Sentiment Analysis For Hindi Language

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

MS by Research
in
Computer Science

by

Piyush Arora
200702027
piyush.arora@research.iiit.ac.in

Search and Information Extraction Lab (SIEL)
Language Technology and Research Center (LTRC)
International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2013

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

This is to certify that the thesis entitled **"Sentiment Analysis For Hindi Language"** submitted by **Piyush Arora** to the International Institute of Information Technology, Hyderabad, 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

Adviser: Prof. Vasudeva Varma

To my Family and Teachers

Acknowledgments

I would like to thank my advisor Prof Vasudeva Varma who has played an important role by helping and guiding in all the aspects ranging from the research area to the other general discussions which helped me a lot to define my problem more concretely and also develop a better positive outlook towards approaching the problem. Another set of people with whom I have been fortunate to be associated with are Prof Rajeev Sangal, Prof Dipti Misra and Dr Prasad Pingali. They all helped me to understand what Research is all about and how to go about for the solution of same, at first it is important to understand the nature of the problem and then moving towards designing the solution. Apart from research, the interactions with all helped me to know more about life, understand and analyze things in a much better way and thus leading to my overall development.

I started my research at LTRC with Machine Translation, the journey has been good enough working from core aspects of language on parsers, machine translation which made a base for my research work. As it is said that "if foundation is strong you can make a big and better building over it" so the credit of this foundation equally goes to the faculty members of LTRC - Dr. V. Sriram, Dr Samar Hussain and Dr. Soma Poul for providing their support and guidance .

Next, I would also like to thank my seniors Pranav Kumar Vasistha, Abhilash Inumella, Sudhir Kolachina, Avinesh PVS, Kushal Dave, Manasi Verma, Sambhav Jain for their guidance, suppport and to the various fruitful discussions that happens with them. The discussions with Pranav, AbhilashI helped me a lot to learn more about myself, my potential and to act upon it. Whenever I faced some problem in general, they were the first person I looked upon for a solution. Discussion with Sudhir and Avinesh sir helped me a lot to know more about research not in just my field across diverse fields. I really admire one note which Avinesh sir gave me while he was guiding me in the Transfer-Grammar Engine- *I was always disappointed when some new bugs came but he made me understand that bugs are good for system, bugs means that you need to improve more and the minor things you didn't account earlier are been getting rectified and these bugs only leads to making things better and developing a better robust system the same scenario can be mapped to our Life.* I didn't spend much time with Kushal Dave and Manasi Verma but the limited time frame helped a lot and motivated me to move forward , try my best. Once you understand the problem then go and give it your best shot, try all the ideas, approaches you

feel so and there will be people to guide you and help you out if you ever get struck.

This acknowledgement would be incomplete without mention of my friends - Akshat Bakliwal, Ankit Patil, Karan Jindal, Rahul Namdev, Manish Aggarwal and a lot more. There is a lot to say about each one of them Akshat Bakliwal- this thesis wouldn't be there without his help, long discussions on the problems understanding them, analyzing them and working towards solving them. Karan Jindal- always motivated me with his hard-work and dedication. Ankit Patil and Manish Aggarwal- had some of the best discussion ranging from Research, Life, College and other. Rahul Namdev- the hardworking guy, he is really a good friend and companion, the long philosophical discussions had helped a lot in the discourse of understanding and leading a purposeful life.

I can't forget my lab mates especially Aditya Mogadala he has been constantly providing his advice, feedback on the work. I also want to acknowledge my wing mates without whom the things wouldn't have been so easy and interesting, their help and motivation persuaded me to learn, study and do research more effectively. I also enjoyed the company of my juniors (Ravi Aggarwal, Naveen Pawar, Dhruv Data and many more) which helped and made my stay pleasant and thus I have been able to give back to my work. I would like to thank all, for their support, encouragement and foremost for making this journey memorable. I couldn't conclude without expressing thanks to IIIT-Hyderabad for providing such a good environment to live and do whatever you want. Working in lab till early morning , playing whenever you feel like and approaching towards professors so easily whenever you have problems, encouraging new ideas and a learning atmosphere where - you do work , make mistakes and learn.

Above all I would like to express my gratitude towards my parents, my sisters who have been constantly encouraging and supportive. The credit of all my work and for my present state goes to them. I am deeply thankful to them for their eternal love and support.

Some lines by William Ernest Henley which has always motivated me and are my inspiration source-

" Out of the night that covers me, Black as the Pit from pole to pole,

I thank whatever gods may be For my unconquerable soul.

In the fell clutch of circumstance, I have not winced nor cried aloud.

Under the bludgeonings of chance My head is bloody, but unbowed.

Beyond this place of wrath and tears, Looms but the Horror of the shade,

And yet the menace of the years Finds, and shall find, me unafraid.

It matters not how strait the gate, How charged with punishments the scroll.

I am the master of my fate:I am the captain of my soul "

Abstract

Sentiment Analysis is an area of focus over the last decade. Increase in user-generated content provide an important aspect for the researchers, industries and government(s) to mine this information. The user-generated content is one important source for various organizations to know/learn/identify the general expression/sentiment of different users on the product.

In this work, we focus on mining sentiments and analyzing them for Hindi language. Hindi is the 4th commonly spoken language[1] in the world. With the increase in the amount of information being communicated via regional languages like Hindi, comes a promising opportunity of mining this information. Mining sentiments in Hindi comes with their share of issues and challenges. Hindi is morphologically rich and is a free order language as compared to English, which adds complexity while handling the user-generated content. The scarcity of resources for the Hindi language brings challenges ranging from collection and generation of datasets. We take up this challenge and work towards building resources- reviews, blogs annotated corpora and subjective lexicon for Hindi language. We propose a technique to build a subjective lexicon given a pre-annotated seed list for a language and its wordnet representing the network/connectivity of words using synonyms and antonyms relations. One of the salient features of this technique is that the method can be applied for any language which has the wordnet available. To show the applicability of the technique on other languages, we experiment our technique on English language in addition to the Hindi language. The lexicon generated by our algorithm is evaluated using the following different metrics.

1. Comparing against Manual Annotation
2. Comparing against similar Existing Resources
3. Classification Performance

In addition to resource creation, we take up the task of sentiment classification in Hindi Language. We work on two different genres of Hindi User-Generated Web Content-

1. Reviews
2. Blogs

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

For both of these genres we present three different approaches for performing sentiment classification such as-

1. Using Subjective Lexicon
2. N-Gram Method
3. Weighed N-Gram

We aim at analysing the merits and demerits of each of the above approaches across the different genres for the sentiment classification task. We discuss in detail the problems and the issues while working with the user-generated content (reviews and blogs) in Hindi language. This research work, throws some light on the main differences between the User-Generated Content in English and Hindi language at linguistic and its representation level and the approaches followed to address the same.

English language provides the option of leveraging the abundant resources and tools that have been developed in the past, the work for Indian Languages has just began since last decade and it is in early stage of research and development, so our focus has been on- "*To effectively mine the subjective information from the user-generated content in Indian languages, overcoming the data scarcity challenges associated with such problems.*".

Contents

List of Figures

List of Tables

Chapter 1

Introduction

User-generated content is an important source of information to mine the sentiment/opinion of people on different products and services. The era of Web 2.0 has resulted in generation of vast amount of user-generated content. The developing technology with ease of reachability and better connectivity has lead to wide spread use of blogs, forums, e-news, reviews channels and the social networking platforms such as Facebook, Twitter. These social networking platforms has exponentially increased the amount of information generated on daily basis. Thus mining the data and identifying user sentiments, wishes, likes and dislikes is one of an important task that has attracted the focus of research community from last decade. The world wide web plays a crucial role in gathering public opinion, these opinions play an important role in making business related decisions. To obtain the factual and subjective information on companies and products, analysts are turning towards web to gather information. Extracting public opinion from this information is a major task. Industrialists spend a large chunk of their revenue on business intelligence to read minds of general public and interpret what they think about their product(s). Sentiment analysis tries to mine information from various text forms such as reviews, news, blogs and classify them on the basis of their polarity as positive, negative or neutral.

In the following sections, we describe some of the basic concepts and definitions to make the discussion in the further chapters simpler and easier to understand.

1.1 Sentiment Analysis - an introduction

Sentiment Analysis deals with analyzing emotions, feelings, the attitude of a speaker or a writer from a given piece of text. "*Sentiment analysis or opinion mining refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials*" (Source: Wikipedia). Sentiment Analysis involves capturing of user's behavior, likes and dislikes of an individual from the generated web content. There is no concrete definition of "Sentiments", but in general they are considered as thoughts, views and attitude of a person arising mainly based on the emotion instead of a reason. Sentiments are considered as the manifestation of our feelings

and emotions. This field of computer science deals with analyzing and predicting the hidden information stored in the text. This hidden information provide valuable insights about user's intentions, taste and likeliness. Sentiment Analysis focus on categorizing the text at the level of subjective and objective nature. Subjectivity indicates that the text contains/bears opinion content whereas Objectivity indicates that the text is without opinion content.

Some examples-

1. **Subjective-** *This movie by Aamir Khan and Kajol is superb.*
   (this sentence has an opinion, it talks about the movie and the writer's feelings about same "superb" and hence it's subjective)
2. **Objective-** *This movie stars Aamir Khan and Kajol.*
   (this sentence is a fact, general information rather than an opinion or a view of some individual and hence it's objective)

The subjective text can be further categorized into 3 broad categories based on the sentiments expressed in the text.

1. **Positive-** *I love to watch Star Tv series.*
2. **Negative-** *The movie was awful.*
3. **Neutral-** *I usually get hungry by noon.* (this sentence has user's views, feelings hence it is subjective but as it does not have any positive or negative polarity so it is neutral .)

1.2 User-Generated Content

Web 2.0 & 3.0 has lead to an exponential increase in the user-generated content. It has provided varied mechanisms for the users to interact with the web. User-generated content is available on the web in various forms such as-

1. **Weblogs-** Collection of blogs, lists of blogs mainly as hyperlinks to be used as a mechanism for recommendation and suggestion to other blogs and websites.
2. **News-** The online-news catering to daily events, activities and other information across the globe with a rapid reachability and that too in a multilingual fashion.
3. **Discussion forums-** The web-pages pertaining to an individual person, products, services, industries etc thus providing a mechanism to link the consumers/users with the other consumers/users and suppliers/companies. The users can interact through different mechanisms such as- providing feedback, requesting information, asking query, expressing opinion etc.
4. **Reviews-** The e-commerce and the entertainment industry has been one of the largest booming industry across the globe. The web presence of same has provided a platform to the consumers

to share their views and feelings about the products and services and thus help the other fellow beings in making optimal choices and decisions.

The technological advancement in the past two-three decades have enabled humans to find different ways to interact with each other. One big revolution that became the part of the internet era is the social network revolution. These social networking sites and other platforms leads to the generation of petabytes of data per week. Some of the popular and widely used social networking platforms with its brief overview-

- *Facebook* - The total monthly active users on facebook are around 850 million. Everyday about 250 million photos are uploaded and 2.7 billions likes are made on it. Facebook stores, accesses, and analyzes 30+ Petabytes of user-generated data
- *Twitter* - Twitter has about 465 million accounts and about 175 millions tweets are done per day.
- *Google+* - It has about 90 millions users and about 675,000 users are been added everyday.

The statistics mentioned above gives an idea about the rate at which the web has been increasing. With such vast data generated regularly, it provides enormous business opportunities to handle this data safely and precisely.

1.3 Web Content in Hindi language

Hindi is the 4th largest spoken language and has 490 million speakers across the world majority of whom are from India [1]. With Unicode (UTF-8) standards for Indian languages introduced, web pages in Hindi language have increased on a rapid pace. There are many websites which provide information in Hindi, ranging from various news websites such as http://dir.hinkhoj.com/, http://bbc.co.uk/hindi to sites providing information regarding the culture, music, entertainment and other aspects of arts for example http://www.webdunia.com/, http://www.virarjun.com/, http://www.raftaar.in/ etc.

There are various weblogs which are written in Hindi. With internet reaching to more and more people within the country and in the other parts, the users and contributors to the increasing web space are rising tremendously. With large pool of information generated everyday, we need to properly mine this information to aid the people by analyzing their opinions, views and take the necessary actions. Due to the development in technology and its easy availability the number of users of internet has been increasing and huge amount of data is generated everyday. Thus it becomes an important aspect to analyze the user's intentions and views to improve the services and lead towards a better ecosystem.

Some of the popular news websites in Indian Languages with respect to the number of viewers is mentioned in Table 1.1 .

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers#More_than_100_million_native_speakers

Table 1.1 Common news web sites in Indian Languages

Name of Site	Language	Web Viewers
Malayala Manorama	Malyalam	5 million
Dainik Bhaskar	Hindi	3.5 million
NavBharat Times	Hindi	2.7 million
Amar Ujala	Hindi	2.1 million
Anand Bazar Patrika	Bengali	1.05 million
Gujarat Samachar	Gujarat	1 million

1.4 Mining Sentiments For Hindi language

As discussed above, the number of users and web content for Hindi language has been increasing tremendously. This part of the web hasn't been explored much in the direction of sentiment and opinion analysis. Different forms of user-generated content can be used for mining the sentiments accompanied by it. News can be classified as subjective and objective based on the content. Large amount of blogs and reviews are generated everyday, we can perform sentiment analysis and can generate summary using same. We can generate gist of the discussion forums based on the sentiments expressed and the objects being talked about. We can also perform sentiment analysis on available stories and poems and summarize them based on the sentiments associated with each entity.

Sentiment Analysis for Hindi has its own challenges and problems as it's a Scarce Resource Language. Scarce resource languages are the languages for which the availability of tools, annotated corpus and other resources is limited and under the development phase. This work, focuses on building resources and performing Sentiment Analysis for Hindi language. We also discuss and address some of the challenges faced while mining sentiments from the user-generated content.

1.5 General Approaches

Much of the work has been done for the English language with less focus been given to the Scarce Resource Languages. Some of the common and popular approaches used for sentiment analysis are-

- **Using Subjective Lexicon**- Subjective Lexicons are a list of words where for each word we have the scores indicating the positive, negative and objective nature of the word. In this approach, for a given text, we sum the score of the polar/subjective terms using the subjective lexicon. Finally, we get a combined positive, negative and objective score which gives an idea about the subjective nature of the text. The category with maximum score defines the polarity of a document.

- **Using N-Gram Modeling**- In this approach, from a given training data we make N-Gram model (uni-gram, bi-gram , tri-gram and combination of same) and perform classification of the test data using the model formed.

- **Using Machine Learning**- In this method, we extract features from the text and learn the model using the training corpus by selecting a set of relevant features. While forming features, information is incorporated at different levels such as syntactic information, lexical information, part of speech information, negation words such as not, no, none which reverses the polarity, abbreviation, punctuations etc to perform supervised or semi-supervised learning .

1.6 Main Challenges

Some of the general challenges while addressing the problem of sentiment analysis are-

- **Unstructured Data**- The data available on the internet is very unstructured, there are different forms of the data talking about the same entities, persons, places, things and events. The web contains data from different sources varying from books, journals, web documents, health records, companies logs, internal files of an organization and even data from multimedia platforms comprising of texts, images, audios, videos etc. The diverse sources of the data makes the analysis more complex as the information is coming in different formats.

- **Noise (slangs, abbreviations)**- The web content available is very noisy. In today's era of 140 characters texting, for their ease people use various abbreviations, slangs, emoticons in normal text which makes the analysis more complex and difficult.
  Eg mvie ws awsummm :D
  The web content reports a large number of spelling variations for the same word. Eg a word **awesome** can be found in various forms as- "*awsum, awssuummm, awsome*" the repetition of the characters can be in any combination.

- **Contextual Information**- Identifying the context of the text becomes an important challenge to address. Based on the context the behavior/use of the word changes in a great aspect.

  Ex-1 The movie was long.

  Ex-2 Lecture was long.

  Ex-3 Battery life of samsung galaxy-2 is long.

  In all the above 3 examples, meaning of long is same- indicating the duration or passage of time. In ex-1 and ex-2 "long" indicates boredness hence a Negative expression whereas in ex-3 "long" indicates efficiency hence a Positive expression.

  With the help of above examples, it's clear that same word with same meaning can have multiple usage depending on the context. So, it becomes important to detect the context to find the subjective information in a text

- **Sarcasm Detection**- *" Sarcasm " is defined as a sharp, bitter, or cutting expression or remark; a bitter jibe or taunt usually conveyed through irony or understatement* (Source: Wikipdeia). It's a

hard task for human beings to interpret sarcasm, making a machine able to understand same is a more difficult task. Some examples of sarcasm -

Ex-1 Not all men are annoying. Some are dead.

Ex-2 What a superb movie by Tushar, I won't watch his movie ever.

- **Word Sense Disambiguation**- The same word can have multiple meanings, and based on the sense of its usage the polarity of the word also changes. Example a word **good** in English has 21 adjective senses, 4 noun senses and 2 adverb senses whose polarity changes with respect to the sense in which it is used as mentioned below in Table 1.2 where Pos, Neg and Obj scores indicates the Positive, Negative and Objective scores for a particular sense of Good.

Table 1.2 Multiple Senses of the word "Good"

Sense	Part of Speech	Polarity Scores	Meaning
good-1	Adjective	Pos-1 Obj-0 Neg-0	morally admirable
good-2	Adjective	Pos-0.375 Obj-0.5 Neg-.125	not left to spoil; "the meat is still good"
good-3	Noun	Pos-0 Obj-1 Neg-0	commodity/articles of commerce

- **Language Constructs**- Each language has its own nature and style of writing which is accompanied by its own challenges and specifications. Challenges while dealing/working with Hindi language are as follows-
  - *Word Order*- Word arrangement in a sentence plays an important role in identifying the subjective nature of the text. Hindi is a free order language i.e the subject, object and verb can come in any order whereas English is a fixed order language i.e subject followed by verb and followed by object. Some examples are mentioned in Table 1.3.

English Sentence	Order of Words
Ram ate three apples	Correct/Valid order SVO
Ate ram three apples	Incorrect/Invalid order VSO
Three apples ate ram	Incorrect/Invalid order OVS
Hindi Sentence	**Order of Words**
राम ने तीन आम खाए	Correct order SOV
तीन आम खाए राम ने	Correct order OVS
खाए तीन आम राम ने	Correct order VOS

Table 1.3 Word Order Variations

Word order plays a vital role in deciding the polarity of a text, in the text same set of words with slight variations and changes in the word order affect the polarity aspect.

– *Morphological Variations*- Handling the morphological variations is also a big challenge for Hindi language. Hindi language is morphologically rich which means that lots of information is fused in the words as compared to the English language where we add another word for the extra information.

As shown in the Table 1.4 the verb "खाया" (khaya) carries far much information apart from

English Sentence	Corresponding Hindi Sentence
"Ram ate food"	राम ने खाना खाया
"Ram is eating food"	राम खाना खा+रहा+हैं
"Swati is eating food"	स्वाति खाना खा+रही+हैं

Table 1.4 Morphological Variations

just the root/padh. It carries the inflection which provide information/idea about the tense, gender and person. Thus with same root there can be many words in a language with varying information i.e multiple variations of same words can have the same root with respect to the sense of tense, gender, person and other information. Some Examples

1) Boy, Girl, Boys - लडका, लडकी, लडके, लडको

2) Eat - खाया, खा रही हैं, खा रहा हैं, खाएगा, खाएगी etc.

– *Handling Spelling Variations*- In Hindi language, same word with same meaning can occur with different spellings, so it's quite complex to have all the occurences of such words in a lexicon and even while training a model it's quite complex to handle all the spelling variants some examples are shown in Table 1.5

संबंध	सम्बंध	सम्बन्ध
महंगा	मंहगा	महँगा
ठंढा	ठंडा	ठण्डा

Table 1.5 Spelling Variations

– *Lack of resources*- the lack of sufficient resources, tools and annotated corpora also adds to the challenges while addressing the problem of sentiment analysis especially when we are dealing with Non-English languages.

In comparison to the English language, Hindi on one side is morphologically rich and on the other side it's a free order language. These properties increase the complexity while analyzing the Hindi web content. The content available on the internet in Hindi language has been in numerous fonts and formats as shown in Table 1.6.

For this research, we focus only on the utf font but believe that this work could be extended to other formats with little tweaking.

Format	Example
English Sentence	Ram ate food
Utf format	राम ने खाना खाया
Wx format	rAma ne KAnA KAyA
Transliterated format	ram ne khana khaya

Table 1.6 Different fonts and formats of User-Generated Content

1.7 Motivation

Understanding emotions, analyzing situations and the sentiments associated with it is the natural ability of a human being. But how efficiently can we train a machine to exhibit same phenomenon becomes an important and vital question to be explored and answered. Sentiment Analysis provides an effective mechanism of understanding individual's attitude, behaviour, likes and dislikes of a user.

Small Story- *"It's believed that when a child is small the mother knows very well what and when he/she is going to need, at what time the child drinks, eats or even the time of difficulty when it cries and the possible causes for same. She works towards rectifying same. She very well knows the difference in the cry for food and cry for getting the diapers changed, thus the mother can analyze and take the necessary action very well"*

Analogous to this small story, how well our lives would be if what we want can be automatically analyzed, suggested and provided to us without putting much efforts? Sentiment analysis provides us with the services and products we want of our taste at our ease. With e-commerce business spreading at a great speed the task of mining opinions on various products becomes an useful resource to guide and help people in making choices and decisions. Mining sentiments and subjective information helps to provide products and services in a personalized fashion and as per individuals taste and likings. With more emphasis laid on personalized information it becomes necessary and important to go about catering information to an individual, based on his likings and taste.
The study of sentiment analysis also provide enough information about how human beings perceive and express information in the form of text to express their feelings and emotions. This wide multi-dimension aspects discussed above, motivated me to take this problem as my Research Problem.

1.8 Applications

Sentiment Analysis has been widely used for understanding the subjective nature of a text. Few areas where Sentiment Analysis can be applied are-

1) **Aid in decision making**- Decision making is an integral part of our life. It ranges from “which products to buy”,“which restaurant to go” to “which bank isurance policies to go for”, “which investments to make”. Sentiment Analysis can be used to decide and select from the available options based on the general opinions expressed by other users.

2) **Designing and Building Innovative Products**- With exposed to tough competition and open to critics through public reviews and opinions, sentiment analysis leads to better analysis of the products in terms of the usability and human-friendly nature. It creates an environment for better and more innovative products.

3) **Recommendation Systems**- Most of the websites we visit have a recommendation system inbuilt to assist us, ranging from sites related to books, online-media, entertainment, music, film industry to other forms of art. These systems uses our personal information, previous history, likes and dislikes and our friends information to make suggestions.

4) **Products Analysis**- With the help of sentiment analysis it has become easier to analyze different products and make the choices accordingly. This kind of analysis also helps to select a product based on its feature specifications. The comparison between two products has also been made quite easier.

5) **Business Strategies**- Much of the business strategies are been guided with respect to the response from the people. Companies aims to satisfy the needs and demands of the users, thus strategic moves of companies are driven through public opinions and views. With the world connected through technology events have a global impact, the issue/failure on one part of the world has an impact on the other corner of the globe. So it becomes quite important to drive products/services according to the public viewpoint.

6) **User Modelling**- From research perspective it has led to better designing of interfaces, more interactive designs where emphasis is on Human Computer Interaction- the mechanism of how a user interacts with the system/machine.

1.9 Problem Statement

This work concentrates on analyzing sentiments for Hindi language. The main focus lies on understanding the challenges, issues while working on Hindi language and the approaches followed while performing sentiment classification for the user-generated content. With very few tools and annotated corpora available, it’s a challenging task to perform sentiment analysis for Hindi language. The work done comprises of resource generation which involves building of annotated datasets and subjective lexicon. We propose an algorithm to create Subjective Lexicon for a language using a pre-annotated seedlist and wordnet of that language. The technique has been

applied for English and Hindi language.

We perform sentiment classification for two genres-

1) *Reviews*- The product and movie reviews in Hindi language are used for performing task of sentiment classification.

2) *Blogs*- About 250 blogs were crawled and are used for performing sentiment classification at blog level exploiting the inherent nature of the blogs. The aim is to perform Blog Sentiment Classification and understand the nature and complications for Hindi language as compared to that of English language

1.10 Contribution of this Thesis

In this research, our primary focus is on Hindi language. As stated above, Hindi language lacks adequate resources to perform sentiment analysis. Our work involves building annotated corpus open for public use and analyzing the same for sentiment extraction.

We proposed an algorithm to build subjective lexicon for under-resourced languages, using the wordnet of that language. Our method of building a subjective lexicon is dependent only on WordNet and a small pre-annotated seed list. Using WordNet and simple graph traversal method, we construct the subjectivity lexicon. In our method, initially a small seed list of words is decided along with their polarity and using WordNet this seed list is populated based on the synonyms and the antonyms of the words in the list. Here, we make an assumption that synonyms possess similar polarity and antonyms show opposite polarity. The algorithm has been tested and validated for Hindi and English language.

We perform sentiment classification for the following genres of web content in Hindi language-

1) **Product and Movie Reviews** - We performed experiments with different approaches for sentiment classification of Hindi reviews. In the initial experiments we applied traditional methods such as N-Gram and learning from subjective lexicon. We used various pre-processing tasks and other optimization to capture the subjective phrases more accurately. We used negation handling, performed stemming and morphological analysis to capture the maximum information from the text. We discuss the pos and cons of the approaches used and show how effectively a model of combined lexical and N-Gram features perform while addressing the sentiments classification at reviews level.

2) **Blogs**- We present different methodologies of sentiment classification at blogs level which are widely used. We comment on the merits and demerits of these approaches in the context of the blog sentiment classification. In the first method, we use subjective lexicon(s) to perform the

classification task. We report the results in terms of accuracy of the classification and also discuss where and why these lexicons fails to classify the blogs correctly. In second method, we use N-Gram technique to classify blog sentiments and also discuss the pros and cons of this approach.

Overall the work done in this research has been to understand the nature of user-generated content for Hindi language and to build resources, design algorithms that would lead to the development in the field of sentiment analysis and opinion mining for Hindi language.

1.11 Thesis Summary and Organization

The thesis has been organized as follows- *Chapter 2* discuss the related work done in the area of sentiment analysis which is splitted into three sub parts 2.1 Sentiment Analysis, 2.2 Lexicon Generation and 2.3 Indian Language Sentiment Analysis, *Chapter 3* presents the information about the resources developed and also give an account of the resources that are used for this research. *Chapter 4* focus on Hindi Subjective Lexicon generation with detailed discussion on various aspects ranging from what Hindi Subjective Lexicon is, the algorithm used, the uses of subjective lexicon and the issues related to same. *Chapter 5* discuss the work done towards sentiment classification at Reviews as well as at Blog Level. We present the conclusions and the possible future extensions of this research work in *Chapter 6*.

Chapter 2

Related Work on Sentiment Analysis

Sentiment Analysis has been the focus of research community from last decade. There has been large amount of work done for English language, the work for Indian Languages has just began. The initial lexicon and dictionary based approaches for extracting sentiments has given way to machine learning approaches using the syntactic and semantic features. In this chapter, we discuss the work done in the past in the area of Sentiment Analysis. To make it simple and easy to understand we have divided the related work into three sections-

2.1 Sentiment Analysis

Identifying the sentiment polarity is a complex task, to address the task of sentiment classification various methodologies have been applied earlier. Most common, widely used approaches for identifying sentiments for a given piece of text are as follows-
1) **Syntactic Approaches**- Syntactic approach towards sentiment classification using N-Grams have been used by Bo Pang, Lillian Lee and Vaithyanathan[35]. They used the traditional n-gram approach along with POS information as a feature to perform machine learning for determining the polarity. They used Naive Bayes Classification, Maximum Entropy and Support Vector Machines on a three fold cross validation. In their experiment, they tried different variations of N-Gram approach like unigams presence, unigrams with frequency, unigrams+bigrams, bigrams, unigrams + POS, adjectives, most frequent unigrams, unigrams + positions. They concluded from their work that incorporating the frequency of matched n-gram might be a feature which could decay the accuracy. Maximum accuracy achieved by them among all the experiments they performed was 82.9% which was obtained in unigrams presence approach on SVM.

2) **Semantic/Pattern Mining**- Semantic approaches using part of speech learning has also been used quite popularly for identifying sentiments in a text. Turney and Benamara used this approach for binary classification in [40] and [5]. Much work has also been done in the field of extracting

sentiment expressions using various NLP techniques. Nasukawa and Yi[33], Bloom, Garg and Argamon[7] used techniques like word sense disambiguation, chunking, n-gram to perform binary polarity classification .
Ohana and Tierney [34], Saggion and Funk[38] used sentiwordnet to perform opinion classification. They calculated positive and negative score for a review and based on the maximum score, the polarity of the review was assigned. They also extracted features and used machine learning algorithms to perform sentiment classification. Turney [40] also worked on part of speech (POS) information. He used tag patterns with a window of maximum three words (i.e) till trigrams. In his experiments, he considered JJ, RB, NN, NNS POS-tags with some set of rules for classification. His work is extension to the work done on adjectives alone by Hatzivassiloglou and McKeown[20] they consider patterns like RB, NN/NNS. Given a phrase he calculates the PMI (Point-wise Mutual Information) from the strong positive word "excellent" and also from the strong negative word "poor", and the difference gives the semantic orientation of the phrase. Dave, Lawrence and Pennock devised their own scoring function which was probability based [13]. They performed lexical substitutions for negation handling and used rainbow classifiers to decide the class of the review.

3)**Features/Machine Learning**- Much of the work has also been done towards using machine learning approach for identifying the sentiment expressions. Bo Pang, Lillian Lee and Vaithyanathan [35], Zhang [45], Go, Bhayani and Huang [18] deduced features to perform supervised machine learning . The feature based learning has proved to perform better in comparison to the traditional approaches of syntactic and semantic approaches. The features learned from the N-Grams models along with that of subjective lexicon with little bit of fine tuning perform better as compared to the normal N-Gram and Subjective Lexicon scoring mechanism.

Research in the field of sentiment analysis is done at various levels which are as follows-
1) *Document Level*- The document level analysis deals with classifying the whole document as a single polarity positive, negative or objective. Bo Pang, Lilling Lee and Vaithyanathan[35], Turney [40] performed document level classification .
2) *Sentence Level* The sentence level analysis focus on analyzing the documents at sentence level. The sentences are analyzed individually and classified as objective, negative or positive. The overall document thus has a set of sentences with each sentence being marked with it's corresponding polarity. There has been significant work done by Wiebe, Bruce and O'Hara[42], Yu and Hatzivassiloglou[44], Theresa Wilson [23], Hu and Liu[22] and Kim [30] with respect to the sentence level classification .
3) *Phrase Level*- This analysis involves going much deeper and deals with identifying the phrases in a sentence for a given document and analyze the phrases and classify them accordingly as positive, negative or objective. The phrase level analysis is also known as fine grained analysis going

deep into the text to identify the subjective items/entities and classify same as done by Wilson et al. [43], Agarwal, Biadsky and Mckeown[1] .

The work done in the past in the area of sentiment analysis can be categorized into various genres such as Reviews, News, Blogs analysis etc.
Blogs level analysis- Work done at blog level can be attributed to Chesley [9], Ku, Liang and Chen[27], Zhang [46], Ben He, Macdonald, Jiyin He and Ounis[21], Melville, Gryc and Lawrence [28], Draya, Plantie, Harb, Poncelet, Roch and Trousset[16], Godbole, Srinivasaiah and Skiena[19]
Reviews level analysis- Some of the work done at reviews level for mostly English Language is as follows- Wiebe, Bruce and O'Hara[42], Bo Pang, Lillian Lee and Vaithyanathan[35], P Turney [40], Yu and Hatzivassiloglou[44], Theresa Wilson [23], Hu and Liu [22], Blitzer, Dredze and Pereira[6].
News level analysis- Godbole, Srinivasaiah and Skiena[19], Alexander Balahur[4] worked on News.

All the above categories deal with large text, in addition to this there has been much focus given towards micro-blogs analysis which includes analyzing tweets, forums and chats by Go, Bhayani and Huang[18], Nicholas A and David A[15], Apoorv, Boyi, Ilia, Owen and Rebecca[2], Dmitry, Oren and Ari[14]

Draya, Plantie, Harb, Poncelet, Roche and Trousset[16] tried to identify domain specific adjectives to perform blog sentiment analysis. They considered the fact that opinions are mainly expressed by adjectives and pre-defined lexicons fail to identify domain information. Chesley et al. [9] performed topic and genre independent blog classification, making novel use of linguistic features. Each post from the blog is classified as positive, negative and objective.

Turney [40] worked on product reviews. Turney used adjectives and adverbs for performing opinion classification on reviews. He used PMI-IR algorithm to estimate the semantic orientation of the sentiment phrase. He achieved an average accuracy of 74% on 410 reviews of different domains collected from Epinion. Hu and Liu [22] performed feature based sentiment analysis. Using Noun-Noun phrases they identified the features of the products and determined the sentiment orientation towards each feature. Bo Pang, Lillian Lee and Vaithyanathan[35] tested various machine learning algorithms on Movie Reviews. They achieved 81% accuracy in unigram presence feature set on Naive Bayes classifier.

2.2 Lexicon Generation

In 1966, IBM developed the General Inquirer system [39] which marked the beginning of sentiment extraction from plain text. This system was termed as content analysis research problem in behavior science and comprised of 11789 words with each word having at-least one instance. In

1997, Hatzivassiloglou and Mckeown[20] developed a method to predict semantic orientation of adjectives. Their idea was predicting the semantic orientation of adjectives based on the nature of conjunctive joining the two adjectives. A log-linear regression model uses these constraints to predict whether conjoined adjectives are of same or different orientations, achieving 82% accuracy in this task when each conjunction is considered independently. Janyce Wiebe [41] showed a method to learn subjective adjectives from a corpora based on methods for clustering words according to distributional similarity and a small amount of manually annotated corpora. For English, a good amount of work is done in the lines of generating subjective lexicon. SentiWordNet [17], [3] was developed in year 2006 by Esuli, Sebastiani and Baccianella. It contains four Part-of-Speech tags namely adjectives, adverbs, verbs and nouns with 2 million words out of which 3% are adjectives. Each word is assigned three scores positive, negative and objective such that sum of the score for each word sums to one.
Positive score + negative score + objective score=1 (1)
for example the word “best” has a score as positive-0.75 negative-0.0 and neutral-0.25 which satisfies the above equation 1.
SentiWordNet was built using WordNet and a ternary classifier. Their classifier is based on “bag of synset” model which uses manually disambiguated glosses available from the Princeton WordNet Gloss Corpus. Banea, Mihalcea and Wiebe[8] proposed a bootstrapping method for building subjective lexicon for under-resourced languages. Their method build a subjective lexicon using a small seed list (60 words), an online dictionary (Romanian Dictionary) and a small annotated corpora. They used word level similarity (LSA and PMI) to filter words. In their bootstrapping method the initial seed list was manually selected and contained 60 words which were evenly distributed among adjectives, adverbs, nouns and verbs.

Kamps, Marx, Mikken and Rijke[25] tried to determine sentiments of adjectives in WordNet. In this work, they divided adjectives into four major categories and used base words (to measure relative distance) depending on the category. For category *Evaluative* their base words were “good” and “bad”, for category *Activity* their base words were “active” and “passive”, etc.The polarity orientation of a word ’w’ belongs to range [-1,1], -1 for words on bad side and 1 for words on good side. Based on this method, they populated a total of 1608 words in all four categories with avg. correctness of 67.18% for English.

Kim and Hovy [31] proposed a method of identifying and analyzing judgment opinions. This was a four step process in which first step was recognizing the opinion. For identifying the opinion they introduced an algorithm to classify a word as positive, negative or objective which was based on WordNet. They made an assumption which was to add synonyms of a word with the same polarity as the source word. To avoid words with multiple meaning (dual nature) they applied a method to identify closeness of a word to each category (positive, negative, objective). For their proposed method to give high recall the initial seed list should be large enough and with wide variety of words.

Delip Rao and Deepak Ravichandran [36] presented an extensive study on the problem of detecting polarity of words. The authors considered bi-polar classification of words i.e. a word can be either positive or negative. They performed semi-supervised label propagation in graph for polarity detection of words. Each of these words represent a node in the graph whose polarity is to be determined. They focused on three languages mainly English, French and Hindi but claim that their work can be extended to any other language for which WordNet is available. Much work has been done towards developing the subjective lexicon for English languages but very few work has been done towards development of subjective lexicon for Indian Languages .

2.3 Indian Language Sentiment Analysis

As far as Indian Languages are concerned, we can see small amount of work done in Hindi and Bengali by Amitava Das and Bandopadhya[10], [11], Dipankar Das and Bandopadhya[12], Joshi, Balamurali and Bhattacharyya[24]. Das and Bandopadhya[10] developed sentiwordnet for bengali language. They applied word level lexical-transfer technique to each entry in English SentiWordNet using an English-Bengali Dictionary to obtain a Bengali SentiWordNet. This process resulted in 35,805 Bengali entries. Das and Bandopadhya[11], devised four strategies to predict the sentiment of a word. First approach, an interactive game which in turn annotated the words with their polarity. Second approach, using Bi-Lingual dictionary for English and Indian Languages. Third approach, wordnet expansion using synonym and antonym relations, but their article missed the approach they followed for this expansion. Fourth approach, learning from pre-annotated corpora.

Dipankar Das and Bandopadhya[12], performed the task of emotion tagging for Bengali words. They classified words in Ekmans six emotion classes (anger, disgust, fear, happy, sad and surprise) along with three types of intensities (high, general and low) for sentence level annotation. Joshi, Balamurali and Bhattacharyya[24] created H-SWN (Hindi-SentiWordNet) using two lexical resources namely English SentiWordNet and English-Hindi WordNet Linking [26]. Using WordNet linking they replaced words in English SentiWordNet with equivalent Hindi words to get H-SWN.

Our work is motivated towards Hindi Language and is related to works by Kim and Hovy [31] and Delip Rao and Deepak Ravichandran[36]. Kim and Hovy restricted their assumption to synonyms, we extend the relation to antonyms. Rao and Ravichandran performed bi-polar classification, we extend classification to third level i.e. objectivity. In this work, we use Hindi WordNet [32] to obtain the polarity of adjectives and adverbs for Hindi Subjective Lexicon. The work done in blogs sentiment classification is slightly related and motivated by the work done by Draya, Plantie, Harb, Poncelet, Roche and Trousset[16]. The product reviews classification is motivated by the work done for opinions classification by Bo Pang, Lillian Lee and Vaithyanathan [35], P Turney [40], Amitava Das[11], and Joshi, Balamurali and Bhattacharyya[24].

Chapter 3

Resources

Hindi is a scarce-resource language, not much work has been done in the area of Sentiment Analysis for Hindi language as discussed in Chapter-2. Our focus is to build resources and make it public/open to be used by others. The annotated data sets for different genres, provide an opportunity to design and develop algorithms/approaches to analyze sentiments from the web text. In this chapter, we discuss the resources which are developed as a part of this research work and the already existing resources that are been used in this research.

3.1 Resources Developed

For performing Sentiment Classification for Hindi language we don't have sufficient annotated datasets and other resources. As a part of this research work we developed the following resources-

1. **Reviews Dataset**- We built a dataset of 700 product and movies reviews with equal number of positive and negative reviews i.e the dataset has 350 positive and 350 negative reviews. The reviews are built using pre-annotated 1000 positive and 1000 negative English reviews from amazon dataset [6]. These reviews are carefully selected and are translated using Google translation. These reviews are then manually annotated and corrected to select the the top 350 positive and 350 negative reviews.

2. **Blogs Dataset**- We build a dataset of 250 blogs which were crawled from BBC Hindi. Most of the blogs belong to the domain of politics, sports and news events. Each blog has an average length of 20 sentences which are manually annotated at three class level-positive, negative and objective. The sentences are annotated by three native speakers of Hindi and the polarity of the sentences is decided based on the maximum vote (inter-annotator agreement).

3.1.1 Hindi Product Reviews Dataset[1]

Amazon Reviews has English reviews which are rated on a scale of 1-5. For the Multi-Domain Sentiment Data the authors classified the reviews which have rating of 4-5 as positive and one with the ratings 1-3 as negative reviews [6]. We carefully choose the reviews with length less than 25 to minimize the translation errors. So after pruning the reviews based on the length, we finally selected 1000 positive and 1000 negative review which are then translated using Google Machine Translation. This set of 1000 positive and 1000 negative reviews are manually annotated and edited by 3 people which are native Hindi speakers. The annotation was at the scale of 1 and 0. 1 if sentence exhibits positivity and 0 if sentence exhibits negativity else discard the sentence/ review. Thus for each review we have 3 scores by different annotators. Out of the 1000 reviews each we selected the top 350 by cross-verifying the reviews which were manually annotated by taking the majority vote.

The main issues we observed while building this dataset- As the word order of Hindi is different from English *Refer* Table1.3, so the translation of sentences is not that good in terms of the readability. The mis-match of vocabulary and lack of identifying the context and word sense disambiguation leads to issues at the grammatical aspects. Some Examples-
English Sentence- Clever and stylish design; a lot of colors to choose from; and an excellent performance, best sport shoe.
Translated Hindi Sentence- चालाक और स्टाइलिश .दिजाइन, रंग का एक बहुत से चुनने के लिए, और एक उत्कृष्ट प्रदर्शन सबसे अच्छा खेल के जूते
Correct Hindi Sentence with re-arrangement of words- चालाक और स्टाइलिश .दिजाइन, बहुत से रंग एक चुनने के लिए, और एक उत्कृष्ट प्रदर्शन सबसे अच्छा खेल के जूते

English Setence- Your company was a pleasure to work with- thanks!
Translated Hindi Sentence- आपकी कंपनी को धन्यवाद के साथ काम करने के लिए एक खुशी थी!
Correct Hindi Sentence with re-arrangement of words- आपकी कंपनी के साथ काम करने के लिए एक खुशी थी धन्यवाद!

Table 3.1 summarizes the data (reviews) generated by translation.

Table 3.1 Product Review Data Summary

Total Positive Reviews	1000
Manually Corrected Reviews	350
Total Negative Reviews	1000
Manually Corrected Reviews	350
Total Annotated Reviews	350 + 350

[1]Product Reviews Dataset has been made public and is freely available for research purposes only.

To address the above challenges we faced while forming the Product-Reviews dataset, we made the following assumption - the translation leads to change in the word order so we selected the sentences if even after translation the two conditions hold-
1) Subjective nature is prevalent or restored
2) The readability of the review is maintained

3.1.2 Hindi Blogs Dataset[2]

Hindi Blogs Dataset is created by crawling popular hindi news site BBC Hindi. This dataset contains 250 blogs belonging to sports, politics and news event. The average number of sentences are 20. The blogs have been annotated at sentence level. The details with respect to the overall sentence variations in terms of the subjective nature is as follows-

Source	BBC Hindi
Number of Blogs	250
Number of Sentences	5929
Number of Positive Sentences	733
Number of Negative Sentences	1427
Number of Objective Sentences	3478
Number of Discarded Sentences	291

Table 3.2 Hindi Blog Dataset

For assigning the polarity to the sentence, three classes (1 , 0, -1) have been used.
1 for Positive- If a sentence in a blog gives an indication of positivity.
-1 for Negative- If a sentence in a blog gives an indication of negativity.
0 for Objective- If a sentence does not show any subjective behavior.

The dataset is annotated by 3 annotators who are native Hindi speakers thus for each sentence we have three annotations. The final polarity is decided by the majority basis. If for a sentence all the three votes differ than we discard those sentences. The humans can also make errors in annotation, the perception and interpretation of one individual towards a particular text might be different with respect to others. So we calculated the kappa statistics which gives an idea about the inter-annotator agreement as shown in Table3.3.

[2]Blogs Dataset has been made public and is freely available for research purposes only.

Kappa (κ) score between annotator 'i' and annotator 'j' (κ_{ij})	
κ_{12}	**0.775**
κ_{13}	0.738
κ_{23}	0.743
Average κ Score	**0.752**

Table 3.3 Kappa Scores for Manual Agreement for Hindi Blog Dataset

3.2 Resources Used

3.2.1 Hindi Wordnet

This resource is one of the valuable and most used resource of the Hindi Language. Hindi Wordnet was developed in 2002 [32], it provides the words corresponding to different parts of speech information like adjective, adverb, noun and verbs linked with other words through syntactic and semantic relations. It provides information regarding how words are related to each other, the words belonging to similar sysnsets, it also provides information about the synonyms, antonyms etc for a given word.

Example - For word **अच्छा**

Synonyms are **बढिया, भला** etc

Antonyms are **बुरा, खराब, अनुचित** etc

similarly we have other semantic and syntactic relations between the words provided by Hindi Wordnet .

3.2.2 Hindi SWN

Hindi SentiWordnet is a subjective lexicon developed by IIT bombay [24]. It contains words with part of speech as Adjective, Adverb, Noun and Verb with each word having 3 scores positive, negative and objective. Sum of positive, negative and objective score sums to 1. The lexicon assigns single score to a word irrespective of the sense in which it is used.

This resource was built using two lexical resources namely English SentiWordNet [17], [3] and English-Hindi WordNet Linking [32]. It contains 4 categories of synsets based on the Part of Speech information which are Noun, Verbs, Adverbs and Adjectives.

3.2.3 IIT B- Movies Reviews

This dataset [24] has about 273 movies reviews which are manually annotated, the dataset has equal number of 125 positive and negative documents where each document has 1-2 reviews.

3.2.4 English Wordnet

Wordnet [29] is a large lexical database of English. It groups words together based on their meanings. It also provides various semantic relations between words. Broadly their are 4 POS categories Adjectives, Adverbs, Nouns and Verbs and within each category there is a semantic network of words.

3.2.5 English SentiWordnet

SentiWordnet [17], [3] is a Wordnet based resource which is automatically generated with polarity scores attached to the senses instead of the terms.
Example- For word **"healthy"** the polarity scores are as follows-
Positive=0.75, Negative=0.0, Objective=0.25 **#sense-1** (having or indicating good health in body or mind; free from infirmity or disease; "a rosy healthy baby"; "staying fit and healthy")
Positive=0.5, Negative=0.0, Objective=0.5 **#sense-2** (financially secure and functioning well; "a healthy economy")
Positive=0.0, Negative=0.0, Objective=1 **#sense-3** (large in amount or extent or degree; "it cost a considerable amount")

3.2.6 English Subjectivity Lexicon

Subjectivity lexicon [43] contains 8221 words where each word is assigned 4 polarities (positive, negative, both and neutral) along with 4 level gradation.

3.3 Summary

In this chapter, we discuss the resources that have been developed as a part of this research work and the resources that have been used in this research work. We discuss the mechanisms, assumptions and other details which are made while making the datasets. These datasets are manually annotated using language experts and the inter-annotator agreement is measured using standard techniques like kappa statistics.

Chapter 4

Subjective Lexicon Generation

Subjective Lexicon has been widely used for analyzing the sentiments in a web text. In this chapter, we discuss in detail different algorithms for developing Subjective lexicon that have been used in the past. We present our Graph based approach to generate Subjective Lexicon using just the WordNet and pre-annotated seed list for a language. We present the lexicon developed using our technique for English and Hindi language. The lexicon generated is evaluated using different metrics- comparing against similar existing resources, manual annotation and classification perfomance of same.

4.1 Background

Subjective Lexicon is a list of words where for each word we have the scores indicating the positive, negative and objective nature of the word. The main three approaches for building the subjective lexicon are as follows-

4.1.1 Using Bi-Lingual Dictionaries

This approach deals with mapping the polarity of the words using the bi-lingual dictionaries . For this method we require the linking or mapping between the words of the two languages and the subjective lexicon or a dictionary of polarity of words for one language out of the two.

4.1.2 Using Machine Translation

In this approach, we use the translation technology to convert the subjective lexicon (a list of polarity of words) from source to the target language. Translation tools such as Google Translation, Language specific translation systems such as Sampark System etc are used for translating subjective lexicon of one language to another.

4.1.3 Using Wordnet

This method uses the inherent structure of the language embedded in the form of the wordnet of that language. It outputs the lexicon capturing the relations and similarity between the words of the wordnet.

In our opinion, there are a few limitations of Using the Bi-Lingual Dictionaries and Using the Machine Translation approaches-

1. Bi-Lingual dictionaries may not account for all the words because of language variations. Words can be used in multiple context in either languages and context dependent word mapping is a tough task, error prone and require manual efforts. Table 4.1 gives an idea when a word in English can correspond to more than one words in Hindi varying in the nature as well as in the semantics and meaning.

Black	Close	Cold	Cool
काला	पास	ठंडा	शीतल
श्याम	नज्दीकी	निरुत्साही	शांत
अंधेरा	समान	हल्का	उदासीन
भयंकर	सूक्ष्म	कठिन	मिलनसार
उदास	बराबर	बेहोश	अच्छे

Table 4.1 English-Hindi Sense Based one to many mapping

2. Using Translation method (from rich resource to scarce resource language) for generating subjective lexicon, there is a high possibility of loosing the context information and have many translation errors. One of the common problem in the translation is the mapping of a word in one language to a phrase in the other language as discussed in Table4.2.

English Words	Hindi Translation
sage-green	उक्त वनस्पति की पत्तियों का रंग
uniformed	एक समान बनाया गया
indecisive	दुविधा में पडा हुआ
well-bound	अच्छी तरह से बाध्य
coaxing	मीठी बातें से मनाना
sad	दुख की बात है
weatherly	वायु के समीप रहने में समर्थ
well-groomed	अच्छी तरह पाला - पोसा हुआ

Table 4.2 Word to Phrase translation

Whereas wordnet and its inherent relations provides much richer information to be mined while deriving a subjective lexicon for a particular language.

Words	Positive Polarity	Negative Polarity	Objective Polarity
नरम	0.55	0.45	0.0
ठण्डा	0.263	0.105	0.632
नाजुक	0.115	0.077	0.808
पतला	0.428	0.572	0.0
मुलायम	0.5	0.375	0.125
कच्चा	0.035	0.07	0.895
चतुर	0.357	0.428	0.215

Table 4.3 Ambiguous Words with scores generated by HSL

4.2 Our Algorithm - Using Wordnet for building Subjective Lexicon

In this section we present our graph based method to build a subjective lexicon for a given language using Wordnet as a resource.

Our algorithm makes a hypothesis of traversing WordNet like a graph where every word in WordNet is imagined as a node in the graph. This graph is an undirected graph and is highly connected but not fully connected. In this graph, nodes are connected to each other based on synonym and antonym relations. Kim and Hovy [31] made a simple assumption that synonyms carry the same sentiment as the word. We extend their assumption to antonyms, we assume antonyms carry opposite polarity. Each node is connected to many other nodes i.e. each node will have many in-links and many out-links. This graph has three basic connected components (positive, negative and objective). Words can be connected using two kinds of edges:
Simple Edge: An edge which connects two words in the same (different) domain and represents a synonym (antonym) relation with a given condition that each word should belong to non overlapping region.
Cross Edge: An edge which connects two words based on synonym (antonym) relation and atleast one among these words lies in the overlapping region. Cross Edges among these connected components produce words which have ambiguous (dual) nature *Refer* Table 4.3.

Figure 4.1 explains Simple Edges and Cross Edges pictorially. In figure 4.1, each circle represents a connected component and an overlapping zone that contains words which are ambiguous by nature.

We use a WordNet and a list of words (seed list) which is pre-annotated based on the polarity. Our seed list contains words belonging to all the three categories positive, negative and objective. Each word in this seed list is expanded on synonym and antonym relations. We can consider this expansion as a Breadth First Traversal of a graph. In this traversal method, all the siblings (nodes at any depth d) are expanded before any node at the next level (depth d+1) is expanded. We make use of queue data structure to maintain the order in which the nodes (words) are introduced or

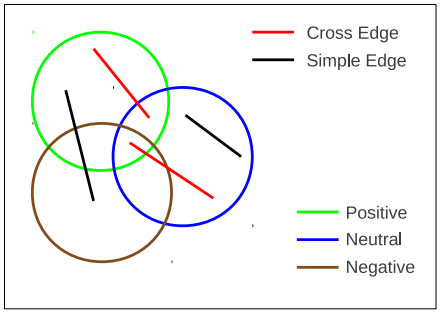

Figure 4.1 Cross Edge and Simple Edge Diagram

expanded. This method helps us to ensure that each node is expanded only once and all the synonyms of a word are traversed at the same time.

In our method we have 2 lists, one is temporary and the other is final list. The initial seed list which contains seed words is copied to temporary seed list with the polarity. Now every time we get a word (a structure which contains a pair of seed and polarity) from the temporary seed list by de-queuing it from the list, we check for this word if it exists in the final seed list or not. If this word is in the final seed list then we don't populate this word further, we just add the current polarity of this word to the polarity in the final list. But if this word is not in the final list, we do three things-

1. Add this word to the final list with the current polarity.
2. Find out all the synonyms of this word and en-queue them in the temporary seed list with the polarity same as the source word.
3. Find out all the antonyms of this word and en-queue them in temporary seed list with opposite polarity. (P ->N, O ->O, N ->P).

We continue this process till all the words in the temporary seed list are explored or in other words till the temporary seed list becomes empty. When the temporary seed list becomes empty the final seed list contains adjectives and against each adjective we have string of P's, N's and O's. Based on this we decide the final polarity of the word. Say for a word 'x' in the final seed list we have string 's' made of P's, N's and O's.
Length of string (s) = Len
Number of P's in s = nP
Number of N's in s = nN
Number of O's in s = nO

Positive polarity of x = nP/Len
Negative polarity of x = nN/Len
Objective polarity of x = nO/len

example for word **X** the polarity string we get is *"PPNNNPPOOONP"*

Len=12, nP=5, nN=4 and nO=3

hence the polarity of the word **X** is *(5/12,4/12,3/12)*

Word **X**- *Polarity* Pos:0.416 Neg:0.333 Obj:0.25

For Pseudo Code Refer Algorithm 1.

4.3 Evaluation Strategies

One of the major task while proposing a new method is the evaluation. In these kind of systems we mainly evaluate by human judgment or by classifying some pre-annotated text. There are few methods which are commonly used for validation-

4.3.1 Human Judgment / Manual Annotation

This method is usually opted for scarce resource languages. In this method, some manual annotators are appointed whose task is to annotate the lexicon generated and later, taking the majority vote of annotators the system generated lexicon is validated.

4.3.2 Validating Against Existing Resources

In this method of evaluation, we find the accordance of lexicon generated with a lexicon which is already proposed and accepted by the research community. This strategy of evaluation is used for languages which are resource rich.

4.3.3 Classification Performance

In this method of evaluation, we classify pre-annotated reviews/blogs using our system generated lexicon and find precision, recall, F1 scores, etc to show the correctness of the lexicon. This strategy is generally used for resource rich languages or for those languages for which we have pre-annotated data.

Algorithm 1 Algorithm for Populating SeedList using WordNet to generate Subjective Lexicon

$InitialSeedList = \{45\ words\ (15\ Objective,\ 15\ positive,\ 15\ negative)\}$
// *Each word is a structure which contains a pair of Seed and Polarity*
$FinalSeedList = \{\}$
$TempSeedList = \{\}$
$TempSeedList = InitialSeedList$
while $TempSeedList \neq EmptyList$ **do**
  $Word = TempSeedList.pop()$
  // *gives the first word in the list and removes it from the list.*
  $Seed = Word[0]$
  $Polarity = Word[1]$
  if $Seed \in$ FinalSeedList **then**
    $FinalSeedList[Seed] = FinalSeedList[Seed] + Polarity$
  else
    FinalSeedList[Seed]=Polarity
    SynonymSet = All the synonyms of Seed
    AntonymSet = All the antonyms of Seed
    for all $synonyms \in$SynonymSet **do**
      $TempSeedList.append(synonym : Polarity)$
      // *Polarity will be P/N/O*
    end for
    for all $antonyms \in$AntonymSet **do**
      $TempSeedList.append(antonym : OppPolarity)$
      // *OppPolarity will be P if Seed has Polarity N*
      // *OppPolarity will be N if Seed has Polarity P*
      // *OppPolarity will be O if Seed has Polarity O*
    end for
  end if
end while
// *Against Each adjective in the FinalSeedList we have a string of* $P's,\ N's$
// *and* $O's$ *which contains the polarity of that word*
for all $adjectives \in$FinalSeedList **do**
  $S = FinalSeedList[i]$ // *Here i is an adjective and S is the string of*
  // *polarity for that adjective*
  $nP = Number\ of\ P's\ in\ S$
  $nN = Number\ of\ N's\ in\ S$
  $nO = Number\ of\ O's\ in\ S$
  $Len = length\ of\ S$ // *Please Note* $nP + nN + nO = Len$
  $PositivePolarity = nP/Len$
  $NegativePolarity = nN/Len$
  $Objectivity = nO/Len$
  //*Please Note* : $PositivePolarity + NegativePolarity + Objectivity = 1$
end for

4.4 Experiments

The algorithm proposed in section-4.2 has been tried and validated for English as well as Hindi Language.

4.4.1 Building Lexicon for English Language

For English language we use English Wordnet [29] and took a list of 45 seed words 15 positive, 15 negative and 15 objective words and using our algorithm (Section 4.2) we traversed the English Wordnet as discussed above Algorithm 1. The algorithm resulted in 9143 adjectives the details are described in Table 4.4.

Positive Adjectives	3067
Negative Adjectives	2966
Neutral Adjectives	3110
Total Adjectives	9143

Table 4.4 English Subjective Lexicon (ESL)

There is a trade off between the size of the seed list and the scope (coverage) of the generated lexicon. The bigger and better the seed list, the better is the generated lexicon. But to have bigger and better seed list, we need more manual exercise to decide the seed words and their polarity. So, here we try to make a trade off between seed list size and generated lexicon size. Table 4.5 provides results for our lexicon with varying seed list sizes. We observed that if the seed list size is too small than the words retrieved are few and increasing the seed list to a large number is costly and requires human efforts. With few seedlist present instead of the complete graph traversal the partial traversal results in varying and false polarities. Thus on an average the seed list of 45 words performed well.

Seed List (size)	**Lexicon (size)**	**Accuracy (%) with SentiWordNet**
2 (Good and Bad)	3564	68.90
15 (5 + 5 + 5)	4765	72.67
30 (10 + 10 + 10)	5378	74.89
45 (15 + 15 + 15)	9143	79.89

Table 4.5 Results for Agreement with SentiWordNet with increasing seed list size

Number of words in SentiWordNet	117659
Number of adjectives in SentiWordNet	21137
Number of words in our Lexicon	9143
Common words in both the lexicons	9143
Words in agreement	7297
Agreement in common words	**79.8%**

Table 4.6 Results for Agreement with SentiWordNet

4.4.1.1 Evaluation

The English Lexicon generated is validated against the English Sentiwordnet [17], [3] and Subjectivity Lexicon [43] which are some of the commonly used subjective lexicons for English Language.

Against SentiWordNet: Based on the polarity scores we assigned each word with a single polarity (the polarity with highest score) for both the lexicons (the one generated by our proposed algorithm and the lexicon of SentiWordNet). Our system generated lexicon achieved an agreement of 79.8% with SentiWordNet. Detailed results are presented in Table 4.6. Most of the cases we missed were due to dual nature of words i.e words showing positive, negative and objective polarity simultaneously eg small, long, short etc. Such words are highlighted in Figure 4.2

Against Subjectivity Lexicon: Subjectivity Lexicon contains 8221 words in total. Each word is assigned one of the four polarities (positive, negative, both, neutral). It's build using the polarity cues that are mentioned in [37]. Considering both polarity and neutral polarity as same, we evaluated our lexicon. Details and results are summarized in the Table 4.7

4.4.1.2 Discussion

The English lexicon generated using our algorithm performed quite well and showed an agreement of about 80% with SentiWordnet and about 81.5% with Subjectivity Lexicon. The coverage in terms of the number of adjectives is also quite good when expansion is performed over the English Wordnet. The main disadvantage is that our lexicon is terms based and not sense based. But the accordance of our algorithm with the other English lexicon indicates that the Wordnet is a well connected network of words and can be used to build subjective lexicon. Main issue and

Number of words in Subjectivity Lexicon	8221
Number of adjectives in Subjectivity Lexicon	3249
Number of words with pos as 'anypos'	1147
Number of words in our Lexicon	9143
Common words in both the lexicons	4210
Words in agreement	3430
Agreement in common words	**81.47%**

Table 4.7 Results for Agreement with Subjectivity Lexicon

error comes while dealing with context sensitive words like long, easy, loose the words lying in the ambiguous area as indicated in Figure 4.2.

4.4.2 Hindi Subjective Lexicon (HSL)

For Hindi language we use Hindi Wordnet [32] and a seed list of 45 adjectival words. Table 4.8 provides a small sample of the seed list we are using for the graph traversal of the wordnet. The lexicon built using the above mentioned approach Algorithm 1 for Hindi Language contains 8048 adjectival words in all. The details with respect to the distribution of the words is described in Table 4.9. For each word, our lexicon provides three scores: positive, negative and objective. Sum of these scores equals to 1.

Example word **"नरम"** in HSL has a score of (0.55 , 0.45, 0.0)
Positive score + Negative Score + Objective Score= 0.55 + 0.45 + 0.0 =1

Seed Words	**Polarity**
अच्छा, वीर, लाजवाब, बुद्धिमान, उत्कृष्ट	Positive
बुरा, स्वार्थी, कुण्ठित, प्रचंड, क्रोधी	Negative
वृद्ध, जवान, विदेशी, शहरी, ग्रामीण	Objective

Table 4.8 Sample of a seedlist

Positive Adjectives	2521
Negative Adjectives	3060
Neutral Adjectives	2467
Total Adjectives	8048

Table 4.9 Hindi Subjective Lexicon (HSL)

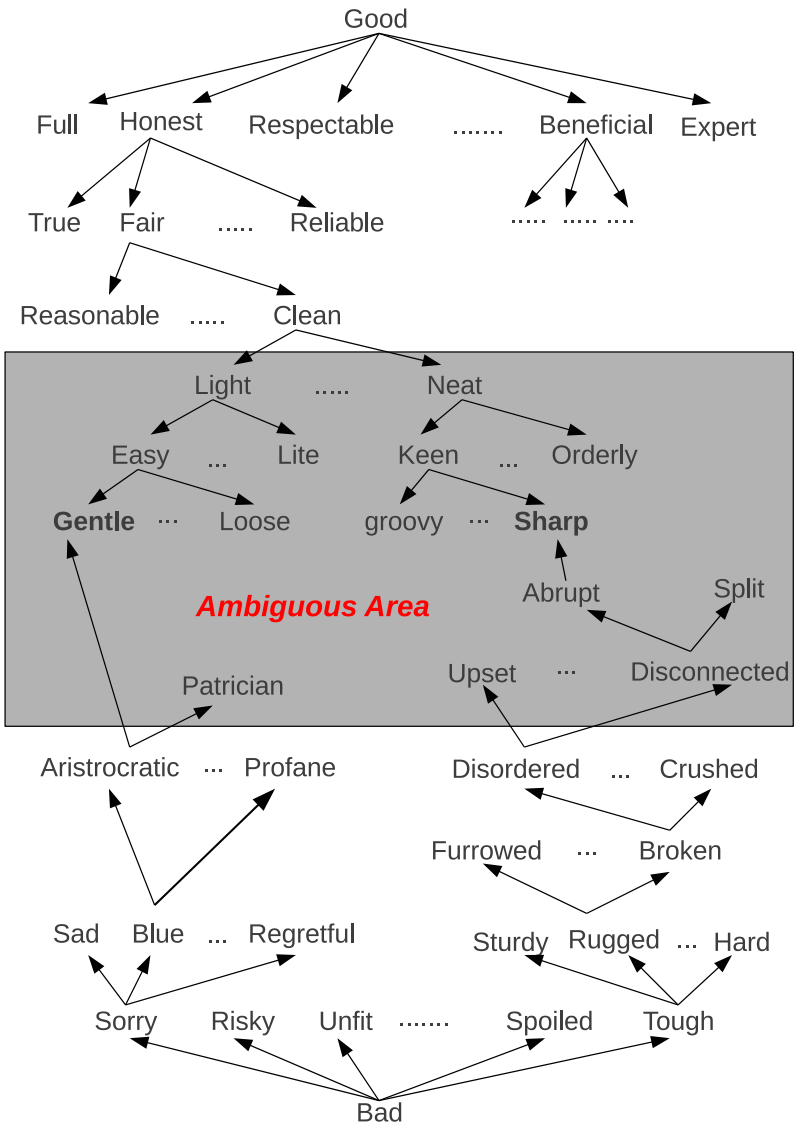

Figure 4.2 English Graph Traversal

Agreement of each annotator with our lexicon	
Annotator 1	66.08%
Annotator 2	64.01%
Annotator 3	**68.45%**
Overall Agreement of our lexicon with the annotators	**68.01%**

Table 4.10 Results for Manual Agreement with Hindi Lexicon

Kappa (κ) score between annotator 'i' and annotator 'j' (κ_{ij})	
κ_{12}	0.775
κ_{13}	0.727
κ_{23}	0.742
Average κ Score	**0.748**

Table 4.11 Kappa Scores for Manual Agreement for Hindi Lexicon

4.4.2.1 Evaluation

We validate this Hindi Subjective Lexicon by all the three different methods of evaluation discussed earlier-

Human Judgment- In this method of evaluation, we hired three manual annotators who are language experts in Hindi. We asked each annotator to tag the words generated by our system on the scale of 3 (negative:-1, neutral:0, positive:1). After getting the list annotated by all the annotators, we had three votes for each word and we took the majority call. Table 4.10 reports accordance of Hindi lexicon generated using our system with manual annotation. We calculated the Kappa (κ) Score to find a measure of inter-annotator human agreement the results are shown in Table 4.11

Reason behind low mutual agreement among the annotators is that many words in Hindi show ambiguous nature as shown in Table 4.3. The polarity of these words depends on the sense in which they are used. This ambiguous nature is highlighted in Figure 4.3.

Classification- For this evaluation strategy, we performed classification of product reviews dataset described in Chapter 3. On this data, we performed unigram presence and simple scoring method classification.
In unigram presence method, for a given sentence/review we count adjectives of positive, negative and objective polarity. The polarity with maximum count is assigned to the sentence/review.
In simple scoring method, for a given sentence/review we summed the positive, negative and objective scores of each adjective. The polarity with dominant score is assigned to the sen-

Table 4.12 Results for Product Review Classification using Lexicon generated by our approach

Method	Accuracy
Adjective Presence	
Baseline	65.50
Baseline + Negation Handling	68.67
Baseline + Stem	67.17
Baseline + Stem + Negation Handling	70.80
Adjective Scoring	
Baseline	67.33
Baseline + Negation Handling	70.00
Baseline + Stem	71.00
Baseline + Stem + Negation Handling	**74.10**

tence/review.

From every review we identified the adjectives[4] and scored those adjectives using our lexicon. If an adjective is missing from our lexicon we considered the stemmed variant of same word. In addition to stemming we also performed negation handling. We identified the words with tag "NEG" using sliding window of 6 words and swapped the polarity (positive and negative) of adjectives in this range. Our sliding window, looked upto 3 words in both the directions (left and right) of this word. Table 4.12 reports the results of classification.

We also applied this method of subjective lexicon generation for Hindi Adverbs, we took a seed list of 75 words, 25 each positive, negative and objective. After expanding based on the algorithm stated above Algorithm 1 we got a list of 888 adverbs with each word been assigned with 3 scores positive, negative and objective.

4.4.3 Other proposed algorithms for generating Hindi polarity lexicon

4.4.3.1 Lexicon Generation using English-Hindi Linkages

Joshi et al.[24] proposed a method to generate a subjective lexicon for Hindi. Their method depends on two lexical resources: English SentiWordNet and English-Hindi WordNet Linking [26], the algorithm proposed in their research finds polarity of each Hindi word using SentiWordNet and English-Hindi linkage. The Hindi Sentiwordnet developed in [24] has been discussed above in Chapter-3.

Using this mentioned approach, we generated the subjective lexicon using a English-Hindi dictionary Shandanjali [5] and used the English Sentiwordnet for assigning scores to the Hindi words

[4]The adjectives were identified using Part of Speech Tagger

[5]http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html

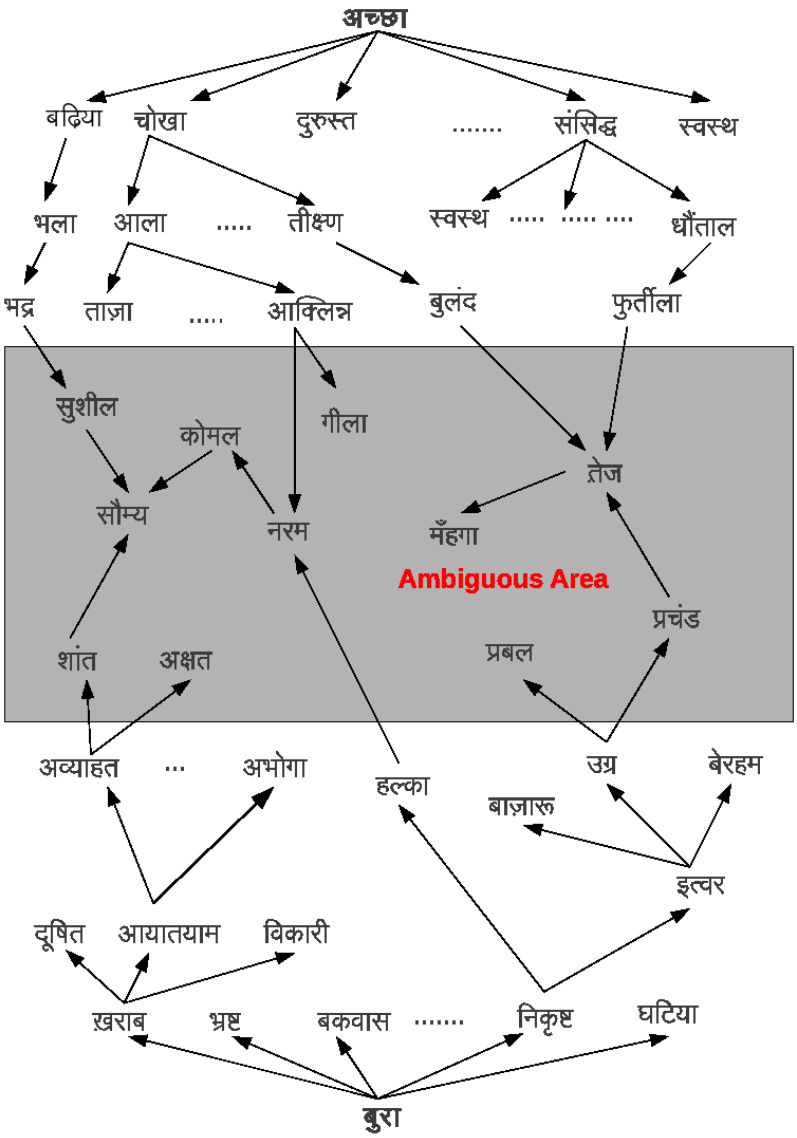

Figure 4.3 Hindi Graph Traversal

mapped from English to Hindi using the Shabdanjali dictionary[5]. This lexicon contains 4335 adjectives and 1279 adverbs. One observation we made made while generating this lexicon is that most of the adverbs are derived from adjective and nouns like luck->luckily, happy->happily, pretty->prettily etc. Table 4.13 details number of adjectives and adverbs across different lexicons. Our lexicon generated using bilingual dictionary differs from the lexicon generated by [24] as they used English-Hindi wordnet linking and the bilingual dictionary used by us is the Shabdanjali dictionary which is a synset based manually generated dictionary.

4.4.3.2 Lexicon Generation using Machine Translation

In this method we translated the English Sentiwordnet to Hindi using google translation and the translated words in Hindi are assigned the polarity as that of the source word in English Sentiwordnet. Because of translation errors and the main issue of phrasal translation many of the words are been wrongly translated and hence the lexicon generated is far poor, some of the examples are mentioned in the Table 4.2.

Type	Our Lexicon (HSL)	Bi-Ling Dict	HSWN	Translated Dict
Approach Used	Wordnet Transversal	Using Shandanjali	Wordnet Linking	Google Translation
Adjectives	8048	4335	4861	12086
Adverbs	888	1279	294	1509

Table 4.13 Distribution of Words across different Lexicons

4.5 Discussion HSL vs HSWN vs Bi-Ling Dict vs Translated Lexicon

The Hindi wordnet has about 15k adjectives and our algorithm is able to traverse and capture about 8k common words which is quite good in terms of the coverage. The English-Hindi linking used from shabdanjali [5] has about 5656 unique words in English which are mapped to 6670 adjectives in Hindi where one word in English can be mapped to many words in Hindi as per sense based mapping as shown in Table 4.1. The lexicon generated using shandanjali English-Hindi word mapping performs better in comparison to the lexicon generated using English-Hindi wordnet linkings, in terms of both the coverage and diversity of words. The English Sentiwordnet has about 21k adjectives but the translation results in about 12k adjectives . Most of the adjectives have noise or phrase translation hence the coverage as well as the performance is quite poor in the lexicon generated directly using translation.

In conclusion, the lexicon built using the wordnet approach is better as compared to the other two in terms of the coverage i.e the number of lexical items (adjectives and adverbs) and also performs

fairly well for the classification problem as shown in Table 4.12.

4.6 Applications of Hindi Subjective Lexicon

The Hindi Subjective Lexicon provides the polarity information of about 8048 adjectives and 888 Hindi adverbs. The lexicon generated can be widely used as follows-
1) **Business Purposes**- The lexicon can be used to predict the subjectivity of a text and thus can help in making business decisions , capturing the user's perspectives and likelihood. It can be used for research purposes for analyzing content coming from web belonging to different domains and genres as blogs, news, forums etc. The lexicon provides an idea about the polarity expressed in the text and thus can aid in performing sentiment analysis.

2) **Research Purposes**- This lexicon can also be used to study the pattern and similarities measure between the text coming from similar places and person to have a periodic analysis of same. This lexicon can also be used to compare with subjective lexicon of other languages and deduce interesting language phenomenon between them.

4.7 Limitations of the Subjective Lexicon

One of the major limitation of this lexicon is that the current version of this algorithm does not perform Word Sense Disambiguation. This is so because WordNet's available (for scarce resource languages) doesn't provide the information of most used senses for a word and various senses in which it can be used. Hindi WordNet (specifically) provides information about synonyms and antonyms but lacks the information on most commonly used senses. Thus the lexicon that we have generated is not sense based but terms based so it gives a rough idea about the term polarity as the exact polarity comes with the sense information which is missing at present. Scope of the system proposed above is dependent on the initial seed list used to populate the WordNet. If we choose the seed list in a careful manner with the help of linguistic experts the results and scope of the Lexicon thus generated would be better.

4.8 Hindi Subjective Lexicon-HSL

The lexicon generated using our proposed Algorithm 1 described in Section-4.2, has about 8048 adjectives and 888 adverbs. The lexicon generated using this algorithm performs better and also contain more entries. For adjectives we used 45 seed words and expanded it using the Hindi

Wordnet which has about 15k adjectives which resulted in about 8048 adjectives. For adverbs we used 75 adverbs which resulted in 888 adverbs only, this statistic indicates the nature in terms of the connectivity between the words in the wordnet for adjectives and adverbs. The network/ connectivity of words in the Hindi Wordnet is better for Adjective part of speech as compared to Adverb part of speech. In Hindi Wordnet, the adjectives are well connected and exhibit strong connection chains whereas in the case of adverbs they form discrete unconnected clusters. The inter-annotator agreement for adjectives is about 68% and the average kappa score is about 75% the low score is because of the words falling in the ambiguous areas Figure 4.3. For these ambiguous words Table 4.3 it becomes hard to assign the subjective nature as the nature of the words depends on the context.

Some of the examples where the polarity assigned by our algorithm differ from the human annotation is described in Table 4.14

Word	Polarity Given by Annotator	Polarity Given by our System
जखमी	Objective	Negative
पिछला	Objective	Negative
शीतल	Positive	Objective
कम	Objective	Negative
नम	Objective	Positive
तीखा	Negative	Objective
मस्त मौला	Positive	Negative
मार्मिक	Positive	Objective
बेफिक्र	Negative	Positive

Table 4.14 Errors in HSL in comparison to the Manual Annotation

Table 4.14 itself indicates how complex is the task of identifying the polarity of a word irrespective of the context. The nature of the word changes to a great aspect in terms of the writer/reader/speaker/listener involved and the situation/circumstances in which the word is expressed/uttered. Some examples to discuss the need of incorporating the context information-

Example-1

1. **सेब नरम हैं**
2. **राम बहुत नरम हैं**

In the first example the word **"नरम "** indicates the positive aspect that the apples are soft whereas in the second it represent the negative aspect that the person is very soft, delicate.

Example-2

1. केला सख्त हैं
2. मास्टरजी बहुत सख्त हैं

In the first example the word **"सख्त"** indicates the positive aspect that the banana is hard whereas in the second it represent the negative aspect that the teacher is strict that too with an intensifier **"बहुत"**.

So the examples discussed above clearly indicates that it becomes important to understand the context and score the sentences accordingly and hence perform Sentiment Analysis in a selective manner.

The lexicon generated by our algorithm performs reasonably well on the classification problem as indicated in Table 4.12. The main issue while classification comes in handling the morph variations and the spelling variations.

4.9 Summary

In this chapter, we presented three different approaches for building subjective lexicon 1) Using Bilingual Dictionaries, 2) Using Machine Translation and 3) Using Wordnet. We discuss in detail the limitations, issues with each one of them. We proposed an algorithm for traversing wordnet like a graph to generate the subjective lexicon, the algorithm has been tried and experimented for English and Hindi language. The method can be used for generating Subjective Lexicon for any language using the wordnet and a seedlist of that language. The lexicon generated using our algorithm for English language has been compared against existing lexicons such as SentiWordnet and English Subjective Lexicon. The Hindi Subjective Lexicon has been evaluated using three different metrics 1) Comparison with manual annotation, 2) Comparison with lexicons developed using other approaches (using translation technology and using bi-lingual dictionaries) and 3) Classification Performance. The Hindi Subjective lexicon showed significant improvement in terms of the coverage and performance as compared to the other lexicons. The Hindi Subjective Lexicon generated at present, is a term/word based list and not a sense based mapping.

The lexicon generated is one of the first of its own kind in terms of exploiting and harnessing the wordnet knowledge and can be extended in the future to cover the other part of speech information. The approach used can also be explored for other Indian Languages. In future we can come up with sense based scores which will help to ease the task of Word Sense Disambiguation (WSD) and better words selection and hence will aid in the problem of sentiment analysis.

Chapter 5

Sentiment Classification For User-Generated Content in Hindi

User-generated content has been increasing tremendously, different forms of web content such as reviews, blogs, news provide large amount of data to be mined and analyzed. This chapter focuses on analyzing sentiments from two different genres of user-generated content 1)Reviews and 2)Blogs. In this research we focus on Hindi Products and Movie Reviews dataset and on Hindi Blogs dataset. We tried different approaches such as using Subjective lexicon discussed in Chapter-4 and N-Gram approaches to mine and analyze the sentiments from a given text.

5.1 Hindi Product and Movie Reviews Analysis

We performed sentiment classification for Hindi reviews. The review dataset that has been used for this task is the products reviews dataset developed as a part of this research work and the movie documents dataset [24]. The size of the datasets we used for reviews classifications is as follows-

1. Product Reviews Dataset- It has about 350 positive and 350 negative reviews, it has a total of 700 reviews as discussed above in Chapter 3.
2. Movies Review Dataset- It has about 125 positive and 125 negative documents where each document has 1-2 reviews, it has a total of about 273 reviews as discussed in Chapter 3.

The main challenges while analyzing reviews are handling unstructured language, noise and spelling mistakes but as both of the datasets we used are already manually annotated and verified so there is less noise and spelling mistakes. The main issue we faced while handling Hindi reviews is the morphological variations *Refer* Table 1.4 and multiple spellings *Refer* Table 1.5 for a given word .

5.1.1 Framework Tools

This section provides a brief detail of the tools and machine learning algorithms used in the experiments.

5.1.1.1 Weka

We used weka toolkit for performing classification. Weka[1] is an open source toolkit developed by Machine Learning Group at University of Waikato it provides many algorithms for classification, clustering, data mining and also support other functionalities for performing text analytics.

5.1.1.2 Naive Bayes

Naive Bayes Classifier uses Bayes Theorem, which finds the probability of an event given the probability of another event that has already occurred. We used the already implemented Naive Bayes implementation in Weka toolkit.

5.1.1.3 Support Vector Machine (SVM)

This classifier constructs N-dimensional hyper-plane which separates data into two categories. SVM takes the input data and for each input it predicts the class. We used libSVM2 classifier which is available as an add on to the Weka toolkit.

5.1.2 Approaches

To address the problem of sentiment classification of Hindi reviews we followed 3 approaches-

5.1.2.1 Using Subjectivity Lexicon

In this approach, we used the subjectivity lexicon developed in Chapter-4 to perform sentiment analysis. In this set of experiments, we match all the words in a sentence with the adjectives and adverbs lexicon whereas in the earlier experiments discussed in Chapter-4.4.2.1, we were matching only the adjectives and adverbs marked by POS tagger in a sentence. As the performance of POS tagger isn't that good, so to minimize the error caused by POS tagger we are matching all the words in a sentence/review.

On the reviews data, we performed unigram presence and simple scoring method classification.

[1]http://www.cs.waikato.ac.nz/ml/weka/

In unigram presence method, we count unigrams of positive, negative and objective polarity in a sentence/review and assigned the polarity for which the count is highest.
In a simple scoring method, we summed the positive, negative and objective scores of each word in a sentence/review and assigned the polarity of the dominant score.
Stemming[3]- If a word was missing from our lexicon we considered the stemmed variant of that word for scoring. Table 5.1 provides some examples with respect to the word and their stemmed forms.

Table 5.1 Words and their stemmed (root) words

Word(s)	Stemmed Word	Morph Output (root word)
छोटे	छोट	छोटा
अच्छी	अच्छ	अच्छी
अच्छे	अच्छ	अच्छा
बड़ी	बड़	बड़ी
हल्के	हल्क	हल्का
लंबे	लंब	लंबा

As such there are no standard stemmers available for Hindi language and stemming itself creates some issues and makes errors so we performed morphological analysis of the word to identify the root word and assign the polarity of the root word to the input word. Table 5.1 provides some examples of the mophological analysis. We used the Hindi morphanalyzer developed at LTRC, IIIT-Hyderabad [4].

In addition to stemming and mophological analysis we also performed negation handling.
Negation Handling - Certain words like **" नहीं ,न "** reverse the polarity/sense of the sentence so they need to be tackled carefully.
Example **"सेब अच्छा नहीं हैं"** in this sentence without incorporating the presence of word "**नहीं**" the review will be classified wrongly as positive.

We identified the words with tag NEG (marked by Hindi Shallow Parser [2]) and swapped the polarity (positive and negative) of adjectives (adverbs) in the sliding window of 6 words. Our sliding window, looked upto 3 words in both the directions (left and right) of this word. Apart from the lexicon developed in this research we also approached towards the reviews classification problem using the Hindi lexicon generated by the other two methods-
1) *Bi-Lingual dictionary*
2) *Translated Dictionary*

[3]We have used the rule-based stemmer for our experiments
[4]http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallowparser.php

We classified the product dataset which has 700 reviews and movies dataset which has about 273 reviews using the lexicons (HSL, HSWN, Bi-Lingual dict, Translated lexicon) discussed in Chapter-4. We divided our experiments into 2 phases-
1)Only Adjectives- In this method, we score the reviews/sentences using just the adjective lexicon.
2)Adjectives+Adverbs- In this method we score the reviews/sentences using both the adjective and the adverb lexicon .
Table 5.2 shows the classification accuracy using different lexicons. In the table 5.2 NH indicates Negation Handling, Stem indicates using Stemmer and Morph indicates using Morph Analyzer.

5.1.2.2 Using N-Gram Modeling

In our experiments we performed N-Gram modeling in two ways-
1) *Traditional N-Gram Modeling (Word-Word feature vector formulation)*- In this method we do the traditional bag of word formulation where we formed the feature vector for the unique N-Grams of the entire reviews dataset. For each training review we form a feature vector comprising of 1's and 0's where 1 and 0 indicates that a particular feature is present or absent. Thus the training model consists of samples with values 1's and 0's and the class to which they belong. Given a new testing sample we form the feature vector and try to learn the class of the sample using the model formed by the training set.

We experimented with different N-grams variation and their combination (unigram, bigram , unigram + bigram and unigram + bigram + trigram). The reviews dataset has just 700 reviews which is a small dataset so in another set of experiments we try to prune the features on term-frequency basis. The results are discussed in Table 5.3 and 5.4 .

2) *Weighing the N-Gram based on relevance*- In this method we divided the pre-annotated data into two parts namely training set and testing set, we performed 3-fold and 5-fold cross-validation. After dividing the data we form trigrams, bigrams and unigrams on the training data and store them in an individual n-gram dictionary. We create two separate models each for positive and negative polarity. For testing every review we create trigrams in the similar manner. Then we check if this trigram exists in our positive and negative trigram dictionary. If it exist then, we increase the count of trigram matched else we break this trigram into two bigrams. These bigrams thus formed are cross checked in the bigram dictionary, if found then the bigram match count is increased otherwise each bigram is further split into two unigrams. These unigrams are then checked against the unigram dictionary. Refer Figure 5.1 for diagrammatic representation of the algorithm.

Method	Product Reviews (%)	Movie Reviews (%)
Adjective Presence		
HSWN (Baseline)	67.50	**68.04**
Bi-Lingual Dict (Baseline)	**70.30**	67.30
Translated-Dict (Baseline)	62.30	61.05
Our Lexicon (HSL)		
Baseline	73.15	77.65
Baseline + NH	73.87	77.65
Baseline + Stem	75.70	70.70
Baseline + Morph	79.14	77.92
Baseline + Stem + NH	76.74	71.4
Baseline + Morph + NH	**81.57**	**78.02**
Adjective Scoring		
Baseline	73.96	**71.42**
Baseline + NH	74.88	70.32
Baseline + Stem	77.44	68.03
Baseline + Morph	78.14	70.32
Baseline + Stem + NH	78.41	68.50
Baseline + Morph + NH	**79.14**	69.23
Adjective +Adverb Presence		
HSWN (Baseline)	69.30	68.4
Bi-Lingual Dict (Baseline)	**71.40**	**70.3**
Translated-Dict (Baseline)	63.90	62.40
Our Lexicon (HSL)		
Baseline	73.70	77.12
Baseline + NH	73.89	78.12
Baseline + Stem	76.03	71.79
Baseline + Morph	80.7	78.12
Baseline + Stem + NH	77.34	72.52
Baseline + Morph + NH	**81.87**	**79.12**
Adjective + Adverb Scoring		
Baseline	74.62	**73.6**
Baseline + NH	74.96	71.8
Baseline + Stem	78.27	68.8
Baseline + Morph	80.54	73.26
Baseline + Stem + NH	79.03	69.23
Baseline + Morph + NH	**81.28**	71.4

Table 5.2 Results for Review Classification

Method	Features	NB(3)	NB(5)	SVM(3)	SVM(5)
Unigram	1934	76.8	78.57	83.57	83.28
Bigram	6063	69.57	70.14	74.28	75
Uni+Big	7998	77.85	78.14	81	**84.42**
Uni+Big+Tri	15581	78.42	78.57	81.14	82.48

Table 5.3 Classification Accuracy using N-Gram and its combination for Product Reviews

Method	Features	NB(3)	NB(5)	SVM(3)	SVM(5)
Unigram	1934	79	77	82.58	**84.42**
Unigram (TF=2)	664	77.28	79	82	**82.57**
Unigram (TF=3)	415	77	78.85	**80.57**	**80.57**

Table 5.4 Classification Accuracy using N-Gram with feature pruning for Product Reviews

We also propose a scoring function which gives priority to trigram matching followed by bigrams and unigrams.

Score = x * Count Tri-grams + y * Count Bi-grams + z * Count Uni-grams

here x = 7/11, y = 3/11, z = 1/11,

Count N-gram = Number of N-grams matched (N = Uni/Bi/Tri).

The values 7, 3, 1 are chosen to ensure that

(1) score for matching a trigram >score for matching 2 bigrams.

(2) score for matching a bigram >score for matching 2 unigrams.

In the scoring function we have given the least possible integer value to unigram, bigram and trigram keeping the above constraints in mind. The rationale behind having these constrains while deciding the values of x, y, z was that higher n-gram carries more weight then a lower n-gram and also matching of a higher n-gram should be weighed more than matching of two lower n-grams. Then we have normalized these values on a scale of 0 to 1. So the final x, y, z parameters are x=7/11, y=3/11 and z=1/11. The results are discussed in Table5.5, this approach was carried on the combination of unigram+bigram+trigram model performed over 3 fold and 5 fold cross-validation.

Approach	Accuracy
Weighed N-Gram (3 fold)	76.2
Weighed N-Gram (5 fold)	77.7

Table 5.5 Classification Accuracy using Weighed N-Gram function

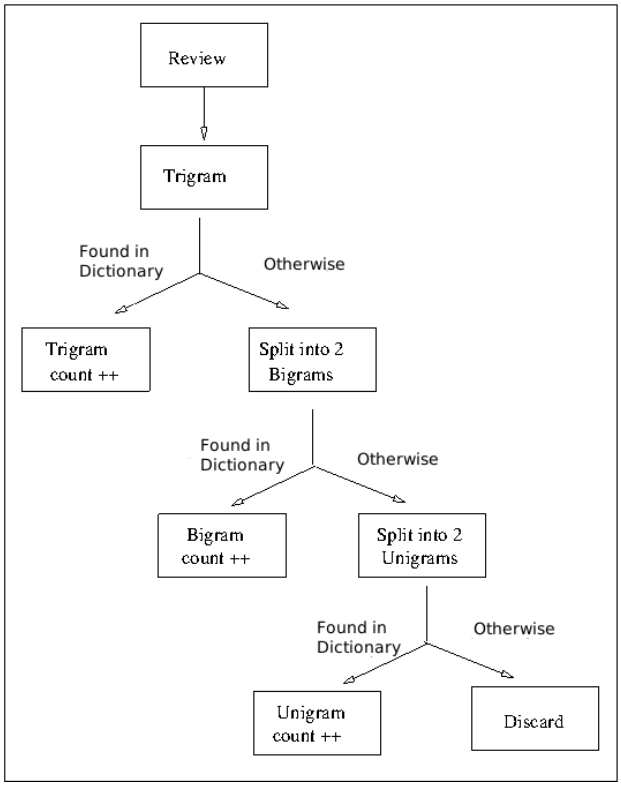

Figure 5.1 Weighed N-Gram Algorithm

5.1.2.3 Combining Lexical and N-Gram Features

The two approaches 1) Using Subjective Lexicon and 2) Using N-Gram discussed above provides valuable information, the lexicon scores gives an idea of the prior and the N-Gram scores gives an idea of the semantic measure. In the above two approaches (Weighed N-gram and Subjective Lexicon) we devised scoring function and calculated the polarity of an opinion but it might be the case that the function we used is biased, so in this approach we formed feature vector incorporating the N-Gram information and also including the subjective lexical information[5], we used machine learning algorithms for performing sentiment classification. We used WEKA toolkit for classification of the testing set (reviews). The feature vector was devised for both the approaches.

The feature vector for a particular review is of the format “f1, f2, f3, f4 f5 ,f6, f7, f8, class”
pos_score, neg_score, pos_tri, pos_bi, pos_uni, neg_tri, neg_bi, neg_uni, class
The features that were used for classification are-

[5]We have used Hindi subjective lexicon(HSL) generated by our algorithm for performing combined N-Gram + lexical experiment

- Feature-1 It provides the positive score of a review which is calculated using the Hindi Subjective Lexicon.
- Feature-2 It provides the negative score of a review which is calculated using the Hindi Subjective Lexicon.
- Feature-3 It denotes the number of positive trigrams matched .
- Feature-4 It denotes the number of positive bigrams matched .
- Feature-5 It denotes the number of positive unigrams matched .
- Feature-6 It denotes the number of negative trigrams matched .
- Feature-7 It denotes the number of negative bigrams matched .
- Feature-8 It denotes the number of negative unigrams matched .
- Class Actual class of the review positive or negative.

The results for this approach are mentioned in the below table-

Method	**Features**	**NB(3)**	**NB(5)**	**SVM(3)**	**SVM(5)**
Weighed N-Gram	6 (only N-Grams)	65.8	67.7	71.2	72
Weighed N-Gram + Lexical	8 (N-Grams + subjective scores)	72.14	72.57	79.85	**80.4**

Table 5.6 Classification Accuracy using Weighed N-Gram and Weighed N-Gram+Lexical Approach

5.1.3 Results Analysis

Results in Table 5.2 highlights that the lexicon generated using translating the sentiwordnet performs poorly over both Products and Movie reviews dataset. The bilingual dictionary performs better than the lexicon developed in [24]. The lexicon generated by our algorithm performs reasonably well over both the datasets we used for experiments. Including adverbs along with adjectives leads to an increase of 1-2% in the classification problem.

Using the proposed strategy for negation handling, it shows a slight (about 0.5-1%) improvement in classification of reviews. We proposed the use of stemmer and morph analyzer to identify the root word for adjectives which were present in the reviews but went missing from our lexicon. Stemming and Morph both showed an improvement of 3%- 6% in classification of reviews. Table 5.1 lists a few mapping of words to their stemmed and morphed form. For product reviews the highest accuracy we obtained was 81.87% and for movie reviews the highest accuracy we obtained was 79.12% for adjective and adverb presence approach, in the presence of negation

handling and mophological analysis of the words.

Using N-Gram modeling the highest accuracy we obtained is 84.42% using SVM classifier on performing 5 fold cross-validation. Table 5.4 describes the results with pruning the features with heuristics like Term Frequency (TF). The gradual decrease of features also results in the droping of the classification accuracy. With just a small set of 700 reviews, our weighed N-Gram approach performed quite well and showed an accuracy of 80.4% with SVM classifier *Refer* Table5.6.

5.1.4 Discussion

The problem of sentiment classification has been approached using three methods. We described and presented the results above, the first approach of scoring the review using lexicon provides a prior subjective score of the review and the N-Gram model helps in attaining the score with respect to the words usage and distribution. We tried the traditional N-Gram and also weighed N-Gram approach towards performing sentiment classification of Hindi reviews, the results were quite good but the data available is very less to experiment with. When we increase the N-Gram to 3 or beyond it was not able to match and the training wasn't much effective. We also performed negation handling, stemming, mophological analysis to decrease the information missing from the lexicon and the N-Gram model. The stemming, morph-analysis and negation handling lead to an increase of about 7% on adjective presence and about 6% on adjective scoring over the baseline.

The results of the lexicon approach are dependent on the coverage of the lexicon, if the coverage is less, then for most of the reviews the lexicon will output no results or null values. Lexicon can be used as a prior giving some information about the subjective aspect but using this as a whole single mechanism of classification has issues until it captures the N-Gram information. While N-Gram approach show promising results but with a small dataset of 700 reviews even learning the optimum model becomes a challenge. For future experiments we need to have some more data to experiment using semi-supervised approach for classifying the input.

As indicated the main challenges in the Hindi language includes morphological variations and multiple spellings. We tried to handle these issues by using morph analyzer and stemmer but even the performance of same is not that good. Apart from the structure view point other challenges come with identifying the context and incorporating the contextual information into the algorithm.

5.2 Hindi Blogs Analysis

A blog is a personal diary published on web (generally) by individuals or a group of individual. Blogs usually contain personal opinions or experiences of the blogger. In this section, we explain the techniques which are used to perform sentiment analysis, using pre-computed lexicons, using N-Gram Techniques and using supervised classification. The dataset we are using in this research comprises of 250 blogs, the dataset is quite heterogeneous in nature as discussed in table 3.2 it has about 733 positive sentences, 1427 negative sentences and 3478 objective sentences, so the uneven distribution of the polarity makes sentiment classification over blogs more complex.

5.2.1 Using Subjective Lexicon

In this method of sentiment classification we used the subjective lexicons discussed in Chapter-4. As discussed earlier for the review classification we experimented using the four lexicons *Refer* Table 4.13, similarly we perform the blog sentiment classification using the four subjective lexicons.

Given a blog (a collection of sentences), we calculate the polarity of the sentence by summing the polarity of individual words which constitute the sentence. The objective score of a word is calculated as the difference of positive and negative score from one. Overall polarity of the blog is calculated by summing the polarity of individual constituting sentences. We performed negation handling using a window of upto two words to the left of adjective and upto two words to the right of adjective. Words like नहीं ,न etc were checked as negation words which invert the polarity. We also performed Morphological analysis to reduce the missing adjectives and adverbs information.
Table 5.7 reports the results of sentiment classification performed using the subjective lexicon.

Number of Sentences	5638			
Method	**HSL**	**H-SWN**	**Bi-Lingual Dict**	**Translated Dict**
Subjective Lexicon	54.2%	53.6%	55.25%	48.30%
Subjective Lexicon + NH	53.98%	53.11%	55.11%	47.90%
Subjective Lexicon + NH + Morph	54.8%	53.90%	**55.90%**	49.20%

Table 5.7 Classification Accuracy using subjective lexicon

Using Hindi Subjective Lexicons, we also built features for each sentence which are frequencies of positive, negative and objective words and sum of their scores as per the lexicon. Our features for lexical approach are <PosWords, NegWords, ObjWords, PosScore, NegScor, ObjScore, Class

$>$. We tested Lexical features on the complete dataset of 5638 sentences.
Table 5.8 reports the accuracy of our lexical features on SVM and Naive Bayes classifier

	Accuracy	
Features	**SVM(5 fold)**	**Naive Bayes(5 fold)**
Lexical (HSL)	61.11%	58.42%
Lexical (HSL) + Negation Handling	61.56%	58.6%
Lexical (Bi-Lingual Dict)	61.58%	58.21%
Lexical (Bi-Lingual Dict) + Negation Handling	61.6%	58.38%

Table 5.8 SVM and Naive Bayes Classifier on Lexical Features

Both of the approaches discussed above perform poorly for the problem of sentence level blogs classification. The lexicon approach showed the maximum accuracy of 55.90% which is very less as compared to the baseline of 61.6% considering the majority class i.e objective. The feature based lexical approach showed an accuracy of 61.6%. Much of the information is missing and is not captured by the lexicon approaches effectively.

Limitations- Classification using lexicon(s) have several limitations which are listed below.

- **Poor Coverage:** The lexicons used in this research had limited coverage. In our dataset, we had unique 1974 adjectives out of which we found only 925 adjectives in HSL and 775 in Bi-Lingual dict. Both of these lexicons contain only base words without any morphological variants as mentioned in Table 5.1. Most of the morphological variants are not listed in the lexicon and thus while handling the data it becomes important to perform Morphological analysis and stemming.

- **Context Dependency:** Lexicons in general are unable to capture the context dependency. Lexicons only look at the token at a time and fail to relate the token with other preceding and successive tokens. Blogs exhibit context dependency not only within the sentence but across multiple sentences.

5.2.2 Using N-Gram Technique

In this method, we performed the bag of word formulation where we formed the feature vector for the N-Grams as discussed earlier in Section 5.1 (product review classification). In this technique, we form a feature vector of unique N-Grams using the annotated training corpus, each sample

is represented using 1's and 0's which denotes the presence or absence of a feature. We form a training model of unigram from this corpus using positive, negative and objective sentences. For a testing sample we form a feature vector of 0's and 1's and assign the class to the sample using the training model formed. As described in Table 3.2, our dataset comprises of unequal number of positive, negative and objective sentences. We randomly selected 733 negative and objective sentences to balance the training data. We perform this experiment for five cycles i.e we randomly select 733 negative and objective sentences 5 different times and perform the experiments, so as to avoid the random biased selection.

Table 5.9 shows Classification Accuracy for N-Gram Technique. We sampled equal number of sentences from negative and objective set 5 times and each time applied 5 fold cross validation.

Number of Sentences	733 (each)
Random Set	Accuracy
Sample 1	54.35 %
Sample 2	54.00 %
Sample 3	**55.45 %**
Sample 4	53.75 %
Sample 5	53.20 %
Averaged	**54.15 %**

Table 5.9 Blogs Classification Accuracy using N-Gram

5.2.3 Weighed N-Gram and Lexical approach

Supervised Machine learning is widely used in performing the task of sentiment analysis. Here, we use SVM and Naive Bayes classifier from weka toolkit to perform blog sentiment analysis. We built N-Gram and lexicon based features and test them on SVM and Naive Bayes classifier. We use frequency of matching trigrams, bigrams and unigrams for each class as N-Gram features. $<$Tri Pos, Bi Pos, Uni Pos, Tri Neg, Bi Neg, Uni Neg, Tri Obj, Bi Obj, Uni Obj,Class $>$. We tested N-Gram features on a balanced dataset (733 sentences each) i.e. equal number of positive, negative and objective sentences. Table 5.10 reports accuracy of N-Gram features and N-Gram + Lexical features averaged over five sample sets of 733 + 733 + 733 (positive + negative + objective) sentences.

N-Gram and Weighed N-Gram approaches also perform badly and show maximum accuracy of 54.15% and 56.15% respectively. The data is heterogeneous in nature which adds more complexity and makes the 3-class (Positive, Negative and Objetcive) classification a difficult task.

Number of Sentences	733 (each)	
Feature Set	SVM	Naive Bayes
N-Gram	54.4%	51.6%
N-Gram + Lexical	**56.15%**	53.5%

Table 5.10 Weighed N-Gram and Lexical Classification Accuracy

As the sentences distribution across different polarities is non-uniform so we also experimented with breaking this problem into two step approach-

1. Step-1 Performing 2 class classification Subjective vs Objective.
2. Step-2 Subjective class is further classified as Positive vs Negative.

So now instead of three classes we have 2 class classification problem and the data available is distributed a little better as compared to the 3 class problem distribution-

Step-1- Subjective (2200) vs Objective (3478) classification

Step-2- Positive (773) vs Negative (1427) classification

We performed the Weighed N-Gram and Weighed N-Gram+Lexical approach on Subjective vs Objective and further Positive vs Negative classification problem. The results are as discussed in table 5.11 and 5.12.

Number of Sentences	2200 (each)	
Feature Set	SVM	Naive Bayes
N-Gram	59.35 %	58.2%
N-Gram + Lexical	**61.71 %**	59.28%

Table 5.11 Subj vs Obj Classification Accuracy using N-Gram and N-Gram + Lexical Features

Number of Sentences	733 (each)	
Feature Set	SVM	Naive Bayes
N-Gram	50.64%	48.44%
N-Gram + Lexical	**51.59%**	49.59 %

Table 5.12 Pos vs Neg Classification Accuracy using N-Gram and N-Gram + Lexical Features

Breaking the sentence level blog classification problem from 3-class (Positive, Negative and Objective) to 2-class (Subjective and Objective) showed a small improvement. The above exper-

iments inidcates that the traditioanl aproaches such as N-Gram and using Subjective Lexicon doesn't work well for the blogs as compared to the product and movie reviews.

Limitations- N-Gram technique classification also fails at many places. In particular to blog sentiment analysis, N-Gram technique fails miserably because of the following reasons-

- **Vocabulary mismatch:** Blogs are written by different people who belong to possibly different culture and zones. There is a huge diversity in the composition of words, vocabulary is influenced by local language words.
- **Diverse Topics:** Blogs are written on diverse topics like politics, sports, current affairs, etc. Each of these domains have their specific words and their usage, so it becomes difficult for the model to capture the essence of the words and also there are many cases where N-Grams are not found in the training model. We need more large datasets to have effective learning.
- **Context Dependency:** N-Gram techniques generally capture the context been talked about within the sentence level. But in particular to Blog analysis, context flows across multiple sentences which is not captured properly by N-Gram models.

5.2.4 Results Analysis

In this research we present two traditional approaches for the sentiment classification, we performed a 3-class (positive, negative and objective class) classification. The first approach uses the prior information and scores the sentences using the lexicon method. Apart from normal scoring using the lexicon, we also considered negation handling to incorporate the effects of words which reverse the polarity. We also perform stemming and morphological analysis as shown in Table 5.1 to minimize the missing information.

The performance of all the lexicon is similar, with best accuracy of 55.90% which is far less than the baseline of 61.6% (assigning default class to all i.e all sentences as objective). The reason being because of the limitation of the lexicon in terms of the coverage and the vocabulary difference between the user-generated content and the lexicon words. The lexical features with negation handling showed an accuracy of 61.6% similar to the baseline as shown in Table 5.8. In another approach we performed N-gram technique, the experiments done with scoring function had an average accuracy of 54.5% as shown in Table 5.9, we also performed supervised learning using the features formed by N-gram approach as discussed in section 5.1.2.2 . Both the approaches have their own limitations, we also performed an experiment to combine the features of both the above approaches. Lexical approach captures the prior information and N-Gram information incorporates context information to some extent, a combined approach perform relatively better. This combined approach using N-Gram and Lexical features showed an improvement of about 2% using SVM on 5 fold cross-validation *Refer* Table 5.10. The experiment was done on 2199

sentences (733 each class), so one reason for low performance could be accounted because of less training data. We also tried with dividing the blogs classification problem from 3 class classification to 2 steps 2 class classification problem-

Step-1 Subjective vs Objective Classification

Step-2 Subjective into Positive and Negative Class

The results as shown in Table 5.11 and Table 5.12 didn't show much improvement as compared to the results of 3 class classification problem *Refer* Table 5.10.

Even it's a challenging task to label the blogs as positive, negative or objective at document level and further labeling the blogs at sentence level. Blogs are written kind of a diary entry, where writers often make comparisons and make referrals to entities or events. Thus the context information needs to be incorporated to deal with classification at sentence level.

5.2.5 Discussion

Using lexicon for blog sentiment classification, we achieved an accuracy of 56% which is very less as compared to general text (reviews) sentiment classification. Accuracy of 55% achieved using N-Gram (N-Gram) method is also below the par level. Accuracy on supervised machine learning algorithms (SVM and Naive Bayes) reached a maximum of 61.6% for N-Gram + lexical (HSL) features. Dividing the 3 class classification problem into 2 class classification even doesn't help much in terms of the sentiment classification at blogs level.

In this research, we found that the nature of blogs is also quite different from the other sources of subjective information such as product reviews, tweets etc. On one hand the data is unequally divided into positive, negative and objective sentences and on the other hand the complexity in blogs increases because of large use of idioms, sarcastic phrases, referral sentences, proverbs etc. In the dataset we used, the vocabulary been used by the blogger is also rich and thus increases the challenge of classification of blogs. The factual descriptive nature of blogs uses a large mention of quotes, proverbs, jokes, idioms etc. In figure 5.2, we quote some sentences from the blog dataset, indicating sarcasm, idioms, proverbs and indirect subjective sentences. Sarcsam is a problem which is even hard for the humans to annotate. It adds more complexity for the system to classify. Solving it is beyond the scope of this work. In this research we are describing the challenges while addressing the problem of sentiment analysis for Hindi blogs.

We described the fact with experiments that the traditional approaches followed for mining sentiments in the text such as N-Gram and using subjective lexicons does not work for the blogs genre as expected. The problem of mining opinion in blogs require some other methodology which uses the inherent properties of blogs. Typically blogs are written around some target entities and

is composed of a mixture of these entities. Generally, blogger expresses his opinion about these entities and thus it would be useful to mine the entities first and then the opinion orientation about them .

In future, we can work on better methods to extract entities from the blogs in the absence of tools like Named Entity Recognizers for Hindi and other Indian Languages and extracting the modifiers which highlights the bloggers opinions towards that entity/object. We can also come up with a robust sense based subjective lexicon which can be used to correctly identity the polarity of the modifiers.

5.3 Summary

In this chapter, we perform the task of sentiment classification for user-generated content in Hindi Language. We worked for two genres 1) Reviews Classification and 2) Blogs Classification. The Hindi reviews has 2 datasets -1) Product Reviews Dataset, developed as a part of this research work, it has about 700 reviews, 2) Movies Reviews Dataset, it has about 273 reviews. The Hindi blogs are crawled from BBC Hindi and have been manually annotated at sentence level as Positive, Negative or Objective. We used various techniques to perform sentiment classification for both of the genres. We tried approaches such as 1) Using Subjective Lexicon 2) N-Gram Modeling and 3) Combining N-Gram and Lexical information. We also performed stemming, morphological analysis and also used negation handling to minimize the missing information. As compared to reviews, blogs performed quite badly indicating that the traditional techniques of sentiment analysis doesn't work effectively for the blogs. The nature of the blogs is quite different as compared to reviews and other genres. Hence in future we need to come up with the effective techniques of mining the entities and capturing events from the blogs and mining sentiments accordingly.

- Sarcasm
  - बात करते हैं तो, **मानो फावड़ा चला रहे हैं**
  - भारत में विकास को लेकर ज़्यादातर बहसें इसी तरह ख़त्म होती हैं- **पहले अंडा या पहले मुर्गी**

- Idioms
  - ऐसे में पीयूष चावला **तुरुप का इक्का** साबित हो सकते हैं.
  - मायावती की पहली प्रतिक्रिया तो जैसे, **नहले पर दहला थी**.

- Proverbs
  - **चित भी मेरी, पट भी मेरी**
  - **गिरते हैं शहसवार ही मैदान-ए-जंग में**

- Indirect Subjective Sentences: These sentences are of special kind. These sentences doesn't have any polar (subjective) word but still they convey some polarity.
  - **ये कश्मीर की आग है**
  - **गुजरात का जवाब नहीं है**

Figure 5.2 Idioms, Proverbs usage in Hindi Blogs

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Sentiment Analysis has been quite popular and has lead to building of better products, understanding user's opinion, executing and managing of business decisions. With rapidly increasing technology, the early approach of word-of-mouth has been shifted towards the mass opinion what the people like and appreciate in majority. People rely and make decisions based on reviews and opinions. The rise in user-generated content for Hindi language across various genres- news, culture, arts, sports etc has open the data to be explored and mined effectively, to provide better services and facilities to the consumers.

The scarcity of resources is one of the biggest challenge while dealing with sentiment analysis for Hindi language. In this work, we focussed on resource creation which includes building of annotated datasets and subjective lexicon for Hindi language. We build Product reviews dataset which has 700 reviews (350 positive and 350 negative reviews) which are manually verified. We crawled 250 blogs from BBC hindi, which are manually annotated at sentence level as Positive, Negaive or Objective.

We designed an algorithm to generate Subjective Lexicon using the Wordnet and a pre-annotated seed list of words for a particular language. This approach can be tried with any language which has wordnet or some other resource in which words are connected using synonyms or antonyms relations. We checked the validity of our algorithm for English and Hindi language and evaluated the lexicon generated thoroughly with multiple strategies such as-

1. Comparison against Manual Evaluation
2. Comparison with already Existing Resources
3. Classification Performance

As a part of this research, we performed sentiment classification on movies and product reviews and on blogs dataset. We also discuss in detail the challenges while working with user-generated content in Hindi language and address some of the problems as discussed in Chapter-1. We worked on three different approaches for performing sentiment classification and discussed the merits and demerits of each approach-

1. Using Subjectivity Lexicon
2. N-Gram approach
3. Weighed N-Gram approach

The main contributions as a part of this research can be divided into three sections-

1. Data Creation: This involves building resources, subjective lexicon and data-set for Hindi language which are made public to be used and developed further by the research community.
2. Language Independent Subjective Lexicon creation: Developing an algorithm which can construct Subjective Lexicon for any language using the wordnet and a pre-annotated seedlist of words of that language.
3. Sentiment Classification: We performed sentiment classification for two genres of Hindi User-Generated Web Content-
   -Reviews- Hindi Product and Movie Reviews Dataset
   -Blogs- Hindi Blogs Datatset

6.2 Future Work

In India, there are 22 official languages and 13 languages have more than 10 million speakers. With multiple sources of data available for each language, it is easy to gather data and analyze them.

In context to Indian Laguages, earlier work done for sentiment analysis has been on Bengali and Hindi, rest all the languages are unexplored. The nature of Indian languages varies a great deal in terms of the script, representation level and linguistic charcteristics etc. So, there is a large amount of work that needs to be done to understand the behaviour of Indian languages and perform the analysis of same accordingly. Subjective lexicon that we developed as a part of this research work requires just the seedlist and a resource in which words form a connected network. We have manually aligned multilingual dictionary available for English and 11 Indian languages (English, Hindi, Telugu, Tamil, Punjabi, Urdu, Malayalam, Oriya, Kannada, Gujarati, Bengali and Marathi). In future, we can try and come up with more focused approach and other heuristics to develop

subjective lexicon for the languages which doesn't even have a wordnet but a structure in form of a multilingual dictionary.

We need to come with a notion of prior sentiment polarity for set of words in the form of a subjective lexicon. Then we can explore and dig in depth regarding the task of sentiment classification for the web text and improve over same. The main issues while working with Indian languages comes- while handling the morphological variations, identifying context, performing word sense disambiguation and handling multiple spellings with each of them in itself is a research problem. Much work needs to be done at this level to address the above challenges that can boost up and help the research in the related areas. As discussed above there is lack of annotated datasets and resources for Indian languages, so it needs considerable focus and time to be given. The basic resources like part of speech tagger, morphological analyzers, named entity recognizers and parsers have also not yet reached the state of art accuracy and needs improvement. Once we have sufficient data to experiment with, various machine learning techniques can be easily used and applied to learn from the text more effectively.

Related Publications

- **Piyush Arora**, Akshat Bakliwal and Vasudeva Varma. *"Hindi Subjective Lexicon Generation using WordNet Graph Traversal"*
  In the proceedings of 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012), New Delhi, India
- **Piyush Arora**, Akshat Bakliwal and Vasudeva Varma. *"Hindi Subjective Lexicon : A Lexical Resource for Hindi Polarity Classification"*
  In the proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey.

Other Related Publications

- **Piyush Arora**, Akshat Bakliwal, Ankit Patil and Vasudeva Varma. *"Towards Enhanced Opinion Classification using NLP Techniques"*
  In the proceedings of Sentiment Analysis where AI meets Psychology (SAAIP 2011), International Joint Conference on Natural Language Processing 2011, November13, Shangri-La Hotel, Chiang Mai, Thailand.
- **Piyush Arora**, Akshat Bakliwal and Vasudeva Varma. *"Mining Sentiments from Tweets"*
  In the proceedings of 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2012), Association of Computational Linguistics (ACL 2012), Jeju, Republic of Korea.
- **Piyush Arora**, Akshat Bakliwal and Vasudeva Varma. *"Entity Centric Opinion Mining from Blogs"*
  In the proceedings of Sentiment Analysis where AI meets Psychology (SAAIP 2012), International Conference on Computational Linguistics (COLING 2012), IIT Bombay, Mumbai, India

Bibliography

[1] A. Agarwal, F. Biadsy, and K. R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams, 2009.

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[3] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[4] A. Balahur, R. Steinberger, M. A. Kabadjov, V. Zavarella, E. V. der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva. Sentiment analysis in the news. In *LREC*. European Language Resources Association, 2010.

[5] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. S. Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007. Short paper.

[6] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[7] K. Bloom, N. Garg, and S. Argamon. Extracting appraisal expressions. In *Proceedings of Human Language Technologies/North American Association of Computational Linguists*, 2007.

[8] R. M. Carmen Banea and J. Wiebe. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of LREC'08*, 2008.

[9] P. Chesley. Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*, 2006.

[10] A. Das and S. Bandyopadhyay. *SentiWordNet for Bangla*. 2010.

[11] A. Das and S. Bandyopadhyay. *SentiWordNet for Indian Languages*. 2010.

[12] D. Das and S. Bandyopadhyay. Labeling emotion in bengali blog corpus a fine grained tagging at sentence level. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 47–55, Beijing, China, August 2010. Coling 2010 Organizing Committee.

[13] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. pages 519–528, 2003.

[14] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[15] N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA, 2010. ACM.

[16] G. Draya, M. Planti, A. Harb, P. Poncelet, M. Roche, and F. Trousset. Opinion mining from blogs. In *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*, 2009.

[17] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06*, pages 417–422, 2006.

[18] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. pages 1–6, 2009.

[19] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[20] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[21] B. He, C. Macdonald, J. He, and I. Ounis. An effective statistical approach to blog post opinion retrieval. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, 2008.

[22] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[23] T. W. Intelligent and T. Wilson. Annotating opinions in the world press. In *In SIGdial-03*, pages 13–22, 2003.

[24] A. Joshi, B. A. R, and P. Bhattacharyya. A fall-back strategy for sentiment analysis in hindi: a case study, 2010.

[25] J. Kamps, M. Marx, R. J. Mokken, and M. D. Rijke. Using wordnet to measure semantic orientation of adjectives. 2004.

[26] A. Karthikeyan. Hindi english wordnet linkage.

[27] L.-W. Ku, Y.-T. Liang, and H.-H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006.

[28] P. Melville, W. Gryc, and R. D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, 2009.

[29] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[30] S. min Kim. Determining the sentiment of opinions. In *Proceedings of COLING*, pages 1367–1373, 2004.

[31] S. min Kim and E. Hovy. Identifying and analyzing judgment opinions. In *Proceedings of HLT/NAACL-2006*, pages 200–207, 2006.

[32] D. Narayan, D. Chakrabarti, P. Pande, and P. Bhattacharyya. An experience in building the indo wordnet - a wordnet for hindi. In *First International Conference on Global WordNet*, 2002.

[33] T. Nasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, K-CAP '03, pages 70–77, New York, NY, USA, 2003. ACM.

[34] B. Ohana and B. Tierney. Sentiment classification of reviews using sentiwordnet. 2009.

[35] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[36] D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 675–682, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[37] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[38] H. Saggion and A. Funk. Interpreting sentiwordnet for opinion classification. In *LREC*, 2010.

[39] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA, 1966.

[40] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, 2002.

[41] J. Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740. AAAI Press, 2000.

[42] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 246–253, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[43] T. Wilson. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354, 2005.

[44] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 129–136, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[45] C. Zhang, W. Zuo, T. Peng, and F. He. Sentiment classification for chinese reviews using machine learning methods based on string kernel. In *Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology - Volume 02*, pages 909–914, Washington, DC, USA, 2008. IEEE Computer Society.

[46] W. Zhang and et al. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (2007)*, 2007.