2.1 A Sample Hindi News Article

2.1.1 Preparation Steps

Gold standard dataset was prepared using two annotators. The aim was to identify subjective sentences and also private states in terms of their functional components. Such as states of writer or person holding opinion, optionally towards a target. For example, for the private state expressed in the sentence "Advani killed Lokpal Bill in Lok Sabha, says Rahul.", the Holder of opinion is "Rahul", the opinion is "killed lokpal Bill", and the target is "Advani". To reduce the complexity, annotators were instructed to find only opinions and their targets that exist in a document.

Subjective sentences annotation covered broad and useful subset of linguistic expressions that are naturally occurring in text expressing opinion and sentiment. For identifying opinion words and targets, annotators identified opinion words in context, rather than judging them out of context. Annotators marked not only adjectives but also different range of words and constituents belonging to verbs and nouns.

Subjectivity in the text can be understood from different levels. Annotators were instructed to mark a sentence subjective as in [46] if it has any of the following markers.

1. Any mention of private states.
2. Any events expressing private states.
3. Any expressive subjective elements.

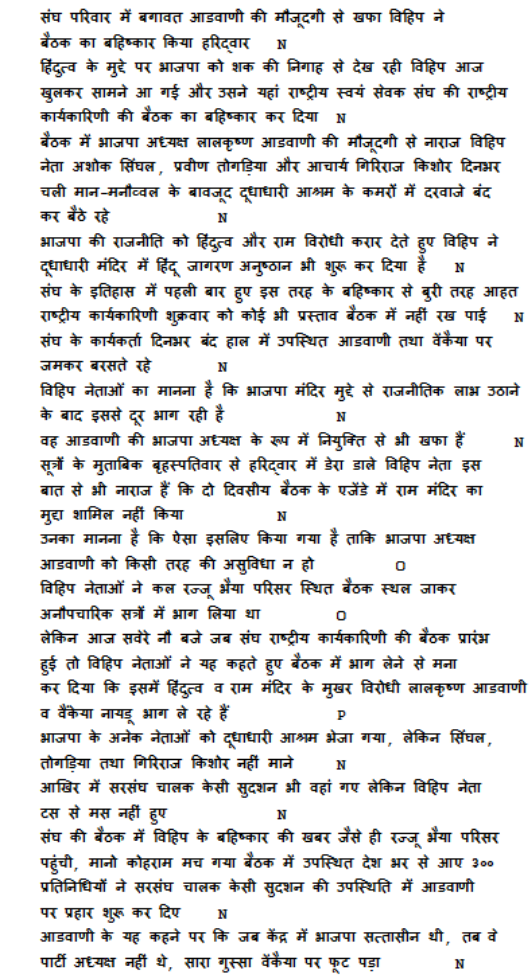

Figure 2.2 Annotated Sample of a Hindi News Article

Figure 2.2 show the sentence-level subjectivity annotations done on the article using the above mentioned approaches. Similarly, Figure 2.3 show an example of private states mentioned in the sentences. Writer of the article had given opinions about people as opinion targets that are mentioned in green and opinions about them in red at word-level. Basically, it tries to find the opinions as the expressive subjective elements at more fine-grained level in a document. From the annotations done on the corpus it is observed that most of the subjective expressions are typically found to be expressive subjective element expressions rather than direct subjective expressions. Wiebe [46] states that direct subjective frames contain more attributes than expressive subjective element frames.

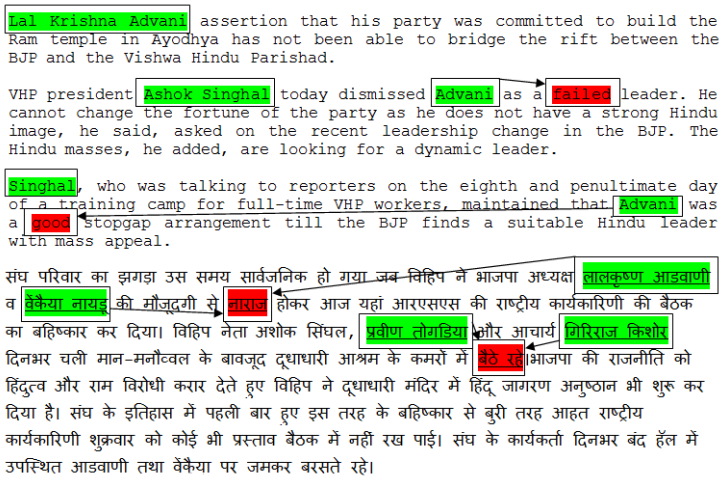

Figure 2.3 Sentences with Opinions and Opinion Targets in Comparable news Corpora

Topics Used for Dataset					
Topic	T1	T2	T3	T4	T5
	78	80	85	90	91

Table 2.1 Topics used for Dataset

2.1.2 Agreement

The inter-annotator agreement is understood by calculating observed kappa scores and agreement percentage. Table 2.2 show the kappa score[4] and the agreement percentage in identifying subjective sentences and opinions, opinion targets tuples averaged over 5 topics.

[4]http://www.vassarstats.net/kappa.html

Inter-Annotator agreement (Unweighted kappa)		
Task	Agree	Observed kappa
Subjective sentences	86.5%	0.689
(Opinions,Opinion targets)	73.7%	0.493

Table 2.2 Inter-Annotator agreement

అవును.. మీరు సరిగ్గానే చదివారు. పాకిస్తాన్ తీవ్రవాద మరియు అంతర్గత రక్షణ శాఖా మంత్రి ఒసామా బిన్ లాడెన్ ని అమెరికా సైన్యం ఒక ప్రత్యేక ఆపరేషన్ లో మట్టుపెట్టింది. పాకిస్తాన్ రాజధానికి అత్యంత సమీపంలో, ఆయన అధికార నివాసంలో జరిగిన ఈ సంఘటనలో, మరెవ్వరూ గాయపడినట్టు సమాచారం లేదు. ఒబామా, తీవ్రవాదం పై తమ దేశం చేస్తున్న యుద్ధం లో ఇది ఒక మైలు రాయిగా వర్ణించగా, పాకిస్తాన్ ప్రభుత్వం మాత్రం ఈ సంఘటనని తీవ్రంగా ఖండించింది. ఇది ఖచ్చితంగా తమ సార్వభౌమాధికారాన్ని భంగ పరచడమే అని అభివర్ణించింది. కాల్పులు జరిగిన సమయం లో ప్రభుత్వ రక్షణ దళం, గ్రీన్ కమాండోలు అక్కడే ఉన్నా, వాళ్ళు రేడియో లో ఐపీయల్ కామెంటరీ వింటూ ఉండటం వల్ల, వెంటనే ఎదురు కాల్పులు చెయ్యలేకపోయినట్టు తెలుస్తోంది. ఇది యాదృచ్ఛికమా, లేక భారత్ హస్తం ఏమైన ఉందా అనే కోణం లో పాకిస్తాన్ ఇప్పటికే దర్యాప్తు మొదలు పెట్టింది.

పరిస్థితి పై, అధికారికంగా స్పందించడానికి ప్రధాని నిరాకరించారు, రేపు వార్తాపత్రికలు చూసి గానీ తను ఏమీ చెప్పలేను అన్నారు; కానీ ఆయన కాఫీ తాగుతూ అన్న మాట విని మా ప్రతినిధి ఇచ్చిన రిపోర్ట్ ఏంటంటే - "ఇది పూర్తిగా పాకిస్తాన్ అంతర్గత వ్యవహారం.. మరియు అమెరికా కి వెలుపలి వ్యవహారం కనుక, తాము స్పందించక్కర్లేదు అన్నారంట. కాకపోతే పరిణామాలని బాగా దగ్గరగా (అంటే బాగా క్లోజ్ గా అన్నమాట) పరిశీలిస్తున్నట్టు తెలిసింది. ఇందుకు గానూ ప్రధాని కార్యాలం పై అంతస్తులో ఒక టెలిస్కోప్ ని కూడా అమర్చడానికి ఈ-టెండర్లు పిలుస్తున్నారంట.. "

Figure 2.4 A Sample Telugu Blog

2.2 Telugu Dataset

Over 75 million[5] people around the world speak Telugu and it stands second only to Hindi in India as to the number of native speakers. As stated earlier in the Introduction, growing Internet penetration in India has increased the participation of native language speakers who are adding content to the Web in their native languages. Blogs are one such medium where people actively participate. A. Mogadala.et.al [30] stated that Blogosphere has doubled every six months for the last three years and good percentage of these blogs are created in Telugu. Generally, these blogs contain opinions of people on different genres like movies, travel experiences, daily activities and current happenings in the society.

The diversity of user generated information in Telugu in Blogs motivated us to create a sentence-level subjectivity analysis dataset to do facilitate research in subjectivity and opinion analysis. Blogs are crawled using Samoohamu[6], a collection containing different Telugu blogs. Once the content is crawled, they are analyzed for the language present in the blogs. If the language found in the body was not Telugu, blogs are discarded. Figure 2.4 shows a sample Telugu blog. Around 2926 blogs along with their comments are crawled belonging to different domains like politics, entertainment, poetry etc. Out of which 250 blogs are randomly picked to annotate subjective and objective labels at sentence-level. Subjective labels constituted positive and negative sentiment labels. Table 2.3 show the number of domains and blogs present in 250 blog dataset.

[5]http://www.cs.ucdavis.edu/ vemuri/classes/freshman/IntroductionToTelugu.htm
[6]http://telugu.samoohamu.com/

Domains in the Dataset	
Domain	Total Blogs
Entertainment	16
Language	82
General-News	19
Politics	28
Spirituality	21
Misc	84
Total	250

Table 2.3 Domains in the Dataset

2.2.1 Preparation Steps

Gold standard dataset was prepared using two annotators who identified subjective and objective sentences existed in the blogs. Approach was similar to Hindi dataset creation. Figure 2.5 show subjectivity annotations done on the blog at sentence-level.

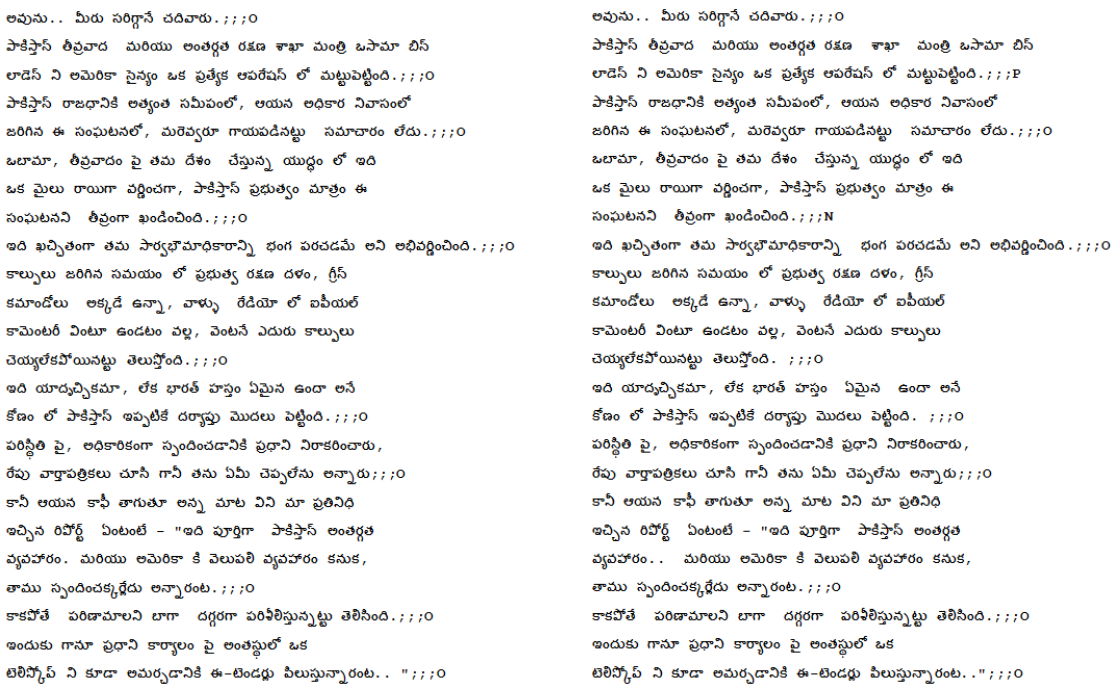

Figure 2.5 Annotated Samples of Telugu Blogs by Annotator-1 and Annotator-2

Inter-Annotator agreement(Unweighted kappa)		
Task	Possible kappa (Maximum)	Observed kappa
Entertainment	0.5378	0.1103
General-News	0.6749	0.1176
Language	-	-
Misc	-	-
Politics	0.3611	0.0201
Spirituality	0.6837	0.0891

Table 2.4 Inter-Annotator agreement

Entertainment-Sentences				
	Positive	Negative	Objective	Total
Positive	12	5	28	45
Negative	13	24	4	41
Objective	100	4	100	204
Total	125	33	132	290

Table 2.5 Entertainment-Sentences-Inter-Annotator-Agreement

2.2.2 Agreement

The inter-annotator agreement calculated using possible and observed kappa in identifying subjective sentences for each domain is given in Table 2.4. It is observed that domains pertaining to Language and Misc had very low kappa scores and is not reported. Agreement in annotating sentences between two annotators for each domain is given by Tables 2.5 to Tables 2.10.

General-News-Sentences				
	Positive	Negative	Objective	Total
Positive	18	0	13	31
Negative	7	34	44	85
Objective	15	16	15	46
Total	40	50	72	162

Table 2.6 General-News-Sentences-Inter-Annotator-Agreement

Language-Sentences				
	Positive	Negative	Objective	Total
Positive	51	1	103	155
Negative	2	6	22	30
Objective	97	9	97	203
Total	150	16	222	388

Table 2.7 Language-Sentences-Inter-Annotator-Agreement

Misc-Sentences				
	Positive	Negative	Objective	Total
Positive	18	0	13	31
Negative	7	34	44	85
Objective	15	16	15	46
Total	40	40	72	162

Table 2.8 Misc-Sentences-Inter-Annotator-Agreement

Politics-Sentences				
	Positive	Negative	Objective	Total
Positive	99	21	149	269
Negative	43	114	414	571
Objective	230	192	230	652
Total	372	327	793	1492

Table 2.9 Politics-Sentences-Inter-Annotator-Agreement

Spirituality-Sentences				
	Positive	Negative	Objective	Total
Positive	36	1	26	63
Negative	5	12	18	35
Objective	66	28	66	160
Total	107	41	110	258

Table 2.10 Spirituality-Sentences-Inter-Annotator-Agreement

2.3 Summary

In this chapter, we explained the approach taken for preparation of datasets for south Asian languages mainly Hindi and Telugu. We achieved this using the help of annotators by educating and guiding them the principles for annotating sentiment datasets. Hindi dataset is prepared using the comparable corpora of news articles on different topics. Each news article is annotated at two levels, one at sentence-level with polarity labels positive, negative and objective and more fine-grained level by identifying opinions on the targets present in the article. Telugu dataset is prepared using the blogs belonging to different domains. Each blog is annotated at sentence level with polarity labels positive, negative and objective. These datasets can be used as the gold standard datasets to facilitate subjectivity or sentiment analysis research in these languages.

Chapter 3

Fine-Grain Opinion Mining in Resource-Scarce Languages with Comparable Corpora

Mining opinions at finer grain in resource-scarce languages(RSL) throw multiple challenges. We need to understand these challenges by asking different questions. Some of these challenges require deep insight into the problems that are proposed by the different questions. Listed below is our hypothesis to solve those problems.

- Can opinion extraction about entities in resource-scarce languages is achieved using minimal language specific resources and tools.
- Can we leverage on resource rich languages English using a comparable corpora.
- Can we surpass word identification tools in resource-scarce languages to identify named entities, adjectives and verbs [18] used for subjective text.
- Can we design approaches which are easily scalable to multiple languages.

3.1 Background

Opinion mining peformed earlier can be broadly categorized into classification and retrieval methods.

3.1.1 Classification Methods

Classification based approaches are proposed for English and mostly for product reviews. W. Du.et.al [9] proposed an iterative reinforcement framework that clusters product features and opinion words simultaneously and iteratively by fusing both their semantic information and co-occurrence information. This approach was designed to outperform the template extraction based approaches. Another approach [25] used probabilistic models. They treat the problem of product feature and opinion extraction as a sequence labeling task and adopt Conditional Random Fields models to accomplish it.

3.1.2 Retrieval Methods

Traditional opinion retrieval systems in TREC [50, 51] tried on large blog collections extract opinions about a query. These systems are designed for retrieving opinions about a single entity i.e. a query. Using this systems on news articles might fail because of the presence of multiple entities with different opinions expressed on them. Some other approaches [6] used opinion weights and proximity of terms to the query directly. They considered proximity-based opinion density functions to capture the proximity information. But these approaches may not be effective in retrieving information from resource scarce language documents due to dependency on language specific features. Some systems developed for languages Romanian [28], Chinese [10, 52] , Japanese [39, 20] only concentrate on identification of subjective texts. Other multilingual systems participated in NTCIR-6 [29] opinion extraction track were dataset specific and their approaches cannot be easily ported to other languages.

This show a need for better and new approaches for opinion mining in resource-scarce languages like european and Indian languages to overcome the above issues.

3.2 Proposed Approach

Opinions about opinion targets are mined using a retrieval approach. Following approach is chosen to overcome the limitations of natural language processing tools and resources in resource-scarce languages. To overcome these issues, we leveraged resource rich language like English to process the text. Opinion targets (mainly people) and possible opinions was extracted from English and ported to Hindi. Even though the experiments are performed on Hindi, proposed approaches are scalable to other languages.

Initially, Named entities, adjectives, verbs from English collection are extracted using named entity recognition tools and POS taggers. It is then transliterated and translated to other languages using conditional random field approach [14] and bilingual dictionaries respectively. Next, cross language ported words are used to identify subjective sentences to create a subjective language model (SLM) and opinion queries in Hindi. Since everything is done offline, we consider this as an offline process which is depicted in the Figure 3.1. Inference Network(IN) framework which support proximity and belief between query words is then used to create structured opinion queries (OQs) with SLM. It is similar to Language Model (LM) with IN [7] for retrieval of documents. This approach is used to confirm the presence of opinions about opinion targets in documents. Since queries are fired for retrieval, this is considered as an online process given by Figure 3.2. In the following sections detailed description for achieving subjective sentence extraction, subjective language model (SLM) and extension of SLM with IN for opinion retrieval is described.

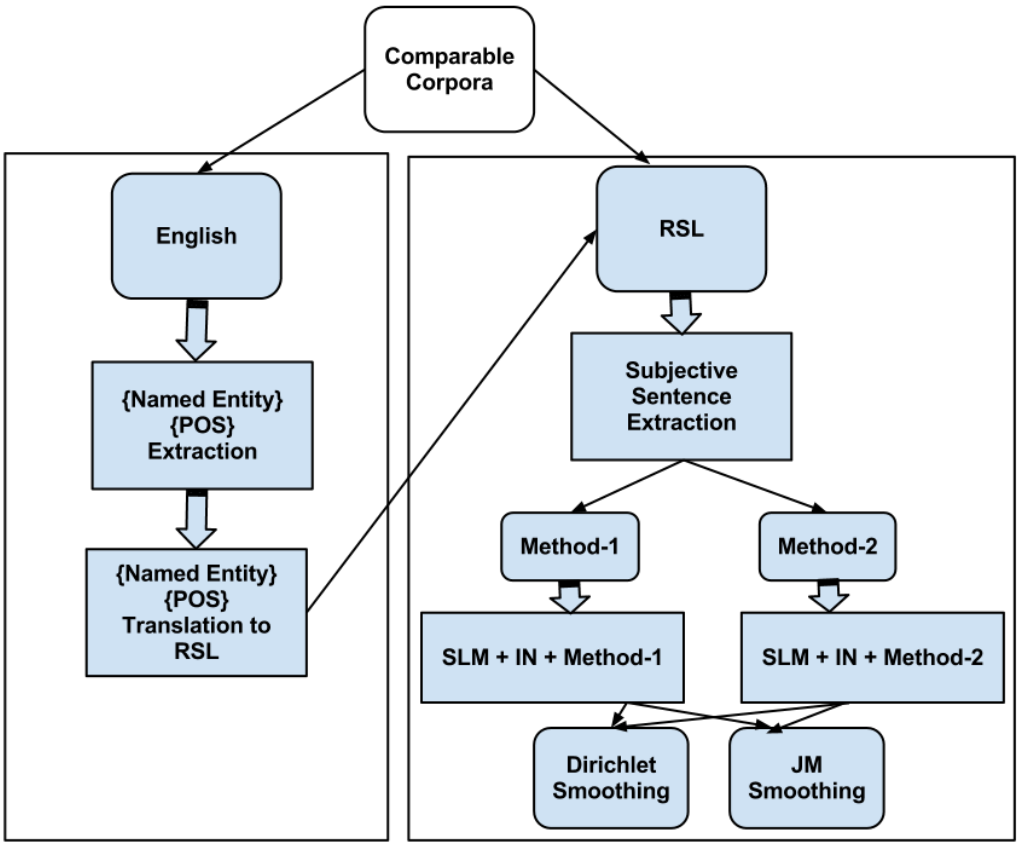

Figure 3.1 Offline Process

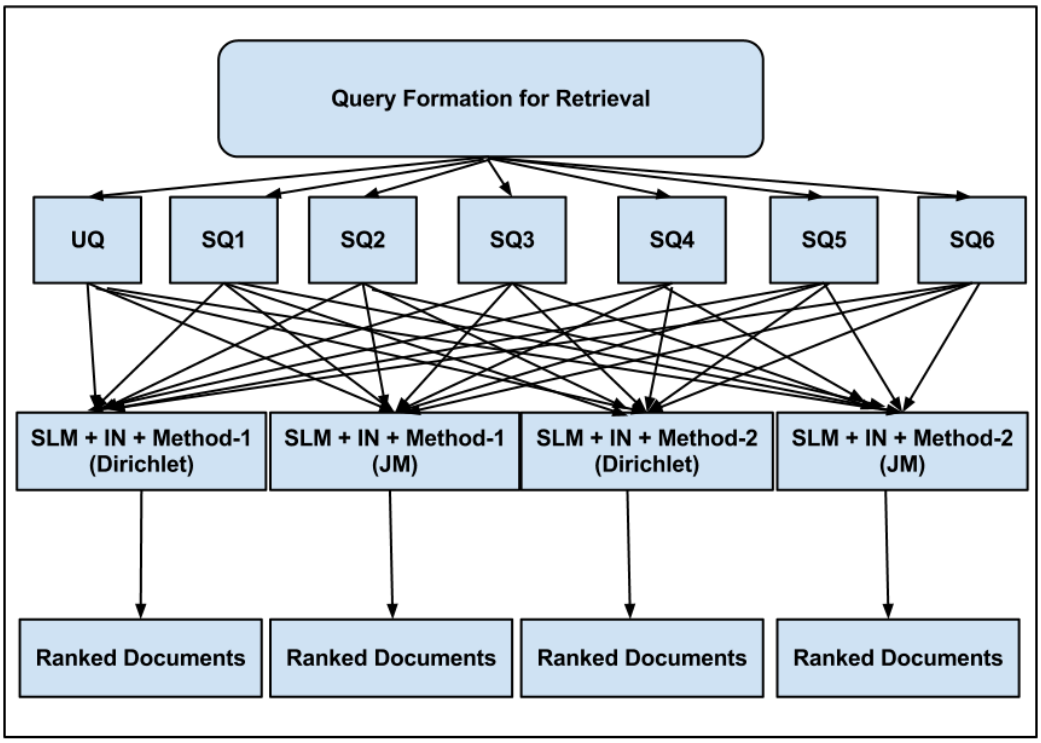

Figure 3.2 Online Process

3.2.1 Subjective sentence extraction

Subjective sentences were selected in the document using two approaches motivated from eariler approach [35]. But we also consider named entities to find opinion targets in the documents. Documents that does not contain subjective sentences are eliminated. Two different approaches are used to extract subjective sentences.

3.2.1.1 Method 1

If two or more strong subjective expressions occur in the same sentence mainly Named entities, Adjectives or Verbs, the sentence is labeled strongly subjective.

3.2.1.2 Method 2

If at least one of Named Entities or Adjectives or Verbs exist in a sentence then it is labeled as weakly subjective sentence.

Difference between performance of *Method 1* and *Method 2* can be observed in the experiments when OQs are used for retrieval.

3.2.2 Subjective Language Model for Opinion retrieval

Subjective language model (SLM) is created for resource scare language documents similar to language model (LM) [21] by selecting **subjective sentences in the documents** using two different methods mentioned earlier.

SLM approach to opinion retrieval is finding probability of an opinion query oq being generated by a probabilistic model based on a document D denoted as $p(oq|D)$. It is done by estimating the posterior probability of document D_i and opinion query oq using Bayes formula given by Equation 3.1.

$$p(D_i|oq) \propto p(oq|D_i)p(D_i) \tag{3.1}$$

where $p(D_i)$ is prior belief that is relevant to any opinion query and $p(oq|D_i)$ is the query likelihood given the document D_i, which captures the particular opinion query oq information. $p(D_i)$ is considered to be multinomial distribution. This assumption helps in choosing better smoothing techniques which is mentioned later.

For each document D_i in the collection c, its subjective language model defines the probability $p(ow_1, ow_2, ..., ow_n|D_i)$ containing opinions and opinion targets given by $ow_1, , ow_n$ as sequence of n query terms. Documents are ranked according to this probability.

The probabilities of the opinion words ow_i in document D_i improves the weight of subjectivity in the document D_i. Equation 3.2 gives probability $p(ow_i|c)$ of finding opinion words ow_i for entire collection c while Equation 3.3 gives probability $p(ow_i|D)$ only for a document D.

$$p(ow_i|c) = \frac{cfreq(ow_i, c)}{\Sigma_{i=1}^{n} cfreq(ow_i, c)} \tag{3.2}$$

$$p(ow_i|D) = \frac{tfreq(ow_i, D)}{\Sigma_{i=1}^{m} tfreq(ow_i, D)} \tag{3.3}$$

Here, $cfreq(ow_i, c)$ represents collection frequency of ow_i in the collection c and $tfreq(ow_i, D)$ is term frequency of the ow_i in a document D. n is total opinion words in collection, while m is total opinion words in a document. The non smoothed model gives maximum likelihood estimate of relative counts. But if the word is unseen in the document then it results in the zero probability. So the smoothing is helpful to assign a non-zero probability to the unseen words and improve the accuracy of word probability estimation in general. In this paper, we used Dirichlet and Jelinek-Mercer smoothing to assign non-zero probabilities to unseen words in the documents and collection. Below are the two smoothing techniques that are used to remove the zero probability scores to unseen words as mentioned in [55].

3.2.2.1 Dirichlet Smoothing

As subjective language model prior is considered to be multinomial distribution, for which the conjugate prior for Bayesian analysis is the Dirichlet distribution. We choose the parameters of the Dirichlet to be μ and the values that needs to be added in the numerator of the Equation 3.3 for smoothing. Equation 3.4 gives values which is Dirichlet parameter multiplied with probability of each opinion word in collection. So after smoothing, the Equation 3.3 is converted into Equation 3.5.

$$\mu p(ow_1|c), \mu p(ow_2|c), ..., \mu p(ow_n|c) \tag{3.4}$$

$$p_\mu(ow_i|D) = \frac{tfreq(ow_i, D) + \mu p(ow_i|c)}{\Sigma_{i=1}^{m} tfreq(ow_i, D) + \mu} \tag{3.5}$$

3.2.2.2 Jelinek-Mercer Smoothing

In Jelinek-Mercer smoothing approach, we consider the mixture of document model $p(ow_i|D)$ and collection model $p(ow_i|c)$ as used in standard retrieval model. This approach takes parameter λ which needs to be estimated. Equation 3.6 gives the mixture model equation.

$$p_\lambda(ow_i|D) = (1 - \lambda)p(ow_i|c) + (\lambda)p(ow_i|D) \tag{3.6}$$

Subjective language model with *Method 1* and *Method 2* forms our extended baseline from basic language model. We will further extend this model with **inference network which allows different types of structured query formulations** as explained below.

3.2.3 Subjective Language model with Inference network for Opinion retrieval

This retrieval model combines SLM with IN [40] to confirm the presence of opinion about opinion targets. This model allows opinion queries containing possible opinions on opinion targets to use prox-

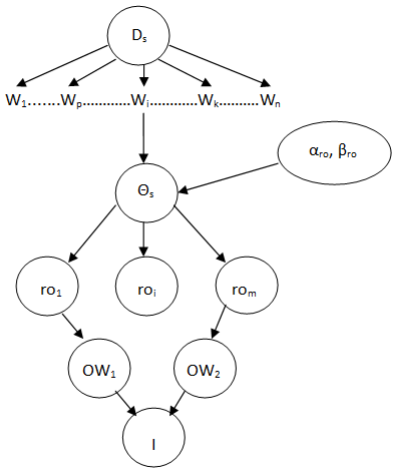

Figure 3.3 Inference Network Representation for SLM

imity and belief information between query terms similar to Indri [41]. We observe in IN framework that documents are ranked according to probability $p(I|D, \alpha, \beta)$ using the belief information need I calculated using a document D and hyper parameters α and β as evidence.

Information need I in our scenario is simply a belief node that combines all of the belief nodes ow_i's containing **opinion evidence** within the inference network into a single belief. In our scenario, belief nodes ow_i's are opinion words in the document. In order to obtain evidence of ow_i's, representation concept nodes ro_i's are used. ro_i's are binary random variables representing only opinion word unigrams from the total features extracted in the document representation. Features here are all word unigrams $w_1....w_n$ present in a document. In order to find the individual belief nodes ow_i's we need to calculate $p(ro_i|D)$ and then combine ow_i's to get information need I. Figure 3.3 shows the representation. Documents are then ranked accordingly using $p(I|D, \alpha, \beta)$.

To achieve it, we assume each subjective document D_s to be in multiple-Bernoulli subjective model θ_s and not multinomial distribution as assumed in previous section. As this model assumption imposes concept nodes ro_i's to be independent. First, we compute $p(\theta_s|D_s)$ which is the model posterior distribution given by Equation 3.7.

$$p(\theta_s|D_s) = \frac{p(D_s|\theta_s)p(\theta_s)}{\int_{\theta_s} p(D_s|\theta_s)p(\theta_s)d\theta_s} \tag{3.7}$$

Where $p(D_s|\theta_s)$ is the likelihood of generating D_s from model θ_s and $p(\theta_s)$ is the model prior. We see this posterior probability $p(\theta_s|D_s)$ is distributed according to multiple-Beta $(\alpha_{ro}, \beta_{ro})$ because beta distribution is the Bernoulli's conjugate prior.

In this IN framework only opinion terms are considered denoted by $optf_{ro}$ out of total terms in a document D_s for independent ro_i's. This is like finding x positive results for n trials. Thus $p(D_s|\theta_s)$ distribution is changed to multiple-Beta($\alpha_{ro}+optf_{ro}, \beta_{ro}+|D_s|-optf_{ro}$) containing opinion terms, where $|D_s|$ is the total opinion word count of the document D_s. Expression 3.8 give estimate of $p(ro|D_s)$ representing individual beliefs ow_i's. It is nothing but a expectation over the posterior $p(\theta_s|D_s)$.

$$p(ro|D_s) = \int p(ro|\theta_s)p(\theta_s|D_s)d\theta_s \tag{3.8}$$

But we know the expectation of a beta distribution given in terms of its parameters is $\frac{\alpha}{\alpha+\beta}$. Therefore, given that $p(D_s|\theta_s)$ is also distributed according to multiple-Beta($\alpha_{ro} + optf_{ro}, \beta_{ro} + |D_s| - optf_{ro}$) the Equation 3.8 is now represented by Equation 3.9.

$$p(ro|D_s) = \frac{optf_{ro,D_s} + \alpha_{ro}}{|D_s| + \alpha_{ro} + \beta_{ro}} \tag{3.9}$$

So for document D_s, $optf_{ro,D_s}$ is the opinion term frequency with ro_i's as features. Thus subjective language model θ_s is estimated based on hyper parameters α_{ro} and β_{ro} combined with the observed document D_s. From these models, concept nodes q_i's are used forming an opinion query. Overall belief I from this opinion query is used for ranking subjective documents.

But there can be chances of zero probability and data sparseness in the model. In order to eliminate this problem we employ smoothing.

3.2.3.1 Dirichlet Smoothing

Dirichlet smoothing is done in order to handle zero probability. α_{ro} and β_{ro} values were chosen given by Equation 3.10 and Equation 3.11 respectively to modify Equation 3.9 to Equation 3.12. $p(ro|c)$ gives beliefs of the representation concept nodes in entire document collection.

$$\alpha_{ro} = \mu p(ro|c) \tag{3.10}$$

$$\beta_{ro} = \mu(1 - p(ro|c)) \tag{3.11}$$

$$p(ro|D_s) = \frac{optf_{ro,D_s} + \mu p(ro|c)}{|D_s| + \mu} \tag{3.12}$$

where μ is free parameter and $optf_{ro,D_s}$ is the opinion terms frequency for feature ro in Document D_s.

3.2.3.2 Jelinek-Mercer Smoothing

This method involves a linear interpolation of the maximum likelihood model with the collection model, using a coefficient λ.

$$p(ro|D_s) = (1 - \lambda)\frac{optf_{ro,D_s}}{|D_s|} + \lambda p(ro|c) \tag{3.13}$$

Thus, this model combines SLM and IN with different smoothing techniques used for opinion retrieval. Queries formed for retrieval are mentioned in next section.

3.2.4 Query Formulation for Retrieval

We understood from the previous section that belief nodes ow_i's combine evidence from the concept nodes ro_i's to estimate the belief that a opinion words are expressed in a document. But actual arrangement of the ow_i's and the way they combine evidence is dictated by the user through the query formulation. So we form structured opinion queries (OQs) to identify opinion targets and opinions in resource-scarce language news articles. These structured OQs contain named entities, adjectives and verbs as opinion words. OQs use proximity and beliefs between query words.

We will see difference in query formulation of opinion queries which use proximity and belief between query words and that don't use below.

3.2.4.1 Opinion queries without proximity and belief

Query terms in OQ are treated as bag of words. Each word in the opinion query is assumed independent without any conditional information. This query model may find relevant documents to query terms but may not extract opinions about opinion targets. In order to rank documents, opinion query likelihood is calculated using subjective language model θ_s given by Equation 3.14.

$$p(oq_1, oq_2|\theta_s) = \Pi_{i=1}^{2} p(oq_i|\theta_s) \tag{3.14}$$

where oq_1 and oq_2 represent opinion targets and opinions on them respectively.

3.2.4.2 Opinion queries with proximity and belief

Indri Query language[1] is used to form OQs. In this approach, only those query operators are selected which use proximity and belief information in their representations. Proximity representations are used

[1] http://www.lemurproject.org/lemur/IndriQueryLanguage.php

Label	Opinion Queries
UQ	NE OExp
SQ 1	#10(NE OExp)
SQ 2	#filreq(NE #Combine(NE OExp))
SQ 3	#filreq(NE #uw10(NE OExp)
SQ 4	#filreq(NE #weight(2.0 #uw10(NE OExp)))
SQ 5	#uw10(NE OExp)
SQ 6	#uw5(NE OExp)

Table 3.1 Opinion Queries

Label	Opinion Queries Explanation
UQ	Unstructured query containing OT (named entity) and opinion expressed (OExp).
SQ 1	Match OT and OExp in ordered text within window of 10 words.
SQ 2	Evaluates combined beliefs of OT and OExp given OT in document.
SQ 3	Match OT and OExp in unordered text within window of 10 words given OT in document.
SQ 4	Match OT and OExp in unordered text within window of 10 words with extra weight of 2.0 and OT given in document.
SQ 5	Match OT and OExp in unordered text within window of 10 words.
SQ 6	Match OT and OExp in unordered text within window of 5 words.

Table 3.2 Explanation of Opinion Queries containing Opinion Target**(OT)** and Opinions Expressed **(OExp)**

to map the opinion targets and opinions expressed on them appearing within ordered or unordered fixed length window of words. To use belief information of opinion words of OQ represented as belief nodes of subjective documents. Opinion words are combined using a belief operator.

Size used for proximity window is intuitive. Each query represented in this framework uses SLM with IN for opinion retrieval. OQs used for retrieval are mentioned in Table 3.1 and their explanation in Table 3.2. Query SQ4 sample representation is given by Figure 3.4

3.3 Experimental Setup

Information about pre-processing done on the collection is mentioned below.

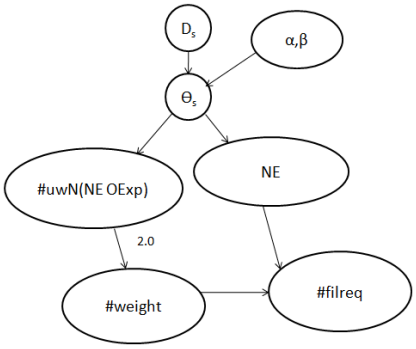

Figure 3.4 SQ4 Representation

English to Hindi Document Analysis	
Translation Coverage(ADJ)(After Exp)	63.9%
Translation Coverage(VB)(After Exp)	65.3%
Transliteration Error	13%

Table 3.3 English to Hindi Document Analysis

3.3.1 Pre-processing on the Collection

English data from the FIRE 2010 collection is used to extract NE's, adjectives and verbs. Stanford NER[2] and POS Tagger[3] is used to extract NE's and adjective, verbs respectively. Manually prepared word-aligned bilingual corpus and statistics over the alignments is used to transliterate English to top 3 Hindi words. For this Hidden Markov Model (HMM) alignment and Conditional Random Fields (CRFs) are used. For HMM alignment GIZA++[4] is used while CRF++[5] is used for training the model. For translation of adjectives and verbs, English-Hindi dictionary shabdanjali[6] is used. Before doing the translation adjectives and verbs are expanded with Wordnet[7]. Table 3.3 shows translation coverage, transliteration errors and Table 3.4 show the word distribution of translated opinions [2] averaged over 5 topics.

[2]http://nlp.stanford.edu/ner/index.shtml
[3]http://nlp.stanford.edu/software/tagger.shtml
[4]http://www.fjoch.com/GIZA++.html
[5]http://crfpp.sourceforge.net/
[6]http://www.shabdkosh.com/archives/content/
[7]http://wordnet.princeton.edu/

Word distribution of translated Opinions				
	Positive	Negative	Neutral	Total
Adjective	151	143	577	871

Table 3.4 Word Distribution of translated Opinions

3.4 Evaluation Metrics

Relevant documents are retrieved for OQs created in each topic. But, documents retrieved may not represent opinion about opinion targets present in OQs. Evaluation is done using recall(R_{oq}), precision(P_{oq}) and F-measure(F_{oq}) for each OQ used for document retrieval to confirm whether OQ terms represent opinion about opinion targets in that document. Equation 3.15 and Equation 3.16 gives the metrics. Similar evaluation is done for subjective sentences using precision(P_s), Recall(R_s) and F-measure(F_s).

$$R_{oq} = \frac{Retrieved\_OQs\_in\_Document}{Total\_OQs\_present\_in\_Document} \tag{3.15}$$

$$P_{oq} = \frac{Relevant\_OQs\_in\_Document}{Retrieved\_OQs\_in\_Document} \tag{3.16}$$

$$F_{oq} = 2 * \frac{P_{oq} * R_{oq}}{P_{oq} + R_{oq}} \tag{3.17}$$

We also calculated mean average precision(MAP) scores to see whether the relevant documents are ranked first for the corresponding OQs. If the MAP scores are high for a model and its corresponding OQ, it can be derived that the model and query is efficient in retrieving more relevant documents first.

3.5 Experiments

In this section we evaluate subjective sentence extraction, identification of opinion about opinion targets and ranking of relevant documents using the proposed approach on the gold standard dataset.

3.5.1 Detecting Subjective Sentences

Subjective sentences are identified using the two methods mentioned in Section 4.2. To analyze the accuracy of proposed methods P_s, R_s and F_s is calculated. Table 3.5 show the average scores for 5 topics. It can observed that Method 2 has 84.1% more recall but 7.3% low in precision compared to Method 1. For opinion mining we feel precision matters more in-order to get accurate and efficient results. This

Topic		Method 1(**M1**)	Method 2(**M2**)
(T1,T2,T3,T4,T5)	P_s	**0.573**	0.534
	R_s	0.543	**1**
	F_s	0.557	**0.696**

Table 3.5 Subjective Sentence Accuracy

Topic		Naive Bayes	SVM	Decision Tree
(T1,T2,T3,T4,T5)	P_s	0.71	**0.73**	0.69
	R_s	0.66	**0.73**	**0.73**
	F_s	0.68	**0.73**	0.71

Table 3.6 Subjective Sentence Classification

analysis was done in next section to analyze the accuracy of opinions retrieved using this two methods. We also did comparison of the following approach with classification methods by 10-cross validation on human annotated sentences using unigrams as features. Table 3.6 show the 10-cross validation results of learning methods.

We can observe that *Method2* achieves 2.3% more F-measure compared to naive bayes, but 4.8% less compared to SVM and 2.0% less compared to decision trees. Similarly, we observe that *Method2* does not fair well in getting good F-measure. But was able to achieve good precision scores compared to learning methods. Since our approaches are unsupervised and achieves decent accuracy in identifying subjective sentences compared to supervised learning approaches. We feel these approaches for resource scarce languages can show significant results in identifying subjective sentences.

3.5.2 Detecting Opinions about Opinion Bearers

In our retrieval approach, query terms are used to confirm their presence in the document. So, SLM is created from sentences obtained using *Method 1*(M1) and *Method 2*(M2). SLM is then extended with IN for forming OQs for retrieval. Our approach is compared with other standard retrieval approaches like LM based retrieval, LM with IN retrieval using lemur toolkit[8]. Since each topic can produce as many OQs given by Equation 3.18 from English collection. Only those queries which retrieved documents

[8]http://www.lemurproject.org/

are considered for evaluation.

$$Total\_OQ's = Total\_NE\_Hindi * OW \tag{3.18}$$

$$OW = (Total\_Adj + Total\_VB's) * (NumofOQ's) \tag{3.19}$$

P_{oq}, R_{oq} and F_{oq} is calculated for 5 topics using baseline LM, LM with IN, SLM and SLM with IN based retrieval using Dirichlet smoothing techniques given by Table 3.7.

Model		UQ	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6
Baseline LM	P_{oq}	0.156	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	**0.270**	-	-	-	-	-	-
SLM+M1	P_{oq}	0.125	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	0.222	-	-	-	-	-	-
SLM+M2	P_{oq}	0.156	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	**0.270**	-	-	-	-	-	-
LM+IN	P_{oq}	0.156	0.076	0.093	0.214	0.214	0.214	0.250
	R_{oq}	1.000	0.406	1.000	0.397	0.397	0.397	0.125
	F_{oq}	**0.270**	0.128	0.171	0.278	0.278	0.278	0.167
SLM+IN+M1	P_{oq}	0.125	0.272	0.093	0.333	0.333	0.333	0.250
	R_{oq}	1.000	0.343	1.000	0.375	0.375	0.375	0.125
	F_{oq}	0.222	**0.304**	0.171	**0.352**	**0.352**	**0.352**	0.167
SLM+IN+M2	P_{oq}	0.156	0.076	0.167	0.214	0.214	0.214	0.250
	R_{oq}	1.000	0.406	1.000	0.437	0.437	0.437	0.125
	F_{oq}	**0.270**	0.128	**0.286**	0.288	0.288	0.288	0.167

Table 3.7 Results obtained using **Dirichlet** Smoothing

Jelinek-Mercer smoothing techniques is also used for the experiments and given by Table 3.8. In each column best performing model and its corresponding query is highlighted.

3.5.3 Documents Ranking

We analyzed the efficiency of OQs and models in retrieving the relevant documents first using mean average precision(MAP) scores. For that we calculated the MAP@4 scores of all the queries which retrieved at-least 4 documents. Average MAP@4 scores are calculated for 5 topics using baseline LM, LM with IN, SLM and SLM with IN based retrieval using Dirichlet smoothing technique given by Table 3.9. While, Table 3.10 show an average MAP@4 scores are calculated for 5 topics using baseline LM, LM with IN, SLM and SLM with IN based retrieval using Jelinek-Mercer smoothing technique. Figure 3.5

Model		UQ	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6
Baseline LM	P_{oq}	0.116	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	0.207	-	-	-	-	-	-
SLM+M1	P_{oq}	0.125	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	0.222	-	-	-	-	-	-
SLM+M2	P_{oq}	0.156	-	-	-	-	-	-
	R_{oq}	1.000	-	-	-	-	-	-
	F_{oq}	**0.270**	-	-	-	-	-	-
LM+IN	P_{oq}	0.116	0.076	0.081	0.204	0.204	0.204	0.250
	R_{oq}	1.000	0.406	1.000	0.397	0.397	0.397	0.125
	F_{oq}	0.207	0.128	0.149	0.269	0.269	0.269	0.167
SLM+IN+M1	P_{oq}	0.125	0.272	0.093	0.333	0.333	0.333	0.250
	R_{oq}	1.000	0.343	1.000	0.375	0.375	0.375	0.125
	F_{oq}	0.222	**0.304**	0.171	**0.352**	**0.352**	**0.352**	0.167
SLM+IN+M2	P_{oq}	0.156	0.076	0.156	0.214	0.214	0.214	0.250
	R_{oq}	1.000	0.406	1.000	0.437	0.437	0.437	0.125
	F_{oq}	**0.270**	0.128	**0.270**	0.288	0.288	0.288	0.167

Table 3.8 Results obtained using **Jelinek-Mercer** Smoothing

show MAP@4 scores calculated for 5 different topics using SLM+IN+M1 model and different OQs using dirichlet smoothing techniques. While, Figure 3.6 shows MAP@4 calculated for 5 different topics using SLM+IN+M1 model and different OQs using jelinek-mercer smoothing techniques.

3.6 Result Discussion

It can observed that the best performing queries from Table 3.7 which used Dirichlet smoothing technique are SQ3, SQ4 and SQ5 of SLM+IN+M1 model in terms of F-measure for retrieving opinions about opinion targets. These queries in SLM+IN+M1 model outperformed LM+IN model by 26.6% in F-measure. Similar comparison was made between the performance of SQ3, SQ4, SQ5 of models SLM+IN+M1 and SLM+IN+M2. It showed that recall of SLM+IN+M1 model is 16.5% low, but performed 22.2% better in F-measure and 55.6% more in precision than SLM+IN+M2. This is contrasting to subjective sentence extraction results. Although the SLM+IN+M2 model had large coverage of sentences. The extracted sentences had weak subjective clues which just improved its recall. But, SLM+IN+M1 model had strong subjective sentences which improved its precision and F-measure.

We can also observe and derive from table 3.7 that unstructured queries (UQ) in all the models achieve 100% recall. But, precision levels vary between the models and are less compared to structured

MAP@4 (Dirichlet)							
Model	UQ	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6
Baseline LM	0.277	-	-	-	-	-	-
SLM+M1	**0.295**	-	-	-	-	-	-
SLM+M2	0.285	-	-	-	-	-	-
LM+IN	0.277	0.294	0.275	0.310	0.310	0.320	**0.322**
SLM+IN+M1	**0.295**	**0.314**	**0.295**	**0.325**	**0.320**	**0.392**	0.275
SLM+IN+M2	0.285	0.312	0.283	0.314	0.310	0.361	0.275

Table 3.9 MAP@4 Values with Dirichlet Smoothing

MAP@4 (Jelinek-Mercer)							
Model	UQ	SQ1	SQ2	SQ3	SQ4	SQ5	SQ6
Baseline LM	0.284	-	-	-	-	-	-
SLM+M1	**0.310**	-	-	-	-	-	-
SLM+M2	0.300	-	-	-	-	-	-
LM+IN	0.284	0.298	0.282	0.302	0.310	0.320	**0.310**
SLM+IN+M1	**0.310**	**0.348**	**0.324**	**0.315**	**0.320**	**0.351**	0.294
SLM+IN+M2	0.300	0.334	0.310	**0.315**	0.315	0.344	0.294

Table 3.10 MAP@4 Values with Jelinek-Mercer Smoothing

queries in the same and across models. This can be attributed to the retrieval of documents containing only query words but not documents which are opinions about opinion targets. This clearly shows the need for structured queries to achieve better performance.

Similar analysis is done for Table 3.8 which uses Jelinek-Mercer smoothing. We can observe that SQ3, SQ4 and SQ5 of SLM+IN+M1 model outperforms other methods and queries in terms of F-measure. These queries in SLM+IN+M1 model perform 30.8% better compared to LM+IN model in F-measure though its recall is 5.8% low. This is observed as LM+IN does not have any constraints in sentence selection, while SLM+IN+M1 have only subjective sentences. There is same difference observed as in Dirichlet smoothing when compared between SLM+IN+M1 and SLM+IN+M2.

Different smoothing methods did not show much difference in best performing queries, although, they had minor differences in queries like SQ2.

From the Figure 3.5 and Figure 3.6 we can observe that opinion query SQ5 in SLM+IN+M1 model performs better than other queries in retrieving relevant document first for different smoothing techniques. This shows the correlation between the F-measure, as we saw that SQ5 outperforms other

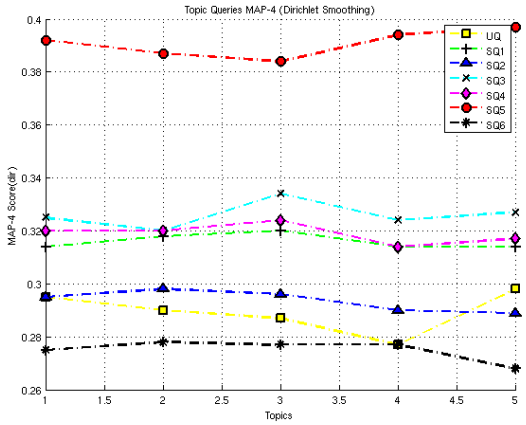

Figure 3.5 MAP@4(Dirichlet) for SLM+IN+M1 Model

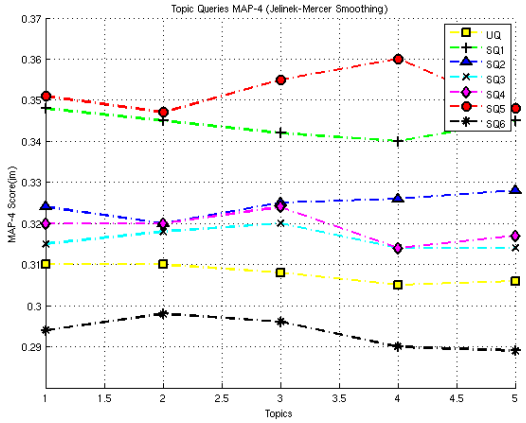

Figure 3.6 MAP@4(JM) for SLM+IN+M1 Model

queries in different models. In conclusion, we can say that SQ5 of SLM+IN+M1 can be used to mine opinions to achieve decent accuracy if the resources in the language are less.

3.7 Summary

In this chapter, we treated opinion mining in resource-scarce languages as a retrieval problem by leveraging a comparable corpora. Adjectives, verbs and named entities are treated as opinion words. Named Entities are depicted as opinion targets, while adjectives and verbs are considered as possible opinions. To extract opinion words resource rich language like English is leveraged and are cross ported to resource-scarce languages to mine opinions. Structured opinion queries are formed using opinions

and opinion targets and are retrieved using LM, LM with IN, SLM and SLM with IN having different smoothing techniques to confirm their presence in the documents. We found that SLM with IN performs better for opinion mining.

Chapter 4

Sentence-level Subjectivity Detection with Feature Selection

Earlier chapter emphasized the need for identification of subjective sentences for fine-grained opinion mining. Solutions provided in the earlier chapter for subjective sentence identification are language dependent and are not scalable. Also, accuracy of classification between objective and subjective sentences needs to be improved. In-order to meet these different challenges, we ask the following questions to improve the performance and accuracy of classification.

- First, can language portability problem be eliminated by selecting language independent features.
- Second, can language specific tools like POS taggers, Named Entity recognizers dependency can be minimized as they vary with language.
- Third, can accuracy of subjective classification is maintained after feature reduction using feature selection methods which are consistent across languages.
- Fourth, can training data be prepared for resource-scarce languages to adhere the needs of subjective and objective text classification.

4.1 Background

Subjective and objective classification task is divided into unsupervised, multilingual and supervised methods.

4.1.1 Unsupervised

Sentiment classification and opinion analysis can be considered as a hierarchical task of subjectivity detection. Improvement of precision in subjectivity detection can benefit the later. Therefore, lot of

work is done for subjective sentence detection to achieve later. Murray [32] proposed to learn subjective expression patterns from both labeled and unlabeled data using n-gram word sequences. Their approach for learning subjective expression patterns is similar to [47] which relies on n-grams, but goes beyond fixed sequences of words by varying levels of lexical instantiation.

4.1.2 Multilingual

In the multilingual space good amount of work is done in Asian and European languages. Several participants in the Chinese and Japanese Opinion Extraction tasks of NTCIR-6 [48] performed subjectivity and sentiment analysis in languages other than English. Banea.et.al and Mihalcea.et.al [27, 5] performed subjectivity analysis in Romanian. While, Mihalcea.et.al [26] performed subjectivity analysis in French, Spanish, Arabic, German, Romanian languages.

4.1.3 Supervised

Furthermore, tools developed for English were used to determine sentiment or subjectivity labeling. For any given target language the text in the language is transfered to English and then an English classifier is applied on the resulting data. The labels were then transfered back into the target language [42]. These experiments are carried out in Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Spanish, and Romanian. Wan [43] constructed a polarity co-training system using the multilingual views obtained through the automatic translation of product-reviews into Chinese and English.

4.2 Proposed Approach

Subjective sentence classification is treated as a text classification task. Mihalcea.et.al [27] used unigrams at word-level as features to classify the subjective and objective sentences in different languages. But, unigrams can occur in different categories (subjective and objective) with equal probability. This hampers the classification accuracy. Also, selecting all possible words in the sentences can create a large index size when considered as entire training set increasing the dimensionality of the feature vector for each sentence.

Feature selection can be applied to efficiently categorize the sentences. It is an important process which is followed for many text categorization tasks [8, 49]. Now to achieve our objective of language independent subjective classification. We use feature extraction, weighing and selection methods that

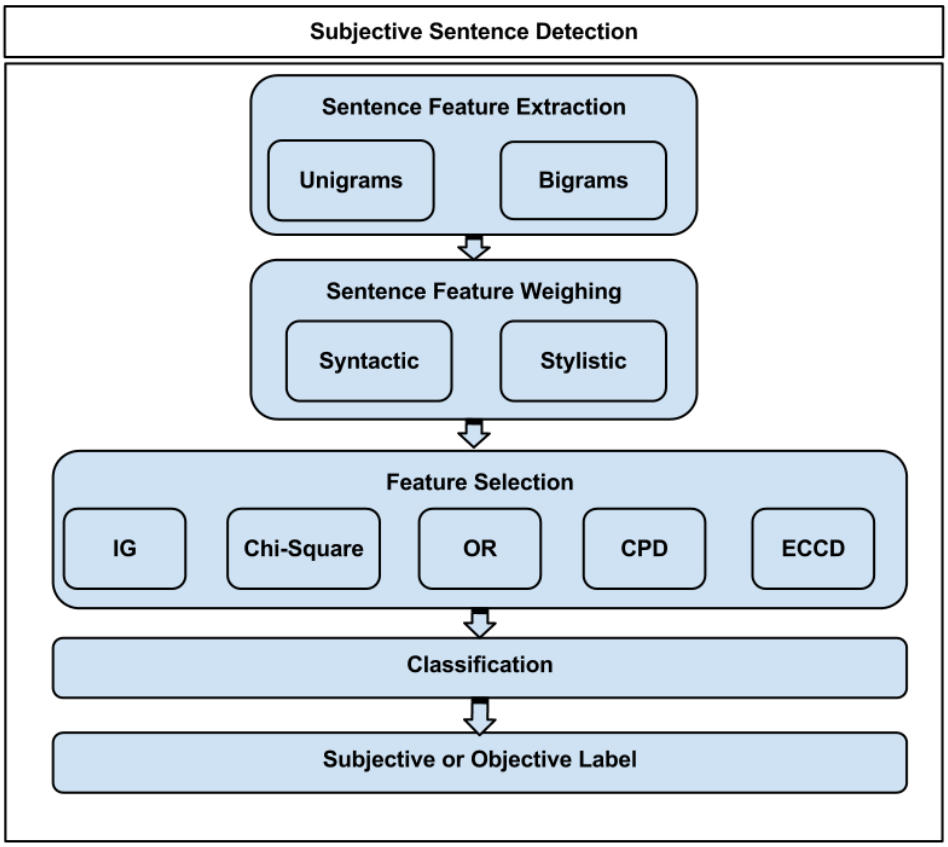

Figure 4.1 Outline of the Approach

are language independent. Figure 4.1 show the outline of the approach used to achieve the objective and subjective sentence classification.

4.2.1 Feature Extraction and Weighing

Features are categorized into syntactic, semantic, link-based, and stylistic features [11] from the previous subjective and sentiment studies. Here, we concentrate more on feature weighing methods based on syntactic and stylistic properties of the text to maintain language independence. Unigrams and Bigrams extracted as features are weighed as given below.

4.2.1.1 Syntactic Feature Weighing

Syntactic features used in earlier works like Gamon.et.al [12] concentrated on word n-grams and part-of-speech (POS) tags. But, POS tagging create dependency on language specific tools. In order to

eliminate the language specific dependencies we will use only word n-grams.

Sentence Representation with Unigram (UF.ISF)

This feature extraction is inspired from vector space model [36] used for flat documents. **UF** represents the unigram frequency at word level in a sentence. While **ISF** represent the inverse sentence frequency of the unigram. For a given collection S of subjective and objective sentences, an Index $I = \{u_1, u_2, ..., u_{|I|}\}$, where $|I|$ denotes the cardinal of I, gives the list of unigrams u encountered in the sentences S.

A sentence s_i of S is then represented by a vector $s_i^{\rightarrow} = (w_{i,1}, w_{i,2}, ..., w_{i,I})$ followed by the subjective or objective label. Here, $w_{i,j}$ represents the weight of unigram u_j in the sentence s_i. Now to calculate the weight $w_{i,j}$ we use the formula similar to TF.IDF.

$$w_{i,j} = \frac{c_{i,j}}{\Sigma_l c_{i,l}} * log \frac{|S|}{|\{s_i : u_j \in s_i\}|} \tag{4.1}$$

where $c_{i,j}$ is the number of occurrences of u_j in the sentence s_i normalized by the number of occurrences of other unigrams in sentence s_i, $|S|$ is total number of sentences in the training set and $|\{s_i : u_j \varepsilon s_i\}|$ is number of sentences in which the unigram u_j occurs at-least once.

Sentence Representation with Bigram (BF.ISF)

This feature extraction is similar to UF.ISF mentioned in the earlier section, but we extract co-occurring words. **BF** represents the Bigrams frequency at word level in a sentence. While **ISF** represent the inverse sentence frequency of the Bigram. For a given collection S of subjective and objective sentences, an Bigram Index $BI = \{b_1, b_2, ..., b_{|BI|}\}$, where $|BI|$ denotes the cardinal of BI, gives the list of bigrams b encountered in the sentences S.

A sentence s_i of S is then represented by a vector $s_i^{\rightarrow} = (wb_{i,1}, wb_{i,2}, ..., wb_{i,BI})$ followed by the subjective or objective label. Here, $wb_{i,j}$ represents the weight of bigram b_j in the sentence s_i. Now to calculate the weight $wb_{i,j}$ we use the formula similar to UF.ISF.

$$wb_{i,j} = \frac{c_{i,j}}{\Sigma_l c_{i,l}} * log \frac{|S|}{|\{s_i : b_j \in s_i\}|} \tag{4.2}$$

where $c_{i,j}$ is the number of occurrences of b_j in the sentence s_i normalized by the number of occurrences of other bigrams in sentence s_i, $|S|$ is total number of sentences in the training set and $|\{s_i : b_j \varepsilon s_i\}|$ is number of sentences in which the bigram b_j occurs at least once.

4.2.1.2 Stylistic Feature Weighing

Structural and lexical style markers can be considered as stylistic features which has shown good results in Web discourse [1]. However, style markers have seen limited usage in sentiment analysis research. Gamon.et.al [12] tried in this direction.

Sentence representation with Normalized Unigram Word Length (NUWL)

This feature extraction considers length of unique unigram words in the sentence. Length of unigram is calculated by the number of characters present in the word. For a given collection S of subjective and objective sentences, an Word Index $WI = \{uw_1, uw_2, ..., uw_{|WI|}\}$, where $|WI|$ denotes the cardinal of WI, gives the list of unigram words uw encountered in the sentences S.

A sentence s_i of S is then represented by a vector $s_i^{\rightarrow} = (lw_{i,1}, lw_{i,2}, ..., lw_{i,I})$ followed by the subjective or objective label. Here, $lw_{i,j}$ represents the weight of unigram word uw_j in the sentence s_i. Now to calculate the weight $lw_{i,j}$.

$$lw_{i,j} = \frac{L_{i,j}}{\Sigma_n L_{i,n}} * log \frac{|S|}{|\{s_i : uw_j \in s_i\}|} \tag{4.3}$$

where $L_{i,j}$ is the character count in the uw_j in the sentence s_i normalized by length of all the unigram words in sentence s_i. $|S|$ is total number of sentences in the training set and $|\{s_i : uw_j \varepsilon s_i\}|$ is number of sentences in which the unigram uw_j occurs atleast once.

4.2.2 Feature Selection

Feature selection methods help in removing the features which may not be useful for categorization. To achieve it, feature selection techniques select subset of total features. But, it is important to reduce features without compromising on the accuracy of a classifier. Most methods like Information Gain(IG) [24], Correlation Feature Selection(CFS) [13], Chi-Squared (χ^2) [3], Odds ratio (OR) [11] does not consider the frequency of the text or term between the categories which leads in reduction of accuracy of a classifier.

In-order to overcome this problem, we used Entropy based category coverage difference (ECCD) [23] feature selection method which uses the entropy of the text or term. f_j is used to represent the text feature extracted (unigram or bigram), c_k for category of the class and c_k^- for the complement of the class. Where j represent number of features and k represents two classes either subjective or objective.

Entropy based category coverage Difference(ECCD)

This feature selection method [23] was proposed to mine INEX XML documents. We use this approach for improving the subjective and objective sentence classification. Let T_j^k be number of occurrences of text feature f_j in the category c_k sentence and, fq_j^k is the frequency of f_j in that category c_k given by Equation 4.4.

$$fq_j^k = \frac{T_j^k}{\sum_k T_j^k} \tag{4.4}$$

So Entropy $Ent(f_j)$ of text feature f_j is given by Equation 4.5

$$Ent(f_j) = \sum_{k=1}^{r} (fq_j^k) * log_2(fq_j^k) \tag{4.5}$$

Entropy equals 0, if the text feature f_j appears only in one category. It means that feature has good discrimination ability to classify the sentences. Similarly, entropy of the text feature will be high if the feature is represented in two classes. If Ent_m represent the maximum entropy of the feature f_j, ECCD(f_j, c_k) is given by following equation 4.6.

$$ECCD(f_j, c_k) = P(f_j|c_k) - P(f_j|c_k^-) * \frac{Ent_m - Ent(f_j)}{Ent_m} \tag{4.6}$$

Where $P(f_j|c_k)$ and $P(f_j|c_k^-)$ are probability of observing the text feature f_j in a sentence belonging to category c_k and c_k^- respectively.

The advantage of ECCD is that higher the number of sentences of category c_k containing feature f_j and lower the number of sentences in other category containing f_j, we get higher value for equation 4.6. It means f_j becomes the characteristic feature of that category c_k which helps in better feature selection. A feature selection method which is similar to ECCD is mentioned below.

Categorical Proportional Difference (CPD)

CPD [37] is a measure of the degree to which a text feature contributes to differentiating a particular category from other categories in a text corpus. We calculate the CPD for a text feature f_j by taking a ratio that considers the number of sentences in subjective category c_k in which the text feature occurs and the number of sentences in objective category c_k^- in which the text f_j also occurs. Equation 4.7 shows the details. Certain threshold of CPD score is kept to reduce the number of features.

$$CPD(f_j, c_k) = \frac{P(f_j, c_k) - P(f_j, c_k^-)}{P(f_j, c_k) + P(f_j, c_k^-)} \tag{4.7}$$

	Subjective	Objective
f_j	A	B
f_j^-	C	D

Table 4.1 Contingency Table

FS	Representation
$IG(f_j, c_k)$	$-\frac{A+C}{M}log(\frac{A+C}{M})+$ $\frac{A}{M}log(\frac{A}{A+B}) + \frac{C}{M}log(\frac{C}{C+D})$
$\chi^2(f_j, c_k)$	$\frac{M(A*D-B*C)^2}{(A+B)*(A+C)*(B+D)*(C+D)}$
$OR(f_j, c_k)$	$\frac{D*A}{C*B}$
$CPD(f_j, c_k)$	$\frac{A-B}{A+B}$
$ECCD(f_j, c_k)$	$\frac{(A*D-B*C)*Ent_m-Ent(f_j)}{(A+C)*(B+D)*Ent_m}$

Table 4.2 Estimation Table

4.2.3 Contingency Table Representation of features

Feature selection methods mentioned in earlier section is estimated using a contingency table. Let A, be the number of sentences in the subjective category containing feature f_j. B, be the number of sentences in the objective category containing f_j. C, be the number of sentences of subjective category which do not contain feature f_j given by f_j^- and D, be the number of sentences in objective category c_k^- which do not contain f_j. Let ($M = A + B + C + D$) be the total possibilities. Table 4.1 represents the above mentioned details. Using the Table 4.1 each of the feature selection methods can be estimated. Table 4.2 show the details.

4.3 Experimental Setup

In-order to achieve subjective and objective classification at sentence-level for different languages. We performed our experiments using different datasets.

4.3.1 Datasets

Hindi experiments were performed using sentences from the news corpus [31] tagged with positive, negative and objective sentences mentioned in earlier Chapter 3. Positive and negative sentences are combined into subjective sentences to do subjectivity analysis.

Also, in-order to check the consistency of approach on different languages, we used translated MPQA corpus provided in [26] containing subjective and objective sentences[1] of French, Arabic, and Romanian languages for the experiments. For English, MPQA corpus[2] containing subjective and objective sentences are used.

We also performed experiments on Telugu sentence-level sentiment tagged dataset mentioned in Chapter 2 to check the consistency of approach on different domains.

4.4 Evaluation

To evaluate various feature selection methods, we use F-measure scores which combines precision and recall. Precision (P_s) measures the percentage of sentences correctly assigned to subjective category, and recall (R_s) measures the percentage of sentences that should have been assigned to subjective category, but actually assigned to subjective category. Using P_s and R_s subjective F-measure F_s is calculated. Similarly, Objective F-measure F_o is calculated using P_o and R_o. After F-measure is determined for both subjective and objective class, the macro-average F-measure $F_{macro-avg}$ is determined by the following Equation 4.8.

$$F_{macro-avg} = \frac{\Sigma_{i=o,s} F_i}{2} \tag{4.8}$$

4.5 Experiments

Initially, 1500 subjective and 1500 objective sentences of English, Romanian, French and Arabic languages are used to perform the experiments. While for Hindi, entire corpus constituting 786 subjective and 519 objective sentences were used. Different feature weighing and selection methods are evaluated with 2 different classifiers to obtain best combination for each language. Table 4.3 to Table 4.7 show the Macro-Average ($F_{macro-avg}$) scores obtained after 10 cross-validation using sequential minimal optimization algorithm [34] for training a support vector machine(SVM) using polynomial kernel and Naive Bayes Multinomial(NBM) classifiers.

Feature space obtained after application of feature selection methods for each language is mentioned in the Table 4.8.

Once the best combination is obtained for each language. It is compared with multilingual space clas-

[1] http://lit.csci.unt.edu/index.php/Downloads

[2] http://www.cs.pitt.edu/mpqa

[SVM]

Feature Selection	UF.ISF	BF.ISF	NUWL
None	0.705	0.660	0.705
CFS	0.710	0.670	0.705
IG,OR,χ^2	0.680	0.665	0.685
CPD	**0.840**	0.805	0.835
ECCD	0.830	0.805	0.830

[NBM]

Feature Selection	UF.ISF	BF.ISF	NUWL
None	0.745	0.690	0.740
CFS	0.730	0.685	0.725
IG,OR,χ^2	0.735	0.690	0.730
CPD	0.855	**0.925**	0.890
ECCD	0.850	**0.925**	0.875

Table 4.3 $F_{macro-avg}$ - English

[SVM]

Feature Selection	UF.ISF	BF.ISF	NUWL
None	0.685	0.665	0.680
CFS	0.715	0.675	0.695
IG,OR,χ^2	0.695	0.635	0.690
CPD	0.845	0.815	**0.850**
ECCD	0.845	0.815	0.845

[NBM]

Feature Selection	UF.ISF	BF.ISF	NUWL
None	0.740	0.685	0.730
CFS	0.735	0.690	0.740
IG,OR,χ^2	0.745	0.685	0.725
CPD	0.865	0.935	0.890
ECCD	0.865	**0.940**	0.885

Table 4.4 $F_{macro-avg}$ - Romanian

sifier proposed in [26][3] along with the baseline constituting simple Naive Bayes classifier with unigram features. Multilingual space constitutes words as features from all languages used for experiments except Hindi.

Scalability of feature selection methods is an issue. In-order to understand the performance of ECCD feature selection method with classifiers. In every iteration 500 sentences are added to each class of initial 1500 subjective and objective sentences limiting to maximum of 3500 to get average scores. Table 4.9 show the comparison of average scores obtained for each language.

Figures 4.2 show precision and recall for subjective sentences obtained using different methods for English and Romanian using different number of sentences. While, Figures 4.3 show precision and recall for subjective sentences obtained using different methods for French and Arabic using different number of sentences.

Telugu sentence-level subjectivity analysis was done on different domains using the combination of ECCD and BF.ISF with NBM classifier. Due to low kappa scores and very low inter-annotators agreement. Experiments were performed on the two datasets prepared by annotators. Table 4.10 and Table 4.11 show the baseline apporoach without any feature selection method and our approach which used feature selection method ECCD.

[3]Note that this research used the entire dataset which had unequal subjective and objective sentences. We used equal number of subjective and objective sentences taken each time from dataset. So, experiments using this method are again performed on our dataset.

	Feature Selection	UF.ISF	BF.ISF	NUWL
[SVM]	**None**	0.695	0.685	0.685
	CFS	0.710	0.695	0.685
	IG,OR,χ^2	0.690	0.685	0.685
	CPD	**0.855**	0.825	0.850
	ECCD	0.845	0.820	0.835

	Feature Selection	UF.ISF	BF.ISF	NUWL
[NBM]	**None**	0.730	0.705	0.725
	CFS	0.730	0.710	0.725
	IG,OR,χ^2	0.725	0.690	0.710
	CPD	0.860	0.940	0.900
	ECCD	0.845	**0.950**	0.885

Table 4.5 $F_{macro-avg}$ - French

	Feature Selection	UF.ISF	BF.ISF	NUWL
[SVM]	**None**	0.670	0.660	0.675
	CFS	0.710	0.665	0.680
	IG,OR,χ^2	0.665	0.645	0.660
	CPD	0.855	0.825	0.850
	ECCD	0.850	0.830	**0.860**

	Feature Selection	UF.ISF	BF.ISF	NUWL
[NBM]	**None**	0.720	0.690	0.710
	CFS	0.730	0.695	0.725
	IG,OR,χ^2	0.720	0.690	0.710
	CPD	0.910	**0.915**	0.910
	ECCD	**0.915**	**0.915**	**0.915**

Table 4.6 $F_{macro-avg}$ - Arabic

4.6 Analysis

Below we see the analysis of the results and the performance difference between SVM and NBM.

4.6.1 Results

It is observed from the Table 4.3 to Table 4.7 that ECCD feature selection and BF.ISF feature weighing method with NBM classifier performs consistently across languages. This behavior is observed due to the capability of ECCD in efficiently discriminating the features belonging to a particular class. Although, UF.ISF and NUWL with ECCD and CPD using SVM classifier has more scores than BF.ISF,

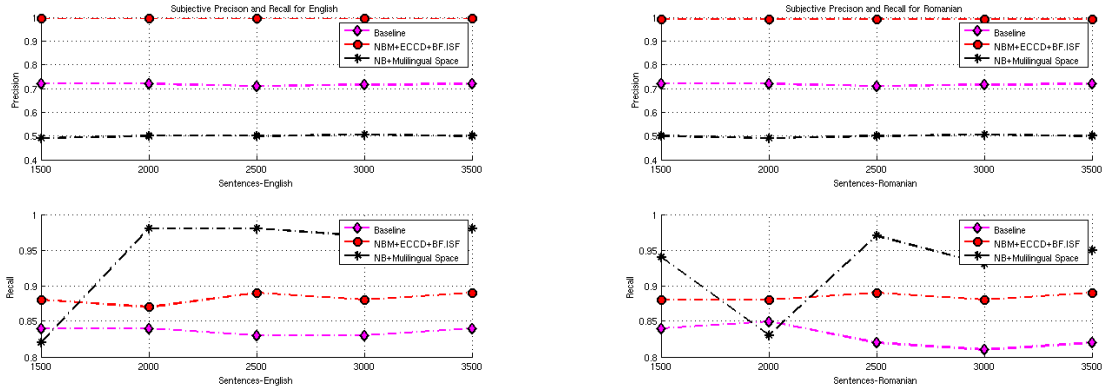

Figure 4.2 Subjective Precision and Recall for English and Romanian

[SVM]

Feature Selection	UF.ISF	BF.ISF	NUWL
None	0.665	0.655	0.650
CFS	0.635	0.655	0.600
IG,OR,χ^2	0.665	0.660	0.650
CPD	0.760	**0.845**	0.755
ECCD	0.735	**0.845**	0.735

[NBM]

Feature Selection	UF.ISF	BF.ISF	NUWL
None	0.615	0.655	0.635
CFS	0.460	0.440	0.460
IG,OR,χ^2	0.580	0.655	0.605
CPD	0.590	0.845	0.660
ECCD	0.555	**0.850**	0.655

Table 4.7 $F_{macro-avg}$ - Hindi

Feature Selection		English	Romanian	French	Arabic	Hindi
None	Unigrams (UF.ISF,NUWL)	100	100	100	100	100
	Bigrams (BF.ISF)	100	100	100	100	100
CFS	Unigrams (UF.ISF,NUWL)	1.8	1.7	1.4	1.2	2.3
	Bigrams (BF.ISF)	8.4	7.2	5.7	3.9	0.7
IG,OR,χ^2	Unigrams (UF.ISF,NUWL)	60	60	60	60	60
	Bigrams (BF.ISF)	60	60	60	60	60
CPD	Unigrams (UF.ISF,NUWL)	66.2	65.5	68.5	71.8	58.1
	Bigrams (BF.ISF)	90	90.1	88.6	92.3	81.1
ECCD	Unigrams (UF.ISF,NUWL)	65.7	65	68	71.4	57.4
	Bigrams (BF.ISF)	90	90	88.5	92.2	80.9

Table 4.8 Feature Space Used(%)

it is not significantly contrasting with the results of NBM classifier. So, NBM classifier with ECCD feature selection and BF.ISF feature weighing method which obtained high $F_{macro-avg}$ scores is selected for comparison with other approaches in Table 4.9. The proposed method not only outperforms on $F_{macro-avg}$ compared to other approaches but also on $P_{macro-avg}$ and $R_{macro-avg}$ in all languages.

For English, proposed methods gains 23.8% over baseline in $F_{macro-avg}$ and 8.0% on $P_{macro-avg}$ on [44]. Similarly, from Table 4.9 it can be deducted that proposed method for Romanian attains 26.1% more $F_{macro-avg}$ than baseline and 155.4% more compared to Multilingual space classifier [26]. Similar observations can be made for other languages. Even though [26] attains high recall for every language. It fails to attain high precision due to presence of large number of frivolous word features which

Language (Method)		P_s	R_s	F_s	P_o	R_o	F_o	$P_{macro-avg}$	$R_{macro-avg}$	$F_{macro-avg}$
English	Baseline	0.720	0.830	0.770	0.800	0.676	0.733	0.760	0.753	0.751
	NBM + BF.ISF + ECCD	**1.000**	0.865	**0.925**	**0.875**	**1.000**	**0.935**	**0.937**	**0.932**	**0.930**
	NB + MultiLingual Space [26]	0.497	**0.927**	0.644	0.350	0.057	0.087	0.423	0.491	0.365
	Wiebe & Riloff [44]	0.904	0.342	0.466	0.824	0.307	0.447	0.867	0.326	0.474
	Chenghua Lin [15]	0.710	0.809	0.756	0.716	0.597	0.651	0.713	0.703	0.703
Romanian	Baseline	0.713	0.830	0.766	0.796	0.663	0.723	0.755	0.746	0.745
	NBM + BF.ISF + ECCD	**1.000**	0.880	**0.940**	**0.890**	**1.000**	**0.940**	**0.945**	**0.940**	**0.940**
	NB + MultiLingual Space [26]	0.497	**0.913**	0.640	0.383	0.063	0.096	0.440	0.488	0.368
French	Baseline	0.703	0.826	0.760	0.790	0.643	0.713	0.746	0.736	0.736
	NBM + BF.ISF + ECCD	**1.000**	0.905	**0.950**	**0.915**	**1.000**	**0.955**	**0.957**	**0.952**	**0.952**
	NB + MultiLingual Space [26]	0.490	**0.913**	0.636	0.370	0.056	0.096	0.430	0.485	0.366
Arabic	Baseline	0.703	0.800	0.750	0.770	0.666	0.713	0.736	0.733	0.731
	NBM + BF.ISF + ECCD	**1.000**	0.845	**0.915**	**0.865**	**1.000**	**0.925**	**0.932**	**0.922**	**0.920**
	NB + MultiLingual Space [26]	0.497	**0.983**	0.656	0.293	0.006	0.016	0.353	0.495	0.336
Hindi	Baseline	0.680	0.900	0.770	0.690	0.350	0.460	0.685	0.625	0.615
	NBM + BF.ISF + ECCD	**0.810**	**1.000**	**0.900**	**1.000**	**0.650**	**0.790**	**0.905**	**0.825**	**0.850**

Table 4.9 Comparison of Average scores between proposed and other approaches

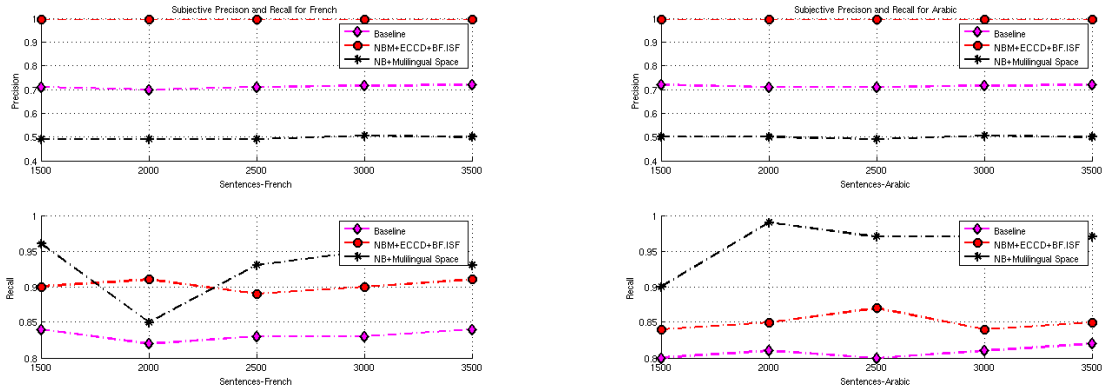

Figure 4.3 Subjective Precision and Recall for French and Arabic

are common for both classes. This being major drawback, ECCD feature selection method eliminates features which attains zero entropy. This reduces the randomness of features and leave only those features which are more eligible for discriminating the classes. Combining ECCD with BF.ISF, a language independent weighing method for Bi-gram features extracted from the sentences. We are able to attain a best classification accuracies which are consistent across languages.

Figures 4.2 and 4.3 also show that increase in number of sentences does not effect the precision of the proposed method, as it still outperforms other methods. But, scalability problem persists for ECCD with BF.ISF for larger datasets, as it may not eliminate less random features due to noise and other constraints. Also, feature selection methods ensure the performance of classifiers is maintained

Domain (Method)		P_s	R_s	F_s	P_o	R_o	F_o	$P_{macro-avg}$	$R_{macro-avg}$	$F_{macro-avg}$
Entertainment	NBM + BF.ISF	0.200	0.940	0.330	0.950	**0.250**	0.390	0.575	0.595	0.360
	NBM + BF.ISF + ECCD	**0.210**	**0.970**	**0.340**	**0.970**	**0.250**	**0.400**	**0.590**	**0.610**	**0.370**
Language	NBM+BF.ISF	**0.07**	0.810	0.120	**0.960**	**0.270**	**0.420**	**0.515**	0.540	0.270
	NBM + BF.ISF + ECCD	**0.07**	**0.840**	**0.130**	**0.960**	**0.270**	**0.420**	**0.515**	**0.555**	**0.275**
General-News	NBM + BF.ISF	0.350	0.900	0.500	0.910	0.390	0.550	0.630	0.645	0.525
	NBM + BF.ISF + ECCD	**0.370**	**0.940**	**0.530**	**0.950**	**0.420**	**0.580**	**0.660**	**0.680**	**0.555**
Politics	NBM+BF.ISF	0.800	**0.230**	0.350	0.420	0.910	0.580	0.610	0.570	0.465
	NBM + BF.ISF + ECCD	**0.900**	**0.230**	**0.370**	**0.440**	**0.960**	**0.600**	**0.670**	**0.595**	**0.485**
Spirituality	NBM+BF.ISF	0.140	0.960	**0.240**	0.960	0.130	0.220	0.550	0.545	0.230
	NBM + BF.ISF + ECCD	**0.140**	**0.970**	**0.240**	**0.970**	**0.140**	**0.240**	**0.555**	**0.555**	**0.240**
Misc	NBM+BF.ISF	**0.310**	0.960	0.460	0.870	**0.110**	**0.200**	0.590	0.535	0.330
	NBM + BF.ISF + ECCD	**0.310**	**0.980**	**0.470**	**0.940**	**0.110**	**0.200**	**0.625**	**0.545**	**0.335**

Table 4.10 Telugu Sentence-level Analysis **(Annotator-1)**

Domain (Method)		P_s	R_s	F_s	P_o	R_o	F_o	$P_{macro-avg}$	$R_{macro-avg}$	$F_{macro-avg}$
Entertainment	NBM + BF.ISF	**0.370**	0.980	**0.540**	0.970	**0.250**	0.390	0.670	0.615	0.465
	NBM + BF.ISF + ECCD	**0.370**	**0.990**	**0.540**	**0.990**	**0.250**	**0.410**	**0.680**	**0.620**	**0.475**
Language	NBM+BF.ISF	0.060	0.990	0.120	1.000	0.140	0.250	0.530	0.565	0.185
	NBM + BF.ISF + ECCD	0.060	0.990	0.120	1.000	0.140	0.250	0.530	0.565	0.185
General-News	NBM + BF.ISF	0.270	0.780	0.400	0.870	0.400	0.550	0.570	0.590	0.475
	NBM + BF.ISF + ECCD	**0.290**	**0.790**	**0.420**	**0.890**	**0.450**	**0.590**	**0.590**	**0.620**	**0.505**
Politics	NBM+BF.ISF	0.370	0.860	0.520	0.680	0.170	0.270	0.525	0.515	0.395
	NBM + BF.ISF + ECCD	**0.400**	**0.930**	**0.560**	**0.830**	**0.200**	**0.330**	**0.615**	**0.565**	**0.445**
Spirituality	NBM+BF.ISF	0.200	0.940	0.330	0.890	0.110	0.200	0.545	0.525	0.265
	NBM + BF.ISF + ECCD	**0.210**	**0.990**	**0.350**	**0.990**	**0.120**	**0.210**	**0.600**	**0.555**	**0.280**
Misc	NBM+BF.ISF	**0.280**	0.980	0.430	0.960	0.190	0.320	0.620	0.585	0.375
	NBM + BF.ISF + ECCD	**0.280**	**1.000**	**0.440**	**0.990**	**0.200**	**0.330**	**0.635**	**0.600**	**0.385**

Table 4.11 Telugu Sentence-level Analysis **(Annotator-2)**

by reducing number of features. But, it does not ensure reduction in fixed percentage of features. As observed our best performing feature selection method ECCD reduces feature size by 10% only.

Experiments performed on Telugu blog sentences belonging to different domains show a slight increase in the accuracy of classification from the baseline approach using bigram features and NBM classifier. When ECCD feature selection method used with the existing baseline approach showed on average increase of 3.38% and 4.87% for each domain using the annotator-1 and annotator-2 datasets respectively. Highest accuracy of classification was attained by "General-News" domain, while least accuracy was observed for domain "spirutuality" for annotator-1 and "Language" for annotator-2 dataset. Differences in the accuracies of classification can be attributed to two major reasons.

- Uneven distribution of samples for a class.

- Due to the complexity in understanding and annotating data. Mislabeling or error in labeling the subjective sentences as objective or viceversa caused the decrease in accuracy.

4.6.2 Performance Analysis between SVM and NBM

From the Table 4.3 to 4.7 it is observed that SVM with some feature selection and weighing methods performs equivalent to the NBM. However, as the number of documents increases the performance of SVM may degrade. It can be derived that, as the training data size increases, it is rare to see SVM performing better than NBM.

4.6.2.1 Training Time behavior

SVM is in a clear disadvantage compared to NBM when processing time is considered. The training time of the SVM is particularly high, especially for larger feature spaces. It is probably attributed to the time taken in finding the proper separating hyperplane.

4.6.2.2 Features behavior

Large feature spaces do not necessarily lead to best performance. So feature selection methods are used to create small feature spaces to build SVM and NBM classifiers. Sometimes, small feature space sizes make SVM perform equivalent to NBM as observed in Table 4.7. Thus, this would explain why SVM is outperformed for small training set sizes and for small feature spaces with large training sets.

4.7 Summary

In this chapter, subjective classification is achieved using combination of feature selection and weighing methods which are consistent across languages. We found that our proposed method which combines ECCD feature selection and BF.ISF feature weighing method used along with NBM classifier performs better than other existing feature selection methods across different languages. It not only outperforms other feature selection methods, but also achieve better scores compared to other approaches.

Chapter 5

Conclusions

In this thesis, we aimed to solve the challenges in mining opinions at fine-grained level from different languages that lack state-of-the-art tools and resources for subjectivity and opinion analysis. To achieve it, we designed approaches that use minimal language specific resources and tools and achieve decent accuracy.

Initially, fine-grained opinion mining problem in resource-scarce language is handled by leveraging resource rich language like English from a comparable corpora. Identification of opinion targets and opinions embedded in a resource-scarce language is treated as a retrieval problem instead of a classification problem. This helped us to surpass the limitation of word identification tools in resource-scarce languages. Only language specific resources that are used in the retrieval approach were bi-lingual dictionaries and a parallel corpus of named entities. Bi-lingual dictionaries were used for translation of adjectives and adverbs treated as possible opinions that are extracted from English collection of comparable corpora. While parallel corpus of named entities is used for HMM character alignment and building a CRF model. Named entities identified in English collection of comparable corpora treated as possible opinion targets are transliterated using the built model. Translated and transliterated words are further used to identify subjective sentences and also to form structured opinion queries in the resource-scarce languages to mine opinions. Once the subjective sentences are identified in a document, a subjective language model(SLM) is built along with Inference network (IN) framework having different smoothing techniques to confirm the presence of opinions on opinion targets in the documents. We found that SLM with IN performs better for opinion mining then other techniques like baseline language model (LM), LM with IN and basic SLM.

But, during the experiments it was observed that accuracy of identifying subjective sentences can be further improved. Though the aforementioned approach to identify subjective sentences is completely

unsupervised, dependency on translated and transliterated words reduce the accuracy. Thus, we used human annotated sentence labels to built a classifier that classifies subjective and objective sentences. Even though it has dependency on human annotations, we made sure that the approach eliminates dependency on language specific word identification tools and resources. Features extraction, weighing and selection is done in such a way that it is consistent across different languages and domains. We experimented our approach on English, south Asian languages Hindi, Telugu and European languages French, Romanian, Arabic.

Overall, we proposed an alternate approach to mine opinions in resource-scarce language by leveraging resource rich language like English and also attempted to solve the sub-problem of sentence-level subjectivity analysis using feature selection methods. Also, to facilitate further research in the field, we build sentence-level polarity annotated datasets in Hindi and Telugu.

Chapter 6

Future Work

Our future research is focused on solving the problems in opinion mining and subjectivity detection by improving theoretical foundations that represent the problem using robust mathematical models. Also, analysis on the results is obtained by extensive experimental evaluation. Below, we present the possible future work required in improving opinion mining and subjectivity analysis in resource-scarce languages.

6.0.1 Fine-Grained Opinion Mining

Proposed retrieval approach for opinion mining needs to be experimented on different resource-scarce languages. This involves designing more complex queries and framework that improves F-measure by retrieving relevant opinions about opinion targets. Currently, our framework depends on a comparable corpora to mine opinions. We need to eliminate this dependency by designing better methods. Also, our approach find only opinion targets and opinions on them. There is a need to mine opinion holders who hold these opinions on the targets. Scalability issues are of concern in transfer or cross-language settings. It needs to be addressed with efficient sampling techniques. Also, language independent techniques to identify words needs to be explored as current technique to identify NE's, adjectives and adverbs depends on dictionaries for translation and transliteration.

6.0.2 Sentence-level Subjectivity Analysis

Subjective sentences were identified in single domain. In a practical setting, identifying labeled data in every domain is not possible. Cross-domain classification creates the necessity. Achieving accuracy in cross-domain setting in multiple languages is a challenging task due to large classification errors. We

need to identify and design solutions using domain adaptation approaches. Also, we need to test the scalability of the subjectivity detection approach on bigger and larger datasets.

6.0.3 Possible Extensions

Subjective analysis can be further used to find the sentiment in the text. Also fine-grained opinion mining can extented to other entites like organizations and events and the user mood can be predicted on them.

6.0.4 Pragmatics

Understanding the context of the information plays an important role in improving the accuracy of opinion mining. In the future work, we would like to extend the opinion extraction approach by adding word sense disambiguation.

Related Publications

- **A. Mogadala** and V. Varma. *Language Independent sentence-level subjectivity analysis with feature selection.* In 26th Pacific Asia Conference on Language Information and Computing (PACLIC 26), Bali, Indonesia. (2012)
- **A. Mogadala** and V. Varma. *Retrieval approach to extract opinions about people from resource scarce language News articles.* In ACM KDD Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM), Beijing, China. (2012)

Also Used Ideas from

- V. Varma and **A. Mogadala**. *Issues and Challenges in Building Multilingual Information Access Systems.* Emerging Applications of Natural Language Processing: Concepts and New Research. Bandyopadhyay, S., Naskar, S.K., Ekbal, A. eds., IGI Global. (2012) **[Book Chapter]**

Bibliography

[1] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS), Volume 26, 12, ACM.*, 2008.

[2] P. Arora, A. Bakliwal, and V. Varma. Hindi subjective lexicon generation using wordnet graph traversal. In *Proceedings of CICLing.*, 2012.

[3] J. Bakus and M. Kamel. Higher order feature selection for text classification. *Knowledge and Information Systems, Volume 9.*, pages 468–491., 2006.

[4] A. Balahur, E. Boldrini, A. Montoyo, and P. Martínez-Barco. Opinion and generic question answering systems: a performance analysis. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers.*, pages 157–160, 2009.

[5] C. Banea, R. Mihalcea, and J. Wiebe. Learning multilingual subjective language via cross-lingual projections. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, Volume 45, 976.*, 2007.

[6] M. J. Carman, S. Gerani, and F. Crestani. Proximity-based opinion retrieval. In *SIGIR*, 2010.

[7] W. B. Croft and D. Metzler. Combining the language model and inference network approaches to retrieval. In *Information Processing and Management: an International Journal,*, 2004.

[8] L. D. D. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217. Association for Computational Linguistics, 1992.

[9] W. Du and S. Tan. An iterative reinforcement approach for fine-grained opinion mining. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 486–493. Association for Computational Linguistics, 2009.

[10] J. Duan, X. Chen, B. Pei, Y. Hu, and R. Lu. A new method for sentiment classification in text retrieval. In *IJCNLP, 1-9.*, 2005.

[11] G. Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research, Volume 3.*, pages 1289–1305., 2003.

[12] M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics, 841, ACL.*, 2004.

[13] M. Hall. Correlation-based feature selection for machine learning. *Phd Thesis, The University of Waikato.*, 1999.

[14] S. Harsh, P. Pingali, S. Ganesh, and V. Varma. Statistical transliteration for cross language information retrieval using hmm alignment model and crf. In *In Workshop on CLIA addressing the Information Need of Multilingual Societies.*, 2008.

[15] Y. He, C. Lin, and R. Everson. Sentence subjectivity detection with weakly-supervised learning. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP), Volume 2, 2.*, 2011.

[16] E. Hovy and S.-M. Kim. Automatic detection of opinion bearing words and sentences. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP)*, 2005.

[17] E. Hovy and S.-M. Kim. Extracting opinions expressed in online news media text with opinion holders and topics. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text at COLING-ACL.*, 2006.

[18] E. Hovy and S.-M. Kim. Identifying and analyzing judgment opinions. In *Proceedings of HLT/NAACL*, 2006.

[19] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the National Conference on Artificial Intelligence.*, pages 755–760, 2004.

[20] T. Inui, H. Takamura, and M. Okumura. Latent variable models for semantic orientations of phrases. In *11th Meeting of the European Chapter of the Association for Computational Linguistics.*, 2006.

[21] L. John and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR*, 2001.

[22] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.*, pages 355–363, 2006.

[23] C. Largeron, C. Moulin, and M. Géry. Entropy based feature selection for text categorization. In *Proceedings of ACM Symposium on Applied Computing, ACM.*, pages 924–928., 2011.

[24] C. Lee and G. Lee. Information gain and divergence-based feature selection for machine learning-based text categorization. In *Information processing & management, Volume 42.*, pages 155–165., 2006.

[25] Q. Miao, Q. Li, and D. Zeng. Mining fine grained opinions by using probabilistic models and domain knowledge. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 358–365, 2010.

[26] R. Mihalcea, C. Banea, and J. Wiebe. Multilingual subjectivity: are more languages better. In *Proceedings of the 23rd International Conference on Computational Linguistics.*, pages 28–36, 2010.

[27] R. Mihalcea, J. Wiebe, C. Banea, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, pages 127–135, 2008.

[28] R. Mihalcea, J. Wiebe, C. Banea, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *EMNLP*, 2008.

[29] T. Mitamura, N. Kando, and T. Sakai. Introduction to the ntcir-6 special issue. In *ACM TALIP, 7(2)*, 2008.

[30] A. Mogadala and V. Varma. Finding influence by cross-lingual blog mining through multiple language lists. In *Proceedings of Information Systems for Indian Languages, Communications in Computer and Information Science, Volume 139, Part 1.*, pages 54–59, 2011.

[31] A. Mogadala and V. Varma. Retrieval approach to extract opinions about people from resource scarce language news articles. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM), ACM KDD*, 2012.

[32] G. Murray and G. Carenini. Predicting subjectivity in multimodal conversations. In *Proceedings of Empirical Methods in Natural Language Processing: Volume 3.*, pages 1348–1357, 2009.

[33] V. Ng, S. Dasgupta, and S. M. N. Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *COLING/ACL, 611–618.*, 2006.

[34] J. Platt. Fast training of support vector machines using sequential minimal optimization. advances in kernel methods - support vector learning. *B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press*, 1998.

[35] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *EMNLP*, pages 105–112, 2003.

[36] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. In *Communications of the ACM, Volume 18.*, pages 613–620., 1975.

[37] M. Simeon and R. Hilderman. Categorical proportional difference: A feature selection method for text categorization. In *Proceedings of the Seventh Australasian Data Mining Conference (AusDM), Volume 87.*, pages 201–208., 2008.

[38] R. Steinberger and A. Balahur. Rethinking sentiment analysis in news: from theory to practice and back. In *Workshop on Opinion Mining and Sentiment Analysis*, 2009.

[39] H. Takamura, Y. Suzuki, and M. Okumura. Application of semi-supervised learning to evaluative expression classification. In *7th International Conference on Intelligent Text Processing and Computational Linguistics.*, 2006.

[40] H. Turtle and W. B. Croft. Evaluation of an inference network based retrieval model. In *Trans. Inf. Syst9(3).*, pages 187–222., 1991.

[41] H. Turtle, T. Strohman, D. Metzler, and W. B. Croft. Indri: A language model-based serach engine for complex queries. In *International Conference on Intelligence Analysis.*, 2004.

[42] L. Vijayarenu, M. Bautin, and S. Skiena. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM).*, 2008.

[43] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1,*, pages 235–243, 2009.

[44] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of Computational Linguistics and Intelligent Text Processing*, pages 486–497, 2005.

[45] J. Wiebe, T. Wilson, and M. Bell. Identifying collocations for recognizing opinions. In *ACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation.*, 2001.

[46] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. In *Proceedings of the Language Resources and Evaluation, volume 39, issue 2-3.*, pages 165–210., 2005.

[47] T. Wilson and S. Raaijmakers. Comparing word, character, and phoneme n-grams for subjective utterance recognition. In *Ninth Annual Conference of the International Speech Communication Association.*, 2008.

[48] Y. Wu and D. Oard. Ntcir-6 at maryland: Chinese opinion analysis pilot task. In *Proceedings of the 6th NTCIR Workshop on Evaluation of Information Access Technologies.*, 2007.

[49] Y. Y. and P. J. O. A comparative study on feature selection in text categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*, pages 412–420. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

[50] K. Yang. Widit trec blog track:leveraging multiple sources of opinion evidence. In *TREC*, 2008.

[51] C. T. Yu, L. Jia, and W. Zhang. Uic at trec 2008 blog track. In *TREC*, 2008.

[52] T. Zagibalov and J. Carroll. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Conference on Computational Linguistics.*, 2008.

[53] T. Zagibalov and J. Carroll. Unsupervised classification of sentiment and objectivity in chinese text. In *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP)*, pages 304–311, 2008.

[54] C. Zhai, K. Ganesan, and J. Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics.*, pages 340–348, 2010.

[55] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. In *ACM Transactions on Information Systems*, pages 179–214, 2004.