Sentiment Analysis of Social Media for Hindi Language

A Thesis
Submitted in the fulfillment of the requirements for the award of the degree of

Doctor of Philosophy

Submitted by
Sujata Rani
(Registration No. 951403004)

Under the supervision of
Dr. Parteek Kumar
Associate Professor
Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology, Patiala

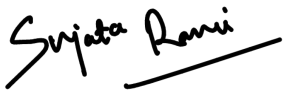
THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

December 2019

Certificate

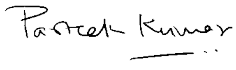
I hereby certify that the work which is being submitted in this thesis entitled "*Sentiment Analysis of Social Media for Hindi Language*", in fulfillment of the requirements for the award of the degree of DOCTOR OF PHILOSOPHY submitted in Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of Dr. Parteek Kumar and refers work of other researchers which are duly listed in the reference section.

The matter presented in this thesis has not been submitted for the award of degree in any other University.



(Sujata Rani)
Regd. No. 951403004

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge and belief.



(Parteek Kumar)
Associate Professor, Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala-147004 (INDIA)
Supervisor

Acknowledgement

First of all, I express my gratitude to the Almighty, Who blessed me with the zeal and enthusiasm to complete this work successfully.

I would like to thank my supervisor *Dr. Parteek Kumar* for his valuable guidance. He has always been there to guide me towards the right direction whenever I got stuck with the ideas. I want to thank him for his valuable evenings and week-ends that he spent with me in finalizing the writings of this thesis work. I also want to thank Mrs. Kumar to provide a conductive environment in her home during discussions on this proposed work.

I express my gratitude to the Doctoral Committee for monitoring the progress and providing valuable suggestions for improvement of my research work. I feel very proud and thankful to the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala for providing all the resources for a good research work. I am also very thankful to *Dr. Maninder Singh* (Head CSE department, TIET, Patiala) and all my colleagues for their constant support and cooperation during my research work.

This thesis would have been impossible without the support of my family. My deep regards to my father *Mr. Labh Chand* and my mother *Mrs. Sunita Rani* for their patience and love. Without them this work would never have come into existence. They have provided me with lessons on honesty and ethics and their humbleness and patience have always amazed me. I feel myself fortunate to have truly understanding husband *Mr. Vaibhav Agarwal* whom I owe my loving thanks for his unflagging love and support throughout writing of this thesis. I would also thank my brother *Ravi Kumar,* my sister *Nandani,* in-laws *Mr. Vivek Agarwal* and *Mrs. Rashmi Mittal,* and my brother-in-law *Mr. Utkarsh Agarwal* for their lifetime love, companionship, and support.

At the end I wanted to thanks all my friends at Palvi Agarwal, Gaganjot Kaur, Mukesh Goyal, Pankaj Garg, Nidhi Kalra, Sachendra Singh Chauhan, Vandana Bhatia, Meenu Singla and Bharti Saneja who have been stress busters for me.

(Sujata Rani)

Abstract

In recent years, due to the availability of voluminous data on web for Indian languages, it has become an important task to analyze this data to retrieve useful information. Because of the growth of Indian language content, it is beneficial to utilize this explosion of data for the purpose of sentiment analysis. There are various applications of sentiment analysis in different domains such as recruitment, education, marketing, policy making, unemployment, fighting riots, terrorism, and education, etc. This research contributes to the development of Hindi sentiment analysis system for aspect, sentence and document level. The system is able to perform the sentiment analysis of Twitter posts. The system is available online at www.hindisenti.com.

Hindi is the official language of India belonging to the family of Aryan languages. It is the 4th most spoken language with 310 million speakers across the world. In India, Hindi is spoken by a total of 422 million speakers; it's about 41% of total population of India. Therefore, there is a need to perform sentiment analysis in Hindi language so that the opinions of users in Hindi can be easily classified and proved useful for the users in decision making. Iin today's life, mostly people share their opinions on social media platforms. This motivated us to explore the field of sentiment analysis on social media for Hindi language. Although there are many differences in language structure of English and Hindi which arise different challenges while performing sentiment analysis on text dataset.

This research work presents the description about the general process of sentiment analysis at different sentiment levels, i.e., aspect/feature, sentence and document level. This research depicts a systematic review in the field of sentiment analysis in general and Indian languages specifically. The current status of Indian languages in sentiment analysis is classified according to the Indian language families. The periodical evolution of Indian languages in the field of sentiment analysis, sources of selected publications on the basis of their relevance are also described. Further, taxonomy of Indian languages in sentiment analysis based on techniques, domains, sentiment levels and classes has been presented. This research work will assist researchers in finding the available resources such as

annotated datasets, pre-processing linguistic and lexical resources in Indian languages for sentiment analysis and will also support in selecting the most suitable sentiment analysis technique in a specific domain along with relevant future research directions.

This thesis presents the architectures of SA system for Hindi language at sentence level and aspect level using ML and lexicon based techniques, respectively. To train the ML algorithms, corpus of reviews and tweets has been collected from online websites and Twitter, respectively. The corpus has been annotated by three Hindi native speakers and has been validated using the statistic kappa measure. The experimental results given by different ML algorithms have been measured using performance measures precision, recall and F-measure. To further improve the accuracy of the system, deep learning based CNN has been applied on the corpus of Hindi reviews. The experimental results suggest that properly trained CNNs can outperform the traditional ML algorithms for sentiment classification.

At aspect level, sentiment analysis has been perfoemd using lexicon-based technique. The system has been experimented on reviews dataset about products and movies in Hindi language. The proposed system uses two lexical resources Hindi Dependency Parser (HDP) and Hindi SentiWordNet (HSWN). It follows an efficient aspect extraction process to extract all the relevant aspects which include three steps, i.e., extraction of frequent nouns, identification of relevant nouns and removal of irrelevant nouns. The sentiment nodes are extracted using HSWN. The system uses HDP to determine the association between the aspect nodes and sentiment nodes. Also, the system generates a dependency graph and assigns the sentiment to the particular aspect having the least distance between sentiment word and aspect word.

This thesis also presents a case study of sentiment analysis for education domain by performing sentiment analysis of student feedback. The students' feedback has been collected from Coursera and SRS of the University using "R" language with natural processing techniques. The sentiments of students have been analyzed in the form of different emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, trust as well as positive and negative sentiments. Two new emotions satisfaction and dissatisfaction are derived from the existing emotions. The direct and indirect assessment

methods of course evaluation have been compared and it has been observed that both the methods provide converging evidence of student learning and teaching quality. Thus, this system can help an organization in improving student learning and teaching quality.

List of Figures

List of Tables

Table of Contents

Chapter 1

Introduction

Sentiment analysis, also known as opinion mining, is the field of study that helps in analyzing people's sentiments, attitudes, opinions, evaluations, emotions and appraisals towards different entities such as organizations, products, services, individuals, events, topics, and their attributes. There are also several names and slightly different tasks, e.g., sentiment analysis, opinion mining, emotion analysis, subjectivity analysis, opinion extraction, sentiment mining, affect analysis, review mining, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. In industries, the term sentiment analysis is commonly used while in academia, both sentiment analysis and opinion mining are frequently employed. However, both the terms represent the same field of study.

This chapter provides a high level view of this thesis. It discusses the fundamental concepts of sentiment analysis, its evolution, its key applications along with the major challenges of this area. It provides the process of sentiment analysis in brief by giving description about different stages. It further provides the motivation to propose sentiment analysis system for Hindi language along with its differences from English. It culminates with discussion of the organization of the rest of the thesis along with its contributions.

1.1 Sentiment Analysis: An Overview

Opinions or sentiments play an important role in decision making process. In earlier days, when a person wanted to take a decision about anything, he used to get the opinion from his friends, collegues or relatives. Similary, when organizations need an opinion about their services or products from public, they conduct surveys or opinion polls. Currently, people can share their opinions or reviews about services or products on nearly everything in various discussion forums, blogs, and social media sites. If a person wants to buy a product, he/she may get opinions or reviews easily available on the web. Therefore, there is not any requirement to conduct surveys to get customer reviews about their services or products because of the availability of the huge of information about the same.

The main goal of sentiment or opinion analysis is to identify whether a text, or a part of it, is objective or subjective. Objectivity shows that the text does not contain any opinionated content whereas Subjectivity shows that the text bears opinionated content. For example, the sentence "Kajol bought Samsung phone today." represents that objectivity as this sentence is a fact and conveys general information instead of an opinion or a view about any person. And the sentence "Samsung phone bought by Kajol is very expensive." represents the subjectivity as this sentence consists of an opinion and it discusses about the phone and the writer's feelings about same "expensive". The subjective text can be further categorized into three broad categories on the basis of the sentiments expressed in the text. For example, the sentence "I love to watch Star TV series." connotes the positive sentiment of writer about "star TV series" and the sentence "The movie was awful." connotes the negative sentiment about movie. In the same way, the sentence "I usually get hungry by noon." connotes neutral sentiment as this sentence consists of user's feelings hence it is subjective, also it does not consist of any positive or negative polarity, so it is neutral.

Sentiments are observed as the demonstration of a person's emotions and feelings. This field of computer science deals with analysis and prediction of the hidden information stored in the text. This hidden information gives valuable insights about user's taste, intentions, and likeliness. Esuli and Sebastiani (2006) define the sentiment analysis problem as having three different aspects: (a) determining the subjectivity of text (i.e., whether the text conveys any opinion or it is a fact); (b) determining the polarity of text, or deciding if a given subjective text expresses a positive or negative opinion; and (c) determining the strength of the polarity of the text (i.e., deciding whether the positive opinion expressed by a text is weakly positive, mildly positive, or strongly positive).

Since past few years, Sentiment and Opinion Analysis are increasingly being used as synonyms. Generally, Sentiment Analysis (SA) is a natural language processing task that uses the computational approach to identify and classify users' opinions from a piece of text into different sentiments such as positive, negative, or neutral and emotions such as happy, sad, angry, or disgusted to determine the user's attitude toward a particular subject or entity. Opinion mining is the mining of opinions of individuals, their appraisals, and

feelings in the direction of certain objects, facts and their attributes. Since, there are not substantial differences between the two therefore, both the terms "Sentiment analysis" and "Opinion mining" have been used interchangeably in this thesis.

To improve the readability of the thesis, the notation given in (1.1) is followed for each Hindi word mentioned in this thesis.

Hindi_Word *transliterated_Hindi_word* 'translated_Hindi_word' (1.1)

According to (1.1), for each Hindi word, its transliteration is given in italic followed by its English translation in inverted commas. For example, in this thesis, the word लड़का in Hindi is mentioned with its transliteration *ladaka* and its English translation '*boy*' as given in (1.2).

लड़का *ladaka* '*boy*' (1.2)

1.1.1 Levels of Sentiment Analysis

Sentiment analysis can be performed at three levels such as document, sentence and aspect/feature. Figure 1.1 represents the classification of levels of SA and the brief description about these levels is given as follows.

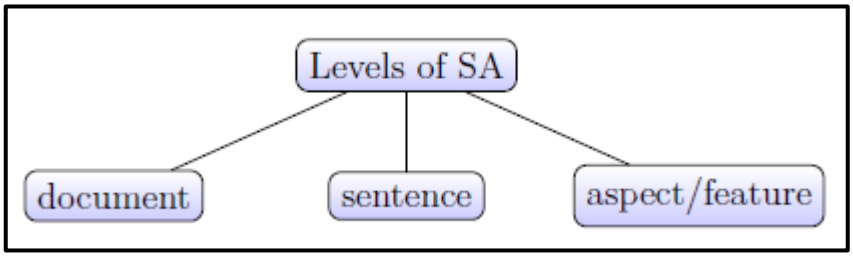


Figure 1.1 Levels of SA

a) **Document Level**- At this level, the whole document is classified into three sentiment classes, i.e., positive, negative or neutral (Pang et al., 2002; Turney, 2002). It is assumed that at this level, the document consists of opinion about any single entity, i.e., produc, service or an organization. Therefore, this level of analysis is not applicable to documents which consist of multiple entities. For

example, in case of a review about product, the sentiment analysis system determines what the review conveys overall positive or negative or neutral sentiment about that product.

b) **Sentence Level**- The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. This level of analysis is closely related to subjectivity classification (Wiebe et al., 1999), which distinguishes objective sentences that express factual information from subjective sentences that express subjective views and opinions. However, subjectivity is not equivalent to sentiment as many objective sentences can imply opinions, e.g., "We bought the car last month and the windshield wiper has fallen off". Though this sentence is objective but connotes negative sentiment about windshield wiper.

   At this level, comparative opinions are also handled as sometimes users do not provide a direct opinion about one product but instead provide comparable opinions. A comparative opinion is generally expressed on two or more objects based on their similarities or differences, and the object preferences are given by the opinion holder based on some of the shared features or attributes of these objects. Comparative opinion is described with the help of comparative or superlative form of an adverb or adjective, but it is not true for all comparative opinions. For example, the sentences, "Coke tastes better than Pepsi" and "Coke tastes the best" are two comparative opinions. Therefore, the goal of the sentiment analysis system in this case is to identify the sentences that contain comparative opinions, and to extract the preferred entity (-ies) in each opinion.

c) **Aspect/Feature Level**- Sometimes reviews consist of different sentiments towards different aspects/features of an entity and overall polarity does not help to identify the exact sentiments of people. Both the document level and the sentence level analyses do not discover what exactly people liked and did not like while aspect level performs finer-grained analysis (Hu and Liu, 2004). Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of

opinion). An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better.

For example, the sentence "Although the service is not that great, I still love this *restaurant*" clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the restaurant (emphasized), but negative about its service (not emphasized). In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence "The iPhone's call quality is good, but its battery life is short" evaluates two aspects, call quality and battery life, of iPhone (entity). The sentiment on iPhone's call quality is positive, but the sentiment on its battery life is negative. The call quality and battery life of iPhone are the opinion targets. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses.

This thesis focuses on all the three levels such as aspect/feature, sentence and document level of sentiment analysis.

1.1.2 Sentiment Analysis Process

Sentiment analysis process is performed through five phases as shown in Figure 1.2. The brief description about these phases is given as follows.

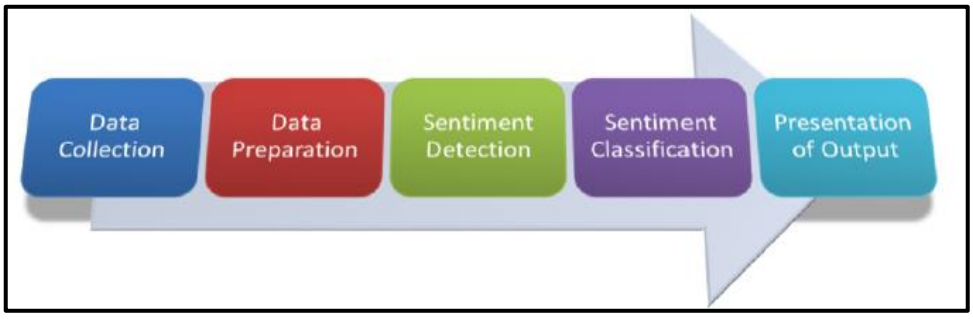


Figure 1.2 Sentiment analysis process

a) **Data Collection**- Sentiment analysis takes advantage of the huge amount of content generated by users over the Internet. There are many sources such as Weblogs, News, discussion forums, reviews and social networking sites for collecting real-time digital data. People share their views or opinions about public issues, companies, products or services etc. on these sources every day. This bulky amount of data can be extracted using API's of these sites. Once the data is extracted, it is then prepared for analysis.

b) **Data Preprocessing-** The data extracted in first phase may be in unstructured and non-processing form. The preprocessing helps in converting noise from high dimensional features to the low dimensional space to obtain as much accurate information as possible from the text. Therefore, in this phase, data is preprocessed to convert it into processing form. The irrelevant and non-textual content which is not useful for identifying sentiment is removed in this phase. The irrelevant data is removed by performing tokenization, removing stop words, normalization which includes casing of characters, negation handling, removal of punctuation marks, special characters, lemmatization, etc.

c) **Sentiment Detection-** The third phase in sentiment analysis process is sentiment detection. In this phase, computational tasks are performed to identify and extract the sentiment or opinion from the textual dataset. Generally, sentiment detection is performed at three levels such as aspect/feature, sentence and document level.

d) **Sentiment Classification**- The fourth phase is sentiment classification which classifies each subjective sentence into classification groups in the textual dataset. This phase can be performed by using machine learning techniques such as Naïve Bayes, Decision Tree, Support Vector Machines and Rule Based or lexicon based techniques. The classifications groups identified can be further classified into different moods like gladness, happiness, pleasure or satisfaction sorrow, regret, sadness etc.

e) **Presentation of Output**- The main aim of the analysis is to extract the meaningful information from the text that is sentiment expressed within the sentence towards a particular entity by an author/writer. Once the analysis is completed, the results of sentiment analysis can be displayed by number of ways

in presentation of output phase. Most commonly used among them are graphical displays such as pie charts, bar charts and line graphs.

The next section of this chapter discusses about the need of sentiment analysis.

1.2 Need of Sentiment Analysis

Very little computational study was carried out on opinions or sentiments prior to the introduction of World Wide Web (WWW) due to limited availability of opinionated text for such analysis. Because of increase in the contents on the WWW, the world has changed and became wealthy in data through advancement of Web. Due to availability of 4G networks for mobiles with high bandwidth and better mobility support, people can easily express their opinions on web (Bae et al., 2009). Therefore, an ocean of data has been generated by citizens on the Internet that did not exist even a few years ago. This data is passively generated by people simply by living their daily lives. This huge amount of data when aggregated and analyzed can reveal significant insights that help the users make faster and more informed decisions. Web 2.0 & 3.0 has led to an exponential increase in the user-generated content by providing varied mechanisms to interact with the users. The web presence of e-commerce and the entertainment industries has also 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 sentiment analysis of this huge data can provide valuable insights.

Social Media as a Huge Repository The technological advancement in the past two-three decades has enabled humans to find different ways to interact with each other. In the current digital environment age, Information Technology (IT) has been used extensively to record, store, and disseminate information digitally (Alansary et al., 2006). IT plays a key role in the success of many organizations and provides value to businesses in any field (Segura et al., 2016). One big revolution that became the part of the internet era is the social network revolution. Unlike, the traditional web-sites and corporate blogs, “Social Media Platforms” are used by the members to share, connect and work together with their peer groups to build durable relationships in the virtual world. This way of communication is known as “Social Networking” and this new medium of

communication is termed as “Social Media” (Parihar, 2012). With the evolution of social media and microblogging mediums, two major changes have occurred. First, it has replaced the traditional media like television and print media which were used as source to get the information about current events and second, it has provided a platform to common people to share their opinions and information (Gupta and Kumaraguru, 2012).

The social media sites lead to the generation of petabytes of data per week. India has 250.8 million Internet users and out of which 106 million are active Social Media users. Figure 1.3 illustrates the prediction of the number of social media users in India from 2015 to 2020.

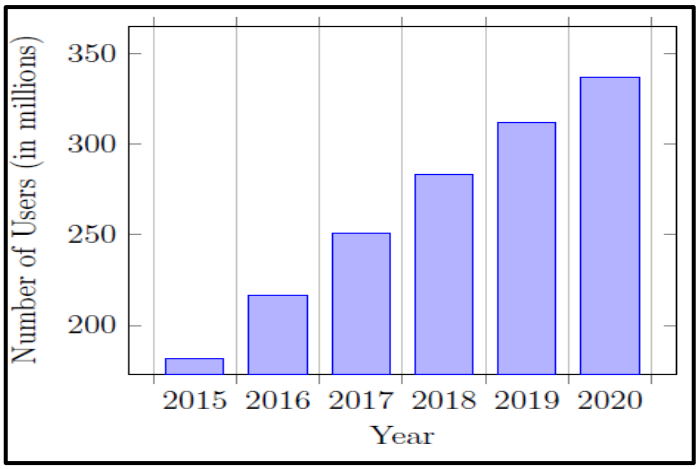


Figure 1.3 Number of social media users in India

The huge amount of data generated on social media gives an opportunity to extract real-time data from publicly shared opinions and issues by people. The major sources of real-time digital data are Facebook, Twitter, LinkedIn and Orkut. Orkut was the first social network to hit it off on a large scale in India. After Orkut, Facebook, Twitter, LinkedIn came into lead as social networking sites. These social media platforms play an important role during crises, providing valuable information to emergency responders and the public, helping reaching out to people in need, and assisting in the coordination of relief efforts (Gupta et al., 2014).

The brief overview about some of these popular and widely used social media platforms is given as follows.

- **Facebook:** With more than 2.38 billion monthly active users, Facebook is the most popular social network worldwide. India ranks first with 260 million users on Facebook according to the statistics published till end of April 2019 (Clement, 2019a). In 2023, the number of Facebook users in India is expected to reach 444.2 million, up from 281 million in 2018. The number of Indian Facebook users is growing 22 percent every six months. Facebook stores, accesses, and analyzes 30+ Petabytes of user-generated data.
- **Twitter:** India ranks eighth among the different countries such as United States, Japan, United Kingdom, Saudi Arabia, Russia, Brazil and Turkey with 7.86 million Twitter users (Clement, 2019b) and about 500 million tweets are done per day.
- **LinkedIn:** India ranks second after the US with a base of over 56 million users and a growth of more than 40 per cent in the last two years.
- **Google+:** It has about 440 million active monthly users and India ranks second after US with 1, 42, 339 users (Agarwal, 2019).

According to ComScore Metrix, March 2011, social networking sites reached 84% of the web audience in India, and taken up 21% of all time spent online. A Pew research study in December, 2012 established that nearly 45% of Indian web users connect on social media to discuss politics. Only Arab countries scored higher than India on this account (Rajput, 2014). Social media played an important role in Anna Hazare's anti-corruption movement, 2011 and 2012. Social media emerged more strongly in late 2012 and early 2013 public protests against the rapes in India. India's 2014 election is being called TwitterElection because it is the largest democratic election in the world till now. In 2015 Delhi election, sentiment analysis tool developed by IIT-B team helped in shaping the Aam Aadmi Party's (AAP) election strategy and to determine the swing in the electorate's sentiment towards AAP at any given point of time.

In November 2016, Twitter recorded 650,000 Tweets in 24 hours and millions more Tweets in the following weeks following Prime Minister Narendra Modi's announcement about the demonetisation of Rs. 500 and Rs. 1000 currency notes to fight against black money and corruption in India. Gurmeet Ram Rahim, the chief of Dera Sachcha Sauda,

was sentenced to 20 years of imprisonment in 2017 after he was convicted of rape. His arrest led to unprecedented violence and Indians across the world took to Twitter to debate their views and opinions on the conviction and the riots. In 2018, #KeralaFloods brought together government agencies, relief organisations, famous personalities and regular people on the platform to help rebuild Kerala. The people utilized Twitter to share information and crowdsource relief and assistance. In the same year, the heinous crimes committed against 8-year-old Asifa Bano led to widespread protests and international attention. Outraged citizens flooded Twitter, expressing their angst and demanding #JusticeForAsifa.

The statistics mentioned above gives an idea about the rate at which the data on the web has been increasing. With such vast data generated regularly, it provides enormous business opportunities to handle this data safely and precisely. This data is very crucial for market analysts, consumers, product developers and many others. But it is very difficult to analyze this enormous and valuable data shared on the web manually. In this scenario, sentiment analysis plays a vital role in extracting the sentiments or opinions of people. As stated by Liu (2010, 2012) and Jawale et al., 2013 in the latest years, sentiment analysis has attracted great deal of concentration from both the academicians and industry persons because of various challenging research issues and support of sentiment analysis for a broad set of applications. Figure 1.4 represents the evolution of trends in sentiment analysis in last 15 years and it shows that research work in sentiment analysis is exponentially increasing since past 15 years.

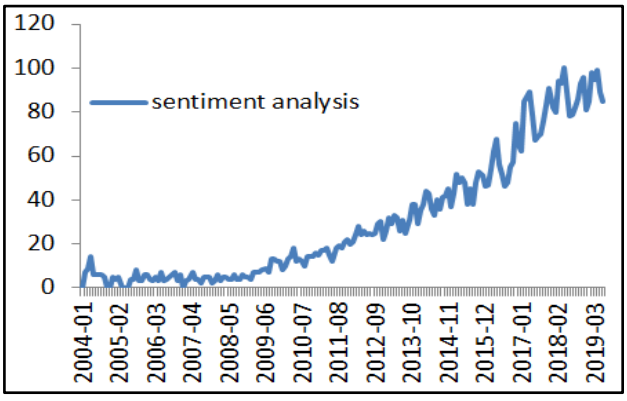


Figure 1.4 Evolution of trends of sentiment analysis

[Source: https://trends.google.com/trends/]

The sentiment analysis of this amount of data when aggregated and analyzed can reveal significant insights that help the users make faster and more informed decisions.

1.3 Sentiment Analysis Applications

The applications of sentiment analysis are endless. Some of these applications are discussed as follows.

a) **Purchasing product or service:** In the age of advertisement, it becomes difficult for people to take a decision while purchasing a product or service. With the help of sentiment analysis technique, people can easily evaluate the opinions of others about any product or service and can compare the competing brands. Today people don't want to rely on external consultant. The sentiment analysis process extracts people opinion form the huge collection of unstructured content from the Internet, analyzes it and then give it to people in highly structured and understandable manner (Rahmath, 2014).

b) **Recruitment:** Sentiment analysis of social media as a recruitment tool provides a way to companies to directly connect to high-caliber employees. The exponential growth of social media sites provides unprecedented access to huge numbers of people who have posted their career experience and other information which is required by the recruiters. Also, the companies can track sentiments among the employees in regard to the organization through social media listening. For example, LinkedIn is a dedicated business networking site used for professional networking. Users can find jobs, people and business opportunities recommended by someone in one's contact network.

c) **Quality improvement in product or service**: By sentiment analysis, manufactures can get the favorable and unfavorable opinions of customers about their products or services. Therefore, they can improve the quality of their products or services. They can also get the online product reviews from websites such as Amazon.com, Zopper.com and rediff.com (Rahmath, 2014).

d) **Policy making:** Sentiment analysis helps the policy makers in making some policy by taking citizen's views and this information can be utilized in creating new citizen friendly policy. Also, the current attitude of public towards some new

government policy can be analyzed by performing sentiment analysis of opinions shared by people on online sites (Rahmath, 2014).

e) **Fighting riots:** Sentiment analysis of social media conversation of people can help to understand the social behavior of people related to sensitive communal issues. It can help us to answer important questions like, Are people talking about riots and religion sentiments that provoke riots over social media; Is there is correlation between sentiments of these conversations with the ground truth information related to riots or communal clashes in the region. These incidents are related with emotions and sentiments of the people. For example, London Metropolitan Police (MET) and the Greater Manchester Police (GMP) analyzed Twitter for crisis communication during the riots in August 2011 (Denef et al., 2013). Also from UK experience, it was found that social media not only plays a vital role in supporting, inciting or encouraging riot but plays a vital role in cleanup operations during the riots.

f) **Enhancement in teaching and learning**: SA plays an important role in many fields including education, where student feedback is essential to assess the effectiveness of learning technologies. Many Universities obtain such feedback via a student response system (SRS) during or at the end of a course to analyze the teacher's performance. Student feedback about teacher performance, the learning experience, and other course attributes can be gathered through social media. Students also comment about their educational experiences in blogs, online forums such as College Confidential (www.collegeconfidential.com), and teacher review sites such as Rate My Professors (www.ratemyprofessors.com). This feedback not only yields useful insights for University administrators and instructors but plays a key role in influencing student decisions on which Universities to attend or courses to take. University administrators also can use this information to integrate some policies, practices and technologies into their operational infrastructure (Manson et al., 2006).

g) **Better policing**: Social media sites offer police departments a way to listen to their citizens and hear what is being said about the department, crime, the quality of life, and events. They also offer the department the ability to shape the

conversation and sentiment analysis can be used to predict it. In London, England, during the G20 protests in April 2009, journalists used Twitter to report what was happening among the crowd. Later that year at an English Defence League protest in Birmingham, the police used Twitter to talk to protesters and point them to the department's Web site and YouTube sites. Those sites featured officers telling the protesters the tactics the police would be using and informing the protesters where they could peacefully protest (Stevens, 2019).

h) **Fighting terrorism**: Today, about 90% of the organized terrorism on the Internet is being carried out through social media. By using these social media sites, the organizations are able to be active in recruiting new friends without geographical limitations. The social media is enabling the terror organizations to take initiatives by making "friend" requests, uploading video clips etc. Aside from recruitment, Facebook is being used by these organizations to gather military and political intelligence (Angelica, 2019).

i) **Reduction in unemployment**: Social media and online user-generated content can be used to enrich the understanding of the changing job conditions by analyzing the moods and topics present in unemployment-related conversations. Hence, sentiment analysis of it can be beneficial for the reduction of unemployment.

j) **Marketing research**: Sentiment analysis techniques can be used in marketing research as these techniques help in analyzing the latest trend used by consumers about some product or services (Parihar, 2012).

k) **Opinion spam detection:** Spams are becoming serious threat for the users of online social networks especially for the ones like of Twitter. As every person shares his views and thoughts on Internet, therefore possibility of spam content on the web has been increased. Spammers publish spurious reviews to promote or demote target online store, including users to buy or not to buy something from particular store to mislead the people. Thus, sentiment analysis can help in classifying the Internet content into "spam" content and "not spam" content (Rahmath, 2014).

The applications of sentiment analysis are endless. In the next section, challenges which arise performing sentiment analysis are discussed along with examples.

1.4 Challenges of Sentiment Analysis

There arise many challenges while performing sentiment analysis. Some of these challenges are discussed as follows (Rahmath, 2014).

a) **Noisy data**: The web content available is very noisy. In today's era of 140 characters texting, people use various abbreviations, slangs, emoticons in normal text for their ease which makes the analysis more complex and difficult. For example, "mvie ws awsummm :D.". The web content reports a large number of spelling variations for the same word, e.g., a word awesome can be found in various forms as- "awsum, awssuummm, awsome" the repetition of the characters can be in any combination. Mostly people often improvise words or use phrases to mean things that were not originally intended. For example, they use improvised words like "OMG" as an exclamation or abbreviated words like "till" to mean "until". But, the existing systems often rely on pre-defined vocabularies, which fail to capture these improvised and abbreviated words. Thus, data available on web presents substantial challenges for sentiment analysis.

b) **Unstructured data**: The rapid growth in web-based activities has led to generation of huge amount of unstructured data which accounts for over 80% of the information. In fact, most individuals and organizations conduct their lives around unstructured data. The web contains data from different sources varying from books, journals, web documents, health records, company's 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.

c) **Sarcasm identification**: It is defined as a sharp, bitter, or cutting expression or remark; a bitter jibe or taunt usually conveyed through irony or understatement. It's a hard task for human beings to interpret sarcasm, making a machine able to understand same is a more difficult task. For example, consider the sentence, "Not

all men are annoying. Some are dead." in which the word "dead" represents sarcasm which is quiet difficult to identify.

d) **Implicit opinion**: Sentiment in a sentence can be implicit or explicit. In case of implicit sentiment, there is no sentiment bearing word present in sentence. But, in case of explicit sentiment, the words present in sentence convey the sentiment. It is difficult to analyze the implicit sentiment in a sentence. For example, the sentence "We had a wonderful time" represents explicit sentiment in this sentence as positive. And the sentence "One should question the stability of mind of the writer who wrote this book" represents the implicit sentiment in this sentence as negative.

e) **Comparative sentences**: A comparative sentence specifies a relation on the basis of similarities or differences of more than one object. The order of words in comparative sentences marks the differences in the determination of the sentiment. For example, "Laptop X is better than Laptop Y" specifies a totally opposite sentiment from "Laptop Y is better than Laptop X".

f) **Multilingual sentiment analysis**: People use different languages while sharing their opinions or views about product, services etc. But, till now, researchers have focused mostly on English due to availability of the lexicon resources and manually labeled corpus for English language. So, there is a need to perform sentiment analysis of Hindi language.

g) **Coreference resolution:** Coreference resolution is the problem of identifying what a pronoun, or a noun phrase refers to. For example, "We watched the movie and went to dinner; it was awful". What does "It" refer to? Coreference resolution may be useful for the feature/aspect based sentiment analysis.

h) **Domain specific**: When disambiguated words are used in different domains, their meaning changes because sentiment depends on the context used in text.

For example, consider the following sentences given in (1.3), (1.4) and (1.5).

"The movie was long." (1.3)

"Lecture was long." (1.4)

"Battery life of Samsung galaxy-2 is long." (1.5)

In all the above examples, meaning of long is same indicating the duration or passage of time. In (1.3) and (1.4) "long" indicates boredness hence a negative expression whereas in (1.5), "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 usages depending on the context. So, it becomes important to detect the context to find the subjective information in a text.

Due to these challenges discussed above, sentiment analysis becomes a difficult process. Although, researchers all over the world are working to handle these issues using different approaches and techniques.

1.5 Sentiment Analysis of Social Media for Hindi Language: The Research Motivation

Hindi is the official language of India belonging to the family of Aryan languages. It is the 4th most spoken language with 310 million speakers across the world which is 4.45% of the world population. In India, Hindi is spoken by a total of 422 million speakers, it's about 41% of total population of India. Therefore, there is a need to perform sentiment analysis in Hindi language so that the opinions of users in Hindi can be easily classified. Our work presented in this thesis is a foray into sentiment analysis for Hindi. Hindi is morphologically rich and is a free order language as compared to English. Usage of Hindi content on the web is in the growing phase, which adds complexity while handling the user generated content. The government of India is promoting the Hindi language offline as well as online by implementing various schemes such as preparation of dictionaries, correspondence course for non-Hindi speaking states, extension programmes which include the programme of neo Hindi writers' scheme, students' study tour, free distribution of Hindi books to Institutions located in non-Hindi speaking areas, book exhibition-cum-sale and scheme of awards to Hindi writers of non-Hindi speaking States. Therefore, there is need to develop a system which can perform sentiment analysis of text shared by people on social media, blogs, discussion forums etc. in Hindi language.

It is well-known that different languages have their own unique ways of expression. All the languages differ generally in their "surface structures" but they all share a common

"deep structure" (Alansary, S., 2012). The basic difference between English and Indian languages is the language structure. For example, English has an SVO (Subject Verb Object) structure, while Hindi follows an SOV (Subject Object Verb) structure. This basic structural difference between English and Indian languages has consequences in deciding the polarity of a text. The same set of words with slight variations and changes in the word order affect the polarity of the words in the text. Therefore, a deeper linguistic analysis is required while dealing with the Indian languages to perform SA. For example, consider the following sentences representing the difference between language structure of English and Hindi.

English: Peter is playing cricket. S V O Hindi: पीटर क्रिकेट खेल रहा है। S O V

The above sentences clearly indicate that English sentences follow the SVO word order, while the Indian language sentences don't follow any word order. The freely word order nature of sentences of Indian language makes the pre-processing difficult. Despite of language structure, there are some other differences between English and Hindi language which make the SA process difficult. These differences (Arora 2013) are discussed as follows.

a) **Null-subject divergence:** A null subject language in linguistic topology is a language in which grammar permits an independent clause known as "null subject" to lack an explicit subject. Some of the null subject Indian languages are Hindi, Tamil, and Telugu etc. whereas English obligatorily requires a subject. Due to this null-subject divergence, SA process becomes difficult. For example, consider the following sentence which represents the null subject divergence between English and Hindi.

English: Long ago, there was a king. Hindi: बहुत पहले एक राजा था। *Long ago one king was*

b) **Handling spelling variations:** In case of Indian languages, the same word with same meaning can occur with different spellings, so it's quite complex to have all the occurrences of such words in a lexicon and even while training a model it's quite complex to handle all the spelling variations. For example, consider the following sentence which shows that the word 'costly' can be written in Hindi with different spelling variations.

English: This phone is very costly. Hindi: यह फोन बहुत मंहगा/महंगा/महँगा है।

c) **Morphological variations:** Handling the morphological variations is also a big challenge for Indian languages. Indian languages are morphologically rich which means that lots of information is fused in the words as compared to the English language where another word is added for the extra information. Indian languages carry the inflection which provides 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.

English: The boys are playing. Hindi: लड़के खेल रहे हैं।
English: The boys killed Ram. Hindi: लड़कों ने राम को मार दिया।

For example, in the above sentences, the same word 'boys' in English has been used for both the morphologically inflected hindi words लड़के *ladake* 'boys' and लड़कों *ladakon* 'boys' which have the same root word लड़का *ladaka* 'boy'.

d) **Paired words:** Sometimes paired words are used in Indian language context. These paired words can be combination of two different opposite, meaningful and meaningless words. For example, the word 'tease' in the following sentence is specified by combining two meaningful ('छेड़') and meaningless ('छाड़') words in Hindi.

English: Karan teased Sangeeta. Hindi: करण ने संगीता के साथ छेड़-छाड़ की।

e) **POS divergence:** As in case of SA, mostly adjectives consist of sentiment in a text. However, sometimes the POS of a word get changed while performing its translation from English to target language. Consider the following sentences in which the words 'wide-eyed' and 'communicative' acting as adjectives in English become adverbs or verbs after translation into Hindi language.

English: The children watched in wide-eyed amazement. Hindi: बच्चे आश्चर्य से आँखे फ़ाड़े देख रहे थे।
English: He was in a bad mood at breakfast and wasn't very communicative. Hindi: नाश्ते के समय वह बुरे मूड में था और ज़्यादा बात चीत नहीं कर रहा था।

These different structural and grammatical challenges of Indian languages make the SA task harder. To understand these rich variations of attributes of the Indian context words, the system needs robust morph analyzer so that the right sense of the word can be mined. Notably, efficient linguistic resources are required to pre-process Indian language context and to take care of spelling and multilingual issues. Also, main challenge is that the text-based information retrieval systems require huge annotation and increase the semantic hole between the view of the user and system understanding (Raghuwanshi and Tyagi, 2019).

1.6 Research Gaps

Some of the research gaps identified for proposed research work are presented as follows.

- India is a land of many languages and usage of these languages reflect in conversations of Indian people over social media. Mostly, Indian people express their sentiments on social media in transliterated form of their language by using QWERTY keyboard. Therefore, direct processing of Indian languages for performing sentiment analysis is not possible and pre-processing of data is an important challenge.

- There is no study to highlight the percentage of people using Hindi language in their conversations on social media which can be helpful to understand the role of social media in Indian context.
- A comprehensive and detailed survey of sentiment analysis for Indian languages is not available so there is need of performing the in-depth analysis of sentiment analysis for Indian languages to get the knowledge about the existing lexical resources and approaches used by researchers.
- Very limited research work has been performed for sentiment analysis of Hindi languages. As there is no labeled data and lexical resources available for performing the sentiment analysis of social media content in Hindi language, there is a need to create these resources for Hindi language.
- Most of the existing work for sentiment analysis has been performed at sentence and document level and a very limited research work has been performed at aspect/feature level.
- Mainly the research work on sentiment analysis has been done using machine learning techniques and there is need to perform the experimentation of sentiment analysis using deep learning techniques to improve the accuracy of the system.
- There is a need of real time system to track the social media conversations for Hindi language and to analyze the different moods of people for its possible use in multiple applications like recruitment, unemployment, marketing research, fighting riots etc.

From the aforementioned research gaps, the following research objectives have been framed.

1.7 Research Objectives

- To study existing sentiment analysis techniques, their solutions for social media and to analyze their adaptability to Indian environment.
- To develop the pre-processing module to handle issues of transliteration and abbreviated text for Hindi language.

- To create labeled data of tweets for Hindi language to build the machine learning module.
- To propose, implement and validate sentiment analysis system for social media by considering Hindi contents to classify the sentiments of users into different sentiment groups.
- To develop a web based interface for monitoring the real time social media traffic.

1.8 Research Methodology

The following methodology has been followed to achieve the objectives of the research work.

To achieve first objective:

- A detailed analysis of Indian language families has been performed to get the knowledge about the research work done in the field of sentiment analysis.
- A review methodology has been followed by identifying the related research studies from the renowned electronic databases as well as the topmost conferences related to the area.
- The impact of research studies has been analyzed by considering the citations of research studies undertaken.
- The annotated lexical and linguistic resources available for different Indian languages have been identified.
- The percentage of status of research works for different Indian languages, techniques, domains, sentiment classes and sentiment levels has been analyzed.

To achieve second objective:

- The issue of transliteration of Hindi text has been handled by integrating the API of Google transliteration with the developed system.
- The abbreviated content has been handled using mapping dictionary of Hindi words.

To achieve third objective:

- The corpra of tweets and movie reviews have been extracted from Twitter and movie reviews websites respectively.
- The extracted corpus has been annotated manually by three native speakers of Hindi and data set has been validated using the kappa statistic measure.
- A corpus of students' feedback has also been collected to present an application of sentiment analysis in the education domain.

To achieve fourth objective:

- The system has been trained on different machine learning and deep learning models on the labeled corpus for sentiment analysis of Hindi content at sentence and document level in Python.
- An aspect-based sentiment analysis system has been proposed to perform sentiment analysis at aspect-level.
- The accuracy of the system has been compared with existing state-of-the-art techniques.

To achieve fifth objective:

- A web-based interface for sentiment analysis of Hindi content on aspect-level, sentence level and document level has been developed which is available at http://www.hindisenti.com/.
- The web-based system is able to extract Hindi tweets based on user-defined Hashtag and analyze the polarity of individual tweets as well as overall polarity of Hashtag.
- The polarity of extracted tweets has been shown by plotting bar charts in the form of three bars, i.e., positive, negative and neutral.

In addition to these objectives, this thesis work also includes a case study as an application to analyze the students' feedback to improve the teaching and learning abilities of teachers and students respectively.

1.9 Research Contributions

Our main contribution in this thesis is the development of sentiment analysis system for Hindi language. The major contributions of the thesis are as follows.

- A comprehensive literature review has been performed to know the state-of-the-art of sentiment analysis over Indian languages and social media. It has been observed that 15 Indian languages have been explored till now with the major research work for Hindi language.
- The available annotated datasets, lexical and linguistic resources have been identified along with the research work on different sentiment levels, domains, sentiment classes and different performance measures used to validate the sentiment analysis systems.
- The corpra of Hindi tweets and movie reviews has been extracted from Twitter and movie reviews websites respectively using the libraries of Python.
- The extracted corpus has been annotated by three native speakers of Hindi and validation of corpus has been performed using kappa statistic measure.
- The pre-processing of corpus has been performed using different pre-processing techniques, tools and resources such as Google Transliterator and mapping dictionary to handle the issue of transliteration and abbreviated content.
- The sentence level sentiment analysis system takes the sentence as input and predicts its polarity in the form of positive, negative and neutral.
- The different machine learning and deep learning models have been trained on the corpus using different feature extraction techniques to perform sentiment analysis at sentence and document level.
- Twitter sentiment analysis system works at document level. It first extracts the tweets on the basis of user-defined hashtag and identifies the polarity of each tweet and overall polarity of hashtag.
- An aspect-based sentiment analysis system takes the Hindi sentence as input. Then, it extracts the nouns and sentiment bearing words from the sentence with

the use of Hindi dependency parser. The system identifies the polarity score of sentiment words and assigns them to the corresponding aspects.

- A case study of students' feedback has been used to present the sentiment analysis system as an application to improve the teaching and learning activities has also been presented and different sentiments such as happy, sad, anger, joy have been analyzed.
- A web-based sentiment analysis system has been developed for Hindi language to perform sentiment analysis at different levels such as aspect, sentence and document level and is available at http://www.hindisenti.com/.
- The developed web-based sentiment analysis can be integrated with other applications for the benefits of society.

1.10 Thesis Organization

The thesis has been structured into 7 chapters. The brief overview about these chapters is given as follows.

Chapter 1 introduces the need of sentiment analysis covering the different levels of sentiment analysis such as sentence, document and aspect/feature. This chapter also includes the discussion about the growth of web content on social media platforms such as Twitter, Facebook, LinkedIn, etc. which acts as huge repository to perform sentiment analysis. The year-wise evolution of sentiment analysis has been presented in graphical form which represents the exponential growth of sentiment analysis in the recent years. This chapter further covers the different applications of sentiment analysis in different domains such as recruitment, education, marketing, policy making, unemployment, fighting riots, terrorism, and education, etc.

The different challenges like noisiness of data, unstructured data, word sense disambiguation, sarcasm detection, implicit opinion, multilingual sentiment analysis, etc. which arise while performing sentiment analysis on natural language text have been discussed in this chapter. The general process of sentiment analysis which consists of five phases such as data collection, data preprocessing, sentiment detection, sentiment classification and presentation of output is also presented in this chapter. Additionally,

this chapter discusses about the motivation behind the research conducted to perform sentiment analysis for Hindi language including the divergence of Hindi from English language as well as the challenges arising while performing sentiment analysis for Hindi language. The last section of the chapter presents the research objectives framed based on the research gaps identified along with thesis contribution and organization.

Chapter 2 includes a systematic review of sentiment analysis over Indian languages and social media. This chapter discusses about the evolution of sentiment analysis. A review methodology has been followed to conduct this systematic review. The review presented in this chapter has been performed by identifying the related research studies from the renowned electronic databases as well as the topmost conferences related to the area. After this, to narrow down the count of selected studies, inclusion and exclusion criteria have been followed and the final research studies have been selected based on formulation of research questions and results have been compiled after performing in-depth analysis.

This chapter also includes the background detailed description about the Indian language families and evolution of Indian languages in the field of sentiment analysis. The extracted outcomes present the year-wise publications and their status from different sources such as journals, conferences, workshops, thesis and online reports. The impact of research studies has been analyzed by considering the citations of research studies undertaken. This chapter presents the detailed description about the online availability of labeled datasets and pre-processing linguistic resources such as shallow parser, POS tagger, dependency parser, morphological analyzer etc. for different Indian languages. The detail about the SentiWordNet (a lexical resource) has also been included covering the discussion about its generation for different Indian languages using different approaches. The different sentiment analysis techniques such as lexicon-based, machine learning and deep learning are discussed in detail in this chapter. The state-of-the-art works based on different parameters such as approach used, corpus type, corpus size, algorithm/technology used and evaluation measures used by researchers to validate their systems are summarized in tabular form. The percentage of status of research works for different Indian languages, techniques, domains, sentiment classes and sentiment levels

has been presented in graphical form which will help the researchers to know about the existing work carried out in the field of sentiment analysis for different Indian languages.

Chapter 3 presents the detailed description about the corpus used for experimentation. Two corpra have been created for the experimentation. The corpra of Hindi movie reviews and tweets have been extracted from Movie Reviews websites and Twitter respectively. This corpus has been annotated manually by three native speakers of Hindi into three classes such as positive, negative and neutral. The annotated corpus has been validated using the kappa statistic measure.

Generally, sentiment analysis is performed at three levels such as document, sentence and aspect level. This thesis includes the performance of sentiment analysis at all the three levels. This chapter presents the architecture of sentiment analysis system which performs the sentiment analysis for Hindi language at sentence or document level. The detailed methodology of sentiment analysis to perform at aspect level is presented in the next chapter. This chapter includes the implementation of different machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Decision Tree, k-Nearest Neighbor (k-NN), Random Forest, AdaBoost, Gradient Boosting and CNN. The CNN model experimented with different parameter settings to analyze its performance has also been presented in detail in this chapter. The results given by the developed sentiment analysis systems have been analyzed using different performance measures such as accuracy, precision, recall, F-measure. The performance of the system has also been compared with the results of the existing systems for Hindi language.

Chapter 4 discusses the development of sentiment analysis system which performs sentiment analysis for Hindi language at aspect level. The aspect-based sentiment analysis system follows the lexicon based approach to perform sentiment analysis. This chapter includes the brief introduction about the linguistic resource like Hindi Dependency Parser and lexical resource SentiWordNet as both of these resources contribute to the development of proposed aspect-based sentiment analysis system.

This chapter presents the example sentences in which sentiment bearing words appear in the form of different POS such as noun, adverb, and verbs along with adjectives which

play a vital role to determine the sentiment polarity. This chapter discusses about the other sources such as a mapping dictionary which maps the Hinglish words and Google API which helps in transliteration of Hindi content. The brief description about the different phases of aspect-based sentiment analysis system has also been presented in this chapter. The proposed system generates a dependency graph from the input sentence and assigns the identified aspects to sentiment bearing words based on the computation of distance. Finally, the sentiment polarity is determined from sentiment score. The performance of the system has also been compared with the results of the existing systems for Hindi language.

Chapter 5 includes the case study taken in the field of education domain to perform sentiment analysis for students' feedback. This chapter presents the sentiment analysis system developed to improve the teaching and learning abilities of teachers and students in education domain. The description about the corpus which has been used as a case study has also been provided in this chapter. The developed system takes the students' feedback in the form of surveys, rating, comments, and pre-process it using different techniques such as tokenization, lowercasing, normalization, stemming, and removal of irrelevant content and transliteration of content. The system classifies the comments into the different emotions such as anger, anticipation, disgust, joy, fear, sadness, surprise and trust as well as sentiment such as positive and negative. Also, two new parameters such as satisfaction and dissatisfaction have also been derived from the existing emotions to analyze the satisfaction level of students.

This chapter also presents the temporal sentiment and emotion analysis of ratings and comments given by students about teacher's performance in lectures and lab sessions of a course. The system has been validated by comparing the direct learning of students which is taken in the form of performance of students in quizzes, tutorials, mid-semester and end-semester examinations with indirect learning taken in the form of surveys and ratings. The analysis presented in this chapter bridges the gap between students and teachers in order to improve the teaching and learning abilities of teachers and students respectively.

Chapter 6 presents the online web-based interface developed for the sentiment analysis of Hindi language. This web-based system is able to perform sentiment analysis of Hindi content at aspect, sentence and document level. For aspect level, the system takes the Hindi sentence as input in text box and shows the dependency graph of that sentence. The aspect-based sentiment analysis system's interface also shows the extracted aspects and sentiment bearing words along with their assignment. For sentence and document level, different machine learning and deep learning models have been trained and the algorithm with highest accuracy has been considered to assign final polarity to the testing sentence. Also, the system is able to extract the tweets about any user-defined Hashtag and analyze it by plotting the bar charts consisting of bars for positive, negative and neutral polarity.

Chapter 7 concludes the research work presented in this thesis and presents the future implications. It has been concluded that the results given by the system are very promising which can be used for the benefits of the society. In future, the accuracy of the system can be improved by extending the dataset as well as re-training of the models by including the predicted dataset by the system itself in order to improve the training of system. Also, the coverage of Hindi SentiWordNet and mapping dictionary can be extended to improve the accuracy. Moreover, in future, system can be improved by integrating the solutions of other challenges such as sarcasm detection, word sense disambiguation which arise while performing sentiment analysis of Hindi text.

Chapter Summary

This chapter presents the description about the sentiment analysis along with sentiment levels, i.e., aspect/feature, sentence and document level. The general process of sentiment analysis is explained in this chapter. The need of sentiment analysis in today's life and role of social media platforms such as Twitter, Facebook etc. for sentiment analysis is also discussed in this chapter. This chapter covers the various applications of sentiment analysis such as recruitment, education, marketing, policy making, unemployment, fighting riots, terrorism, and education, etc. The different challenges which arise while performing sentiment analysis are covered in this chapter. The research motivation for performing the sentiment analysis of social media for Hindi language is explained in detail along with the differences in language structure of English and Hindi. The last section of this chapter presents the research objectives framed based on the research gaps identified along with thesis contribution and organization.

Chapter 2

Literature Review

In this chapter, a system review of sentiment analysis has been presented. This chapter includes the evolution of sentiment analysis and the review methodology followed to perform the detailed and comprehensive survey has also been presented. The survey has been performed by frming different research questions and also an inclusion and exclusion criteria has been followed to include only the relevant studies. This chapter covers the evolution of SA for Indian languages along with its differences from English language. The status of research work on the basis of different Indian languages, SA techniques, levels, classes and domains has also been presented in this chapter.

2.1 Evolution of Sentiment Analysis

The research works on sentiment analysis appeared as early as in 2000 (Pang et al., 2008). Earlier, the term sentiment analysis was primarily used on written paper documents. Today, however, sentiment analysis is widely used to mine subjective information from content on the Internet, including texts, tweets, blogs, social media, news articles, reviews, and comments. This is done using a variety of different techniques, including NLP, statistics, and machine learning methods. Organizations then use the information mined to identify new opportunities and better target their message toward their target demographics. The Obama Administration even uses sentiment analysis to predict public response to its policy announcements. The analysis of sentiments is generally performed at three levels aspect, sentence and document. And the sentiments are classified into mainly three classes such as positive, negative and neutral. The detailed description about the status of research work in the field of sentiment analysis is presented in the following sections.

The next section discusses about the review methodology that has been followed to carry out this systematic review.

2.2 Review Methodology Followed

The steps followed to conduct this review on SA for Indian languages are given as follows.

2.2.1 Development of Review Protocol

This systematic review has been conducted by identifying the related research studies from the renowned electronic databases as well as the topmost conferences related to the area. After this, to narrow down the count of selected studies, inclusion and exclusion criteria have been followed. Then, final research studies have been selected based on the formulation of research questions and results have been compiled after performing in-depth analysis.

2.2.2 Research Questions

The systematic review presented in this chapter focuses on identifying and analyzing the existing literature survey describing different SA methods and techniques used for different Indian languages. It also finds the different lexical and lexicon resources as well as tools which are used by researchers to perform SA for Indian languages.

A set of research questions (listed in Table 2.1) have been formulated in order to conduct a systematic review in an efficient way.

Table 2.1: Research questions for systematic literature review

Research question	**Main motivation**
RQ1: What is the year-wise status and which are the databases of publications since the inception of SA for Indian languages?	Identify the time frame and sources of publications in which the relevant large number of research studies have been published.
RQ2: What is the impact of research studies considered?	Identify the research studies using the citation information to include only the relevant work in this area.

Research question	Main motivation
RQ3: Which Indian languages have been mostly explored until now?	Explore the Indian language families and identify languages for which the majority of the SA research work has been carried out.
RQ4: Which sentiment analysis techniques have been used mostly?	Identify the most commonly used SA techniques.
RQ5: Which annotated dataset, linguistic and lexical resources are available online for Indian languages and which domains have been considered?	The online availability of annotated datasets, linguistic and lexical resources for SA are suggestive of ease of use and development of resources for other Indian languages. Also, identify the domains such as products, movies or social media platforms, etc. in which corpora are available for Indian languages to perform SA.
RQ6: What are the different factors considered while performing SA?	Identify the factors such as sentiment levels (such as aspect, sentence or document) and classes like positive, negative or neutral for which SA is performed.
RQ7: Whether any online tools are available for SA of Indian languages?	Explore the availability of different SA tools that perform online SA for Indian languages.
RQ8: What is the future indirections identified from the literature review?	Identify the unexplored relevant research visions.

2.2.3 Sources of Information

A proper set of e-databases were chosen before starting the search process to identify the relevant research articles. The electronic databases that were selected for identifying the research studies are Google Scholar (www.scholar.google.co.in/), Science Direct (www.sciencedirect.com), ACM Digital Library (www.acm.org/dl) and IEEE Xplore (www. ieeexplore.ieee.org). Most of the papers were published in topmost conferences related to NLP and linguistics, and also Google Scholar covers almost all the papers. The

papers that were redundant on Science Direct, ACM Digital Library, and IEEE Xplore have been excluded before final selection of research articles.

2.2.4 Inclusion and Exclusion Criteria

This systematic survey has been conducted by following the guidelines given by Kitchenham and Charters (2007). A systematical keyword-based advanced search has been followed to retrieve the significant research studies from the e-databases in the time frame 2010-2019 as shown in Table 2.2.

Table 2.2: Keyword-based advanced search

Source	**Keyword**	**Publication Type**	**No. of Results**
Google Scholar	sentiment analysis [in/of/for] [Language]	C/J/M	704
	opinion mining [in/of /for] [Language]	C/J/M	56
IEEE Xplore	Abstract: sentiment analysis [in/of/for] [Langauge]	C	62
	Abstract: opinion mining [in/of/for] [Language]	C	16
Science Direct	Abstract, Title, Keywords: sentiment analysis [in/of/for][Language]	J	6
	Abstract, Title, Keywords: opinion mining [in/of/for] [Language]	J	2
ACM Digital Library	Abstract: sentiment analysis [in/of/for] [Language]	C/J	4
	Abstract : opinion mining [in/of/for] [Language]	C	1

**Language: [Bengali, Hindi, Gujarati, Kannada, Konkani, Malayalam, Manipuri, Nepali, Oriya, Punjabi, Tamil, Telugu, Urdu]

This systematical literature review consists of both qualitative and quantitative research studies from 2010–2019 to ensure the completeness of review as an attempt to work on SA for Indian languages was first commenced in 2010. The keywords "sentiment analysis" and "opinion mining" directed to a large number of results as this field is explored for different languages in different domains. The search has been performed in abstract and title using the search string "Sentiment analysis [in, of, for] [language_name]." For example, some research studies have considered the substring "in

Hindi" or "for Hindi" or "of Hindi" in their title. Therefore, search has been performed by taking this into account so that all of the research studies in this field can be included.

The research studies from various conferences, journals, workshops along with masters and Ph.D. thesis have been included by following an exclusion criterion at different stages shown in Figure 2.1.

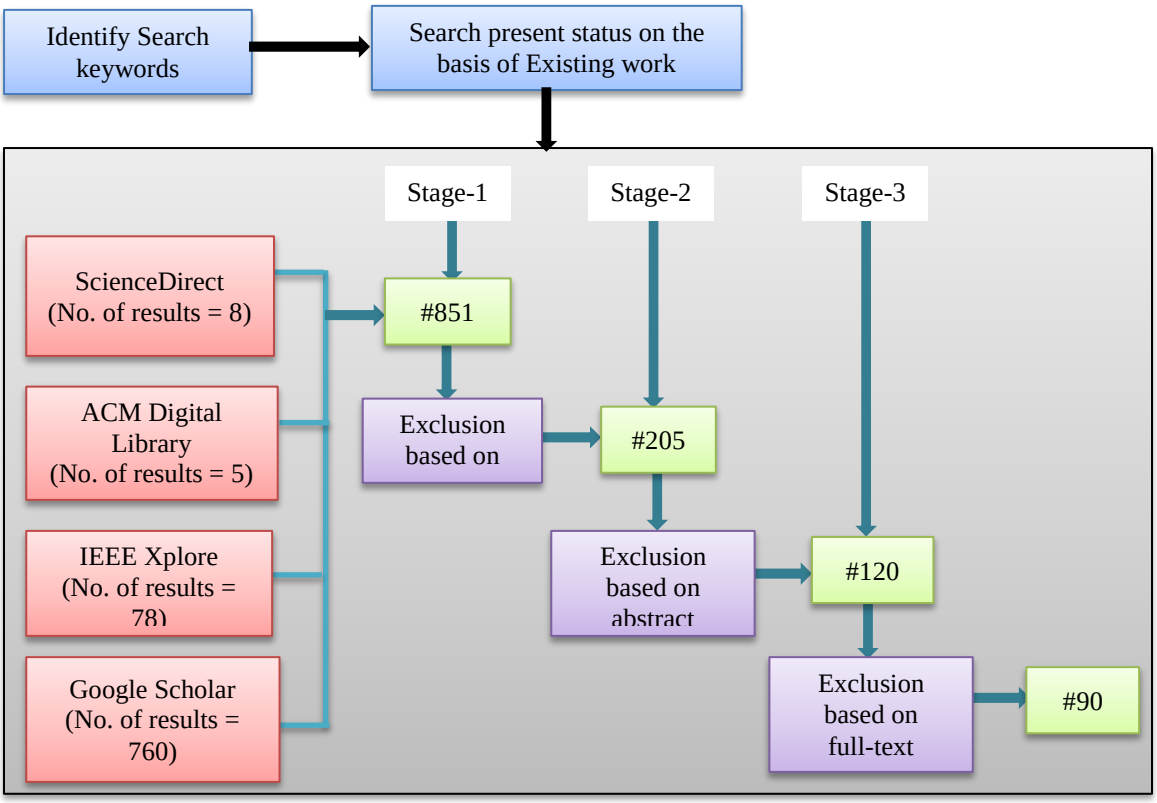


Figure 2.1 Review technique followed

Also, some individual searches have been applied to some conferences and journals related to NLP and linguistics to complete the e-search. Our search returned 851 research studies (shown in Figure 2.1) which were reduced to 205 based on their titles, 120 based on their abstract and 90 on the basis of full-text. After that, these 90 research articles were analyzed in-depth to select a final list of research studies.

2.3 SA for Indian Languages: The Background

2.3.1 Introduction to Indian Language Families

Indian languages belong to several language families and broadly divided into four language families, i.e., Indo Aryan family (Arya), Dravidian family (Dravida), Sino Tibetan family (kirata) and Austroasiatic family (Nishada) as shown in Figure 2.2.

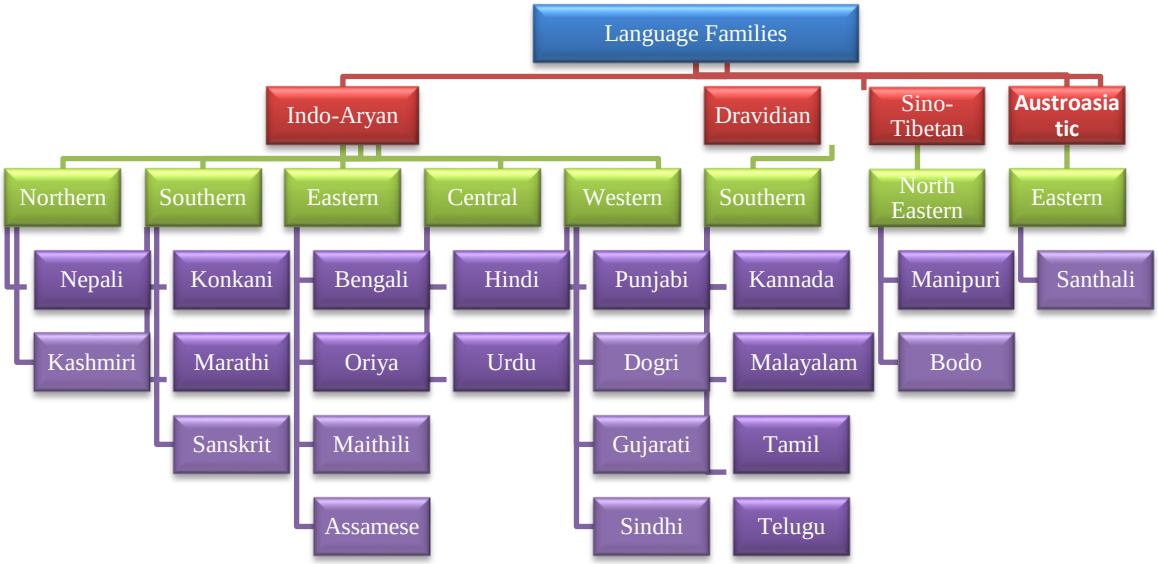


Figure 2.2 Classification of Indian languages

Indo-Aryan language family covers about 74% of the Indian population and 24% of the total Indian population is covered by Dravidian languages. Austroasiatic and Sino-Tibetan languages are the language families' together covering 2% of the population (Ind, 2015). The brief description of these language families is given as follows.

a) **Indo-Aryan language family:** Indo-Aryan language family is part of the Indo-European family of languages and the mostly spoken language family in India. The utmost widely spoken languages of this language family are Hindi, Bengali, Punjabi, Odia (Oriya), Nepali, Konkani, Marathi, Gujarati, Sindhi, Assamese, Dogri, Urdu, Kashmiri and Sanskrit (Ind, 2014).

b) **Dravidian language family:** It is the second-largest language family in Indian language families. This language family is older than the Indo-Aryan language family. The languages of this family are spoken mainly in southern and parts of eastern and central India as well as in parts of north eastern Sri Lanka, Nepal, Pakistan and Bangladesh (Chand, 2016). The major Dravidian languages are Telugu, Tamil, Kannada, and Malayalam.

c) **Sino-Tibetan language family:** The Sino-Tibetan languages are referred as Kiratas in the oldest Sanskrit literature. This language family is also older than the Indo-Aryan language family. These languages have three major sub-divisions such as The Tibeto Himalayan, The North Assamese, and The Assam–Myanomari (Burmese). The main Sino-Tibetan languages are Manipuri and Bodo.

d) **Austroasiatic language family:** The Austric languages are referred as Nisadas in the oldest Sanskrit literature of India and these languages are mainly spoken in the central, eastern and north-eastern India. This ancient language family came into existence before the arrival of Aryans. The most spoken language of this family is Santhali.

Mainly the research work in the field of SA has been done in English as well few other non- English languages such as Arabic, Chinese, etc. A very less contribution exists for Indian Languages such as Hindi, Tamil, Telugu, Bengali, etc. (Kaur and Saini, 2014). The primary reason behind this is the lack of annotated datasets, linguistic as well as lexical resources and tools for Indian languages.

2.3.2 Evolution of Indian Languages for SA

The evolution of Indian languages in the field of sentiment analysis started in 2010 when Joshi et al., (2010) first attempted to work on SA for Indian languages. The authors performed sentiment analysis for Hindi language and later on, researchers started working on different Indian languages such as Bengali, Tamil, Telugu, etc. The year-wise evolution of Indian languages for SA is shown in Figure 2.3.

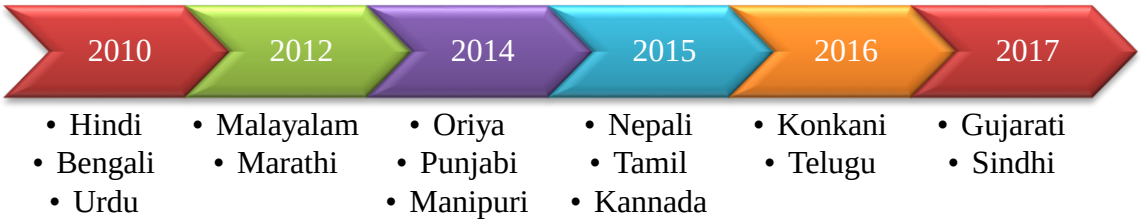


Figure 2.3 Evolution of Indian languages

2.4 Extraction Outcomes

The aim of this work is to identify the available research on SA for Indian languages and is stated in Table 2.1 in the form of research questions. To answer the research question **RQ1**, year-wise status and origin of sources of publications on SA for Indian languages have been explored which are represented in Figures 2.4 and 2.5 respectively.

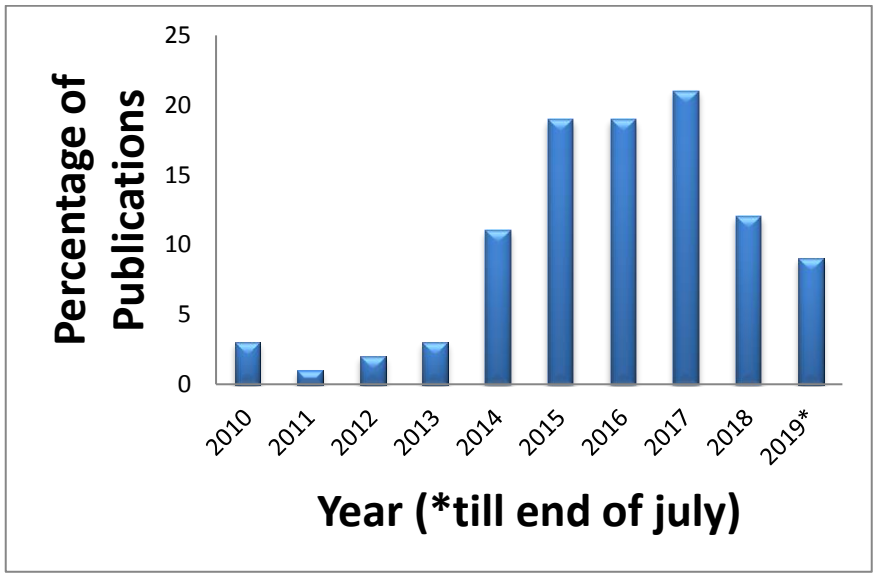


Figure 2.4 Year-wise publications on SA for different Indian languages

The first attempt in this area was commenced in 2010. Therefore, the year-wise status of publications from 2010 to 2019 is depicted in Figure 2.4. From the figure, it has been analyzed that research in this area is continuously growing from the last couple of years. While in-depth analysis, it has also been observed that most of the research articles on Indian sentiment analysis are published in an extensive variety of conference proceedings and journals. The approximate 54% of the research articles are published in conferences,

39% in journals, 4% in workshops, and the remaining 3% are covered by thesis and online reports as shown in Figure 2.5. The highest percentage of research publications came from conferences, followed by journals.

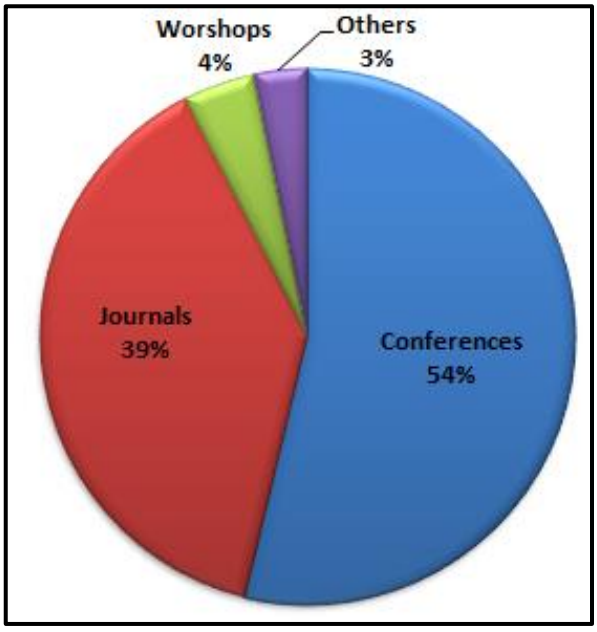


Figure 2.5 Status of publications from different sources

To address the research question **RQ2**, data has been analyzed and is responded through Table 2.3. From this table, it can be observed that the highest cited research study is the work done by Joshi et al., (2010) which has more than a hundred citations. The authors of this research study set the benchmark in this area and other researchers followed the approaches discussed by them to perform SA for other Indian languages.

Table 2.3: Number of publications according to citations

Citation Count	**<=5**	**>5 and <=20**	**>20 and <=100**	**>100**
No of Publications	57	27	8	1

2.5 Preliminaries for SA of Indian languages

2.5.1 Dataset

The first phase to perform SA is the dataset collection. Mainly the social media platforms such as Twitter, blogs, discussion forums and review sites related to products, movies,

and travels have been used to perform SA for Indian languages. Some of the annotated datasets of tweets and reviews are available online for six Indian languages such as Hindi, Bengali, Tamil, Marathi, Telugu, and Urdu as given in Table 2.4.

Table 2.4: Summary about the online available datasets

Sr. No.	**Language**	**Dataset Type**	**Level***	**Dataset details**	**Link**
1.	Hindi (Bakliwal et al., 2012)	Product reviews	S	350 positive, 350 negative	Available on request to author
2.	Hindi (Balamurali et al., 2012)	Travel reviews	S	100 positive and 100 negative	http://www.cfilt.iitb.ac.in/resources/senti/HPLC_tour_downloaderInfo.php
3.	Marathi (Balamurali et al., 2012)	Tourism reviews	S	75 positive, 75 negative	http://www.cfilt.iitb.ac.in/resources/senti/MPLC_tour_downloaderInfo.php
4.	Hindi (Patra et al., 2015)	Tweets	S	168 positive, 559 negative, 494 neutral	http://amitavadas.com/SAIL/index.html
5.	Bengali (Patra et al., 2015)	Tweets	S	277 positive, 354 negative, 368 neutral	http://amitavadas.com/SAIL/index.html
6.	Tamil (Patra et al., 2015)	Tweets	S	387 positive, 316 negative, 400 neutral	http://amitavadas.com/SAIL/index.html
7.	Hindi (Akhtar et al., 2016a)	Movie and Product reviews	A	2,250 positive, 635 negative, 2,241 neutral and 128 conflict	https://www.iitp.ac.in/~ai-nlp-ml/resources.html
8.	Hindi (Akhtar et al., 2016b)	Movie and Product Reviews	A	2,290 positive, 712 negative, 2,226 neutral and 189 conflict	http://amitavadas.com/SAIL/index.html

Sr. No.	Language	Dataset Type	Level*	Dataset details	Link
9.	Hindi (Akhtar et al., 2016c)	Movie Reviews	S	823 positive, 530 negative, 598 neutral, 201 conflict	https://www.iitp.ac.in/~ai-nlp-ml/resources.html
10.	Telugu (Mukku and Mamidi, 2017b)	Reviews, news websites, Twitter, Facebook posts	S	1489 positive, 1441 negative, 2475 neutral	https://drive.google.com/drive/folders/0B8HHvMMuHYdWdnJZZl9rWkY5bk0
11.	Urdu (Khan et al., 2017)	Tweets	S	535 positive, 464 negative	https://github.com/MuhammadYaseenKhan/Urdu-Sentiment-Dataset

*S- Sentence, A- Aspect

These datasets are annotated into different classes such as positive, negative and neutral. As the majority of research work on SA has been done for the Hindi language therefore, aspect and sentence level annotated datasets are available for the Hindi language across various domains. The information given in Table 2.4 helps in attaining the answer to the research question **RQ5**. This table provides a summary of online available annotated datasets for Indian languages.

2.5.2 Pre-processing Linguistic Resources

Some of the pre-processing resources such as shallow parser, Part Of Speech (POS) tagger, dependency parser and morphological analyzer to perform SA for Indian languages along with their online availability are given in Table 2.5. The brief description of these resources is given as follows.

a) **Shallow parser:** Generally, shallow parser provides the analysis of a sentence in the form of morphological structure, Chunking, POS tagging, etc. Shallow parsers for Indian languages are developed under a consortium project funded by the

Government of India (Sha, 2012). Till now, these are mainly available online for nine Indian languages, i.e., Hindi, Punjabi, Urdu, Tamil, Bengali, Telugu, Kannada, Malayalam and Marathi.

b) **Morphological analyzers:** Morphological analyzers give the root word and other features such as gender, number, tense, etc. Thus, morphological analysis is the process of imparting grammatical information of a word given its suffix. The independent morphological analyzers are available online for five Indian languages, i.e., Hindi, Marathi, Kannada, Telugu and Punjabi. However, one can also use shallow parser to perform morphological analysis.

c) **POS tagger:** POS tagging tagging is also known as grammatical tagging or morphosyntactic annotation which takes place at word level and adds morpho syntactic information next to each word in the corpus (Salama and Alansary, 2016). It is a process of classifying and labeling the words of a sentence according to their POS information which includes nouns, verbs, adjectives, determiners, adverbs, and so on. POS tagger generally indicates the status of the word based on the morphological and/or syntactic properties of a language. POS taggers are independently available online for five Indian languages such as Hindi, Kannada, Telugu, Punjabi, and Urdu. However, one can also use shallow parser to extract the POS tagging information.

d) **Dependency parser:** Dependency parsing is the process of revealing the dependency tree of a sentence through labeled links which represent the dependency relationships between words. Researchers have worked on the creation of dependency parsers for various Indian languages such as Telugu, Tamil, Bengali, etc. but presently dependency parser is available online only for the Hindi language.

e) **Sandhi splitter:** Sandhi splitter is a computational tool which shows all possible splitting of a given string. Currently, Sandhi splitter is available online for the Malayalam language.

For the Sindhi language, an online tool has been available at https://sindhinlp.com/ which provides Sindhi parser, stemming, lemmatization and also performs sentiment analysis at

sentence and aspect level. The research studies have been analyzed and on the basis of that, the answer to the research question **RQ5** has been addressed.

Table 2.5 summarizes the details about the online available pre-processing linguistic resources for different Indian languages.

Table 2.5: Online available pre-processing linguistic resources

Sr. No.	Resource	Languages	Online Link
1.	Shallow Parser	Hindi, Punjabi, Urdu, Tamil, Bengali, Telugu, Kannada, Malayalam and Marathi	http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php
2.	Independent Morphological Analyzers	Hindi, Marathi, Kannada and Telugu	http://ltrc.iiit.ac.in/showfile.php?filename=onlineServices/morph/index.htm
		Punjabi	http://www.learnpunjabi.org/punjabi_mor_ana.asp
3.	Independent POS Tagger	Hindi, Kannada and Telugu	http://sivareddy.in/downloads
		Punjabi	http://punjabi.aglsoft.com/punjabi/?show=tagger
		Urdu	https://tech.cle.org.pk/services/text/pos
4.	Dependency Parser	Hindi	https://bitbucket.org/sivareddyg/hindi-dependency-parser
5.	Sandhi splitter	Malayalam	https://github.com/libindic/sandhi-splitter

2.6 SentiWordNet (SWN): A Lexical Resource for SA

Researchers have either manually constructed or used WordNets to create lexical resources for SA. SWN is such a lexical resource that is mostly used for SA. SWN is the result of annotation of all WordNet synsets on the basis of degrees of polarity, i.e., positivity, negativity, and neutrality (Baccianella et al., 2010). WordNets have also been created for a number of Indian languages. For example, IndoWordNet is a linked structure

of WordNets of all major Indian languages and currently supports 19 Indian languages (Bhattacharyya, 2017).

Generally, WordNet and bi-lingual dictionary-based approaches are followed for the creation of SWN(s). In WordNet-based approach, SWN for the target language is developed by mapping the synsets of English SWN along with polarity scores into target language synsets using IndoWordNet. In bi-lingual dictionary-based approach, the polarity scores are extracted from English SWN and assigned to the words of the target language. Das and Bandyopadhyay (2010c) proposed three other approaches such as corpus-based, antonym based and gaming technology to increase the coverage of developed SWN. Presently, SWN(s) are available online for three languages, i.e., Hindi, Bengali, and Telugu at http://www.amitavadas.com/sentiWordNet.php.

Table 2.6 provides the answer for the research question **RQ5** as it gives the information about the online available SWN(s) for different Indian languages along with their development approach and count of synsets.

Table 2.6: Online available SWN(s) for different Indian languages

Language Family	**Language**	**Reference**	**Approach**	**Lexical Resources**	**Synsets**
Indo-Aryan	Hindi	Bakliwal et al., (2012)	WordNet	IndoWordNet (Bhattacharyya 2017), English SWN (Esuli and Sebastiani 2007)	16000
		Das and Bandyopadhyay (2010c)	Dictionary	SHABADKOSH, Shabdanjali	22,708
	Bengali	Das and Bandyopadhyay (2010b)	Dictionary	Samsad Bengali-English dictionary	34,117
	Konkani	Fondekar et al. (2016)	WordNet	IndoWordNet (Bhattacharyya 2017), Hindi SWN	368
	Nepali	Gupta and Bal	Dictionary	English SWN (Esuli	629,

Language Family	Language	Reference	Approach	Lexical Resources	Synsets
		(2015)		and Sebastiani 2007), English-Nepali Dictionary	930
	Punjabi	Kaur and Gupta (2014b)	Dictionary	Hindi SWN	7860
	Gujarati	Gohil and Patel (2019)	WordNet	IndoWordNet (Bhattacharyya 2017), Hindi SWN	6,076
	Odia	Mohanty et al., (2017)	WordNet	Odia, Bengali, Telugu and Tamil WordNet, Bengali, Telugu and Tamil SWN	13, 917
Dravidian	Telugu	Das and Bandyopadhyay (2010c)	Dictionary	Charles Philip Brown English-Telugu Dictionary, Aksharamala English-Telugu Dictionary, English-Telugu Dictionary	30,889
	Kannada	Deepamala and Kumar (2015)	Manually	Hindi SWN	5043
	Malayalam	Anagha et al., (2014)	Dictionary	Hindi SWN	2000

2.7 SA Techniques and Evaluation Measures

2.7.1 SA Techniques

From the comprehensive survey, it has been observed that SA techniques can be classified into three categories such as lexicon-based, machine learning and deep learning as shown in Figure 2.6. The brief description of these techniques is given as follows.

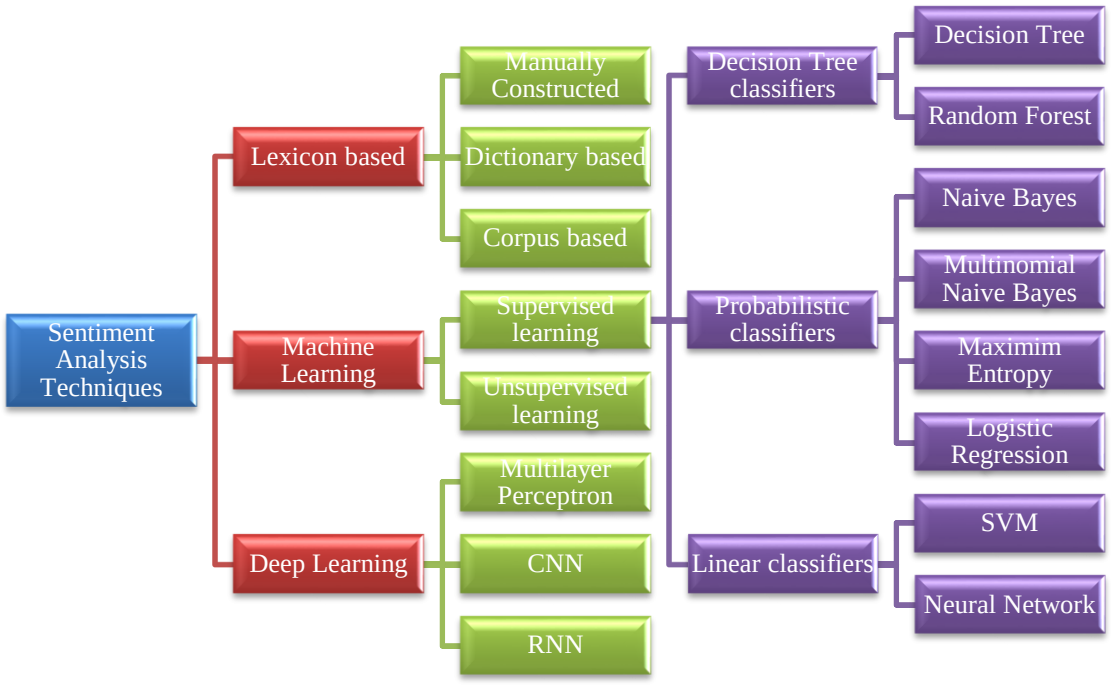


Figure 2.6 Different SA techniques

2.7.1.1 Lexicon Based

This approach is also known as rule-based approach. In this approach, certain rules are followed along with the use of sentiment lexicons to determine the sentiment from the text. The sentiment lexicon consists of words along with their sentiment polarity, e.g., "excellent" as positive, "horrible" as negative (Rehman and Bajwa, 2016; Syed et al., 2010). The sentiment orientation of an unknown document is computed by matching the words in the document to words in the sentiment lexicon and then taking the aggregate of their values using one of the various algorithms. The positive/negative values of the words in the text are aggregated which help in producing the semantic orientation for the entire text. Mainly the three approaches such as manual construction, dictionary-based and corpus-based are followed to construct the sentiment lexicon (Joshi et al., 2010). The manual construction approach is difficult and time-consuming. In this approach, polarities are manually assigned to sentiment words by humans. Dictionary-based is an iterative approach in which small set of sentimental words are selected initially and this set then iteratively grows by adding the synonyms and antonyms from the WordNet. This iterative process continues till no new words are remaining to be added to the seed list. Corpus-

based techniques depend on syntactic patterns in large corpora and can produce sentiment words with relatively high accuracy.

2.7.1.2 Machine Learning (ML)

Machine learning is a subfield of artificial intelligence that makes computers learn learn from data and make predictions on the related data without being explicitly programmed. ML is classified into two classes such as supervised and unsupervised machine learning. In supervised ML, there is a predetermined set of classes into which the documents are classified and training data is available for each class. The system uses any of the classification algorithms such as Naive Bayes (NB), Support Vector Machines (SVM), Decision Tree (DT), k-Nearest Neighbor (k-NN) and trains a model from the given data. This trained model is then used for making predictions and assigning the documents into different sentiment classes (Se et al., 2016). In case of unsupervised approach of ML, no labeled data is provided to models. This approach works on the basis of computation of Semantic Orientation (SO) of specific phrases within the text. If the average SO of phrases is above some predefined threshold, the text is classified as positive and otherwise, it is specified as negative. The brief description of frequently used SA techniques is given as follows.

a) **Naive Bayes:** Naive Bayes classifier is based on Bayes theorem and belongs to the family of probabilistic classifiers. It takes the probability distribution of words in the training dataset and assumes them to be mutually independent. Given a feature vector (x_1, x_2, x_3,….,x_n) and a class variable y, Naive Bayes assigns the class to feature vector according to the Bayes formula given in (2.1).

$$P(y \mid x_1, x_2, ..., x_n) = \frac{P(y)(x_1, x_2, ..., x_n \mid y)}{P(x_1, x_2, ..., x_n)} \tag{2.1}$$

In this formula, $P(y \mid x_1, x_2, ..., x_n)$ represents the posterior probability. *P(y)* is prior class probability and $P(x_1, x_2, ..., x_n)$ is the prior probability of feature set. These prior probabilities are obtained from the training dataset. $P(x_1, x_2, ..., x_n \mid y)$ represents the conditional probability of feature vector $(x_1, x_2, ..., x_n)$ given the class *y*. Naive Bayes finds the probability of each class to be assigned to this

feature vector and assigns the class with maximum probability. It assumes mutual independence among the features as shown in (2.2).

$$P(x_1, x_2, ..., x_n \mid y) = \prod_i P(x_i \mid y) \tag{2.2}$$

b) **Multinomial Naive Bayes (MNB):** Multinomial Naive Bayes is a specific version of Naive Bayes. Whereas a simple NB classifier models the document as the presence or absence of words, MNB takes into account the words counts. Given a class *c*, MNB estimates the conditional probability of a particular word as the relative frequency of the word in that class as given in (2.3).

$$P(t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \tag{2.3}$$

Here, *t* is the term/word and *c* is the class under consideration. The formula given in (2.3) calculates the probability of a word to be classified into a class as the count of that word in the class with respect to count of all the words in that class.

c) **Bernoulli Naive Bayes:** It generates boolean value/indicator about each term of the vocabulary equal to 1 if the term belongs to examining document, if not it marks 0. Non accruing terms in document are takes into document and they are factored when computing the conditional probabilities and thus the absence of terms is taken into account.

d) **Maximum Entropy (ME):** Maximum Entropy classifier is similar to NB classifier except that it doesn't make any assumption about the independence of features. The principle idea behind Maximum Entropy is that it tries to maximize the Entropy and at the same time satisfying the constraints specified. The idea behind Maximum Entropy is to have a model that is as unbiased as possible and thus the probability distribution to be as uniform as possible. Maximum Entropy is when all the events are equally likely to occur and have maximum uncertainty. The formula for Entropy is given in (2.4) and the goal is to maximize *H(P).*

$$H(P) = \sum p(a,b) \log(p(a,b)) \tag{2.4}$$

e) **k-Nearest Neighbor (k-NN):** k-NN is a type of instance-based learning, or also known as lazy learning. k-NN is used for both classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

f) **Support vector machines**: Support Vector Machines work on the concept of a decision plane or a hyperplane. It tries to find a hyperplane which separates the data belonging to two classes as far apart as possible as represented in (2.5).

$$(\vec{w}.\vec{x}) = \sum_i y_i \alpha_i (\vec{x_i}.\vec{x}) + b \tag{2.5}$$

Here, $\vec{x_i} = (x_{i1}, x_{i2}, ..., x_{in})$ is input feature vector, y_i is output class, $\vec{w_i} = (w_{i1}, w_{i2}, ..., w_{in})$ is the weight vector defining the hyperplane and α_i is a Lagrangian multiplier. Once the hyperplane is constructed, the class of any feature vector can be determined. Figure 2.7 shows the working of the SVM.

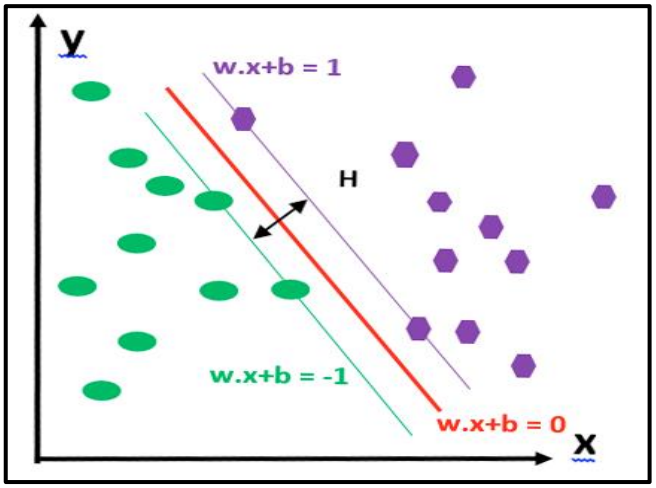


Figure 2.7 Support vector machines

g) **Logistic regression (LR)**: Logistic Regression is a multi-class logistic model which is used to estimate the probability of a response based predictor variables in which there are one or more independent variables that determine an outcome. The expected values of the response based predictor variable are formed based on the combination of values taken by the predictors.

h) **Decision tree (DT):** Decision Tree is a decision support tool that uses a treelike model for the decisions and likely outcomes. A decision tree is a tree in which each internal (non-leaf) node is labeled with an input feature and each leaf of the tree is labeled with a class.

i) **Random forest (RF):** Random Forest is an ensemble of Decision Trees. Random Forests construct multiple decision trees and take each of their scores into consideration for giving the final output. Decision Trees tend to overfit on a given data and hence they will give good results for training data but bad on testing data. Random Forests reduce overfitting as multiple decision trees are involved.

j) **AdaBoost**: AdaBoost was the first successful boosting algorithm developed for binary classification. It is best used to boost the performance of decision trees. AdaBoost algorithms can be used for both classification and regression problem. The weak learners in AdaBoost are decision trees with a single split, called decision stumps. AdaBoost works by putting more weight on difficult to classify instances and less on those already handled well.

k) **Gradient Boosing:** Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

2.7.1.3 Deep Learning

Deep learning is a branch of machine learning inspired by the human brain. It has emerged as a powerful approach for pattern recognition and language processing in recent years. Because of its ability to automatic feature engineering and appreciable accuracy, it

is getting widespread popularity these days. Deep learning refers to the number of layers that comprise the Neural Network (NN). Early NNs were defined with three layers; input, hidden, and output. Adding several hidden layers makes the NN 'deep' and enables it to learn more subtle and complex relationships. The number of 'hidden layers' decides how deep the network is. Neural networks cannot process direct words, but they work on word embeddings or more specifically feature vectors representing those words. One of the abilities of deep learning is that for feature learning, handcrafted features are replaced with efficient algorithms (Seshadri et al., 2016). These are capable of capturing very high-level features from input data. As neural networks learn features from the task in hand they can adapt to any domain. Deep nets can perform better than traditional machine learning approaches if a sufficient amount of training data is given (Akhtar et al., 2016c). Deep learning has found applications in a number of areas like sentiment analysis, computer vision, automatic speech recognition, etc. The brief description of some of the important deep learning models is given as follows.

a) **Multilayer perceptron model**: A Multi-Layer Perceptron model is a feed-forward supervised Artificial Neural Network (ANN) model that learns a function as given in (2.6) by training on a dataset.

$$f(.) = R_m \rightarrow R_o \tag{2.6}$$

Where *m* represents the number of input dimensions and *o* is the number of output dimensions. For a classification problem, the number of nodes in the input layer depends upon the length of the input vector and number of nodes in the output layer depends upon the number of pre-defined classes. There can be any number of hidden layers in between the input and output layer. Input layer takes $(x_1, x_2, ..., x_m)$ as the input vector. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $(w_1x_1 + w_2x_2 + ... + w_mx_m)$ followed by a non-linear activation function as given in (2.7).

$$g(.) = R \rightarrow R \tag{2.7}$$

The output layer then transforms values from the last hidden layer into output. Figure 2.8 shows the working of multi-layer perceptron model.

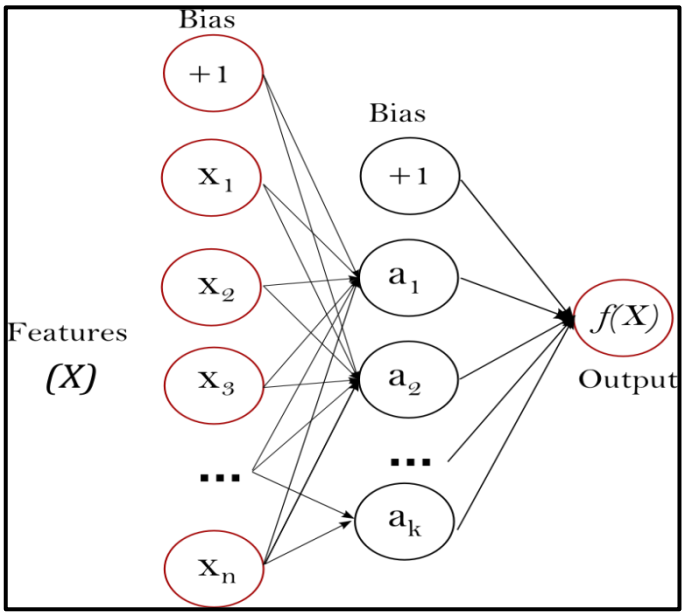


Figure 2.8 Multi-layer perceptron model

b) **Convolutional neural network (CNN):** Convolutional Neural Networks (CNNs) are very much similar to ordinary neural networks. Neurons in CNNs take some input, process it and propagate it further. The difference is that convolutional neural networks explicitly assume input as images. This is the reason they are explicitly used for analyzing image data. Regular neural networks don't scale well to full images. For small dimensions, these are manageable, but as the dimensions grow, more neurons and parameters are required leading to the problem of over fitting.

A CNN is a feed forward neural network which consists of four layers. First is the input layer that represents the sentences over *n x k* dimension, second is convolutional layer, then global max pooling layer and finally fully connected layer producing output results as shown in Figure 2.9. Convolutional layer is the main building block of a CNN as most of the computations are done at this layer. It is a feature extraction layer which extracts the local features through the filters and generates feature map computed by convolution kernel function and then outputs it to pooling layer. The inputs to CNN are sentences or documents

represented as a matrix. Each row of the matrix corresponds to one token, typically a word, but it could be a character. That i, each row is vector that represents a word. Typically, these vectors are word embeddings (low-dimensional representations) like *word2vec* or GloVe, but they could also be one-hot vectors that index the word into a vocabulary.

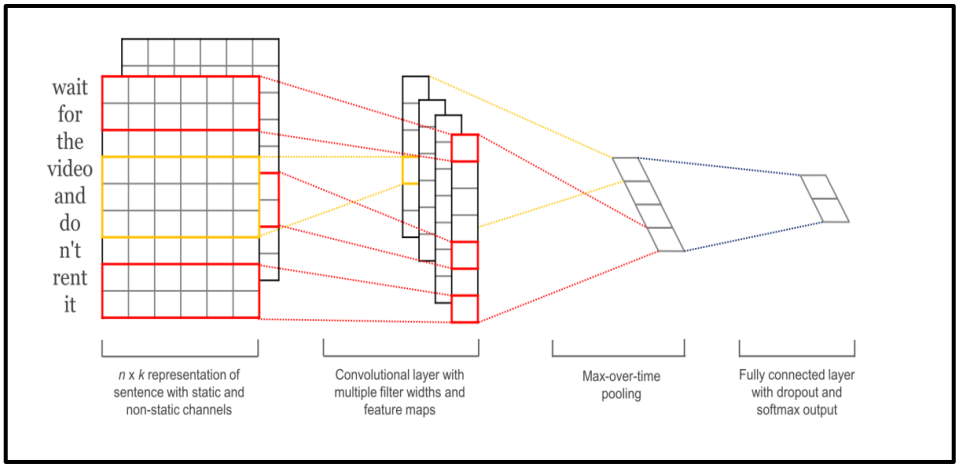


Figure 2.9 Convolutional neural network

c) **Recurrent neural network (RNN):** The major shortcoming of an ordinary neural network, i.e., all inputs and outputs in a layer are considered independent and due to this reason, future events can be predicted from the present state future events Recurrent Neural Networks help in resolving this problem. RNNs make use of loops making information to persist. RNNs have a memory which stores the information calculated so far and this information is used for future predictions. RNNs have applications in a number of areas where ordinary neural network can not solve the problem. For example, given a sequence of words, the next word in sequence can be determined using RNN. Other applications include handwriting recognition and speech recognition. The most common Recurrent Neural Network is Long Short Term Memory (LSTM) in short. The principle architecture of RNN is shown in Figure 2.10. In the figure, h_t is input and x_t is the corresponding output of the neural network. As RNN is unfolded, it becomes similar to a regular neural network.

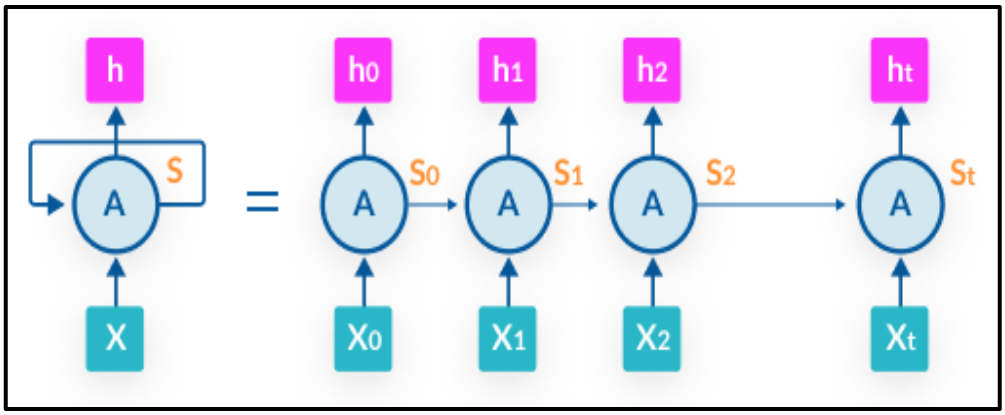


Figure 2.10 Recurrent neural network

2.7.2 Evaluation Measures

Mostly the researchers have used Accuracy (*A*) as evaluation measure. However, accuracy is not a sufficient metric to evaluate the efficiency and effectiveness of a classifier. Therefore, some of the researchers have also used the other metrics such as Precision (*P*), Recall (*R*) and F-measure (*F*) in addition to *A* as these metrics provide much greater insight into the performance features of a classifier. For a sentiment classifier, these four metrics can be defined in terms of True Positive (t_p), False Positive (f_p), True Negative (t_n) and False Negative (f_n) rates. Here, t_p rate represents the positive review and classifier also classifies it as positive, t_n rate represents the negative review and classifier also classifies it as negative. f_p represents the negative review but classifier classifies it as positive, f_n represents the positive review but classifier classifies it as negative. The brief description of the other evaluation measures is given as follows.

a) **Accuracy:** Accuracy *A* can be defined as a ratio of correctly predicted observations to the total observations and is given by using the formula (2.8).

$$A = \frac{t_p + t_n}{(t_p + f_p + t_n + f_n)} \tag{2.8}$$

b) **Precision:** The precision *P* can be defined as in terms of the exactness of a classifier and is given by the ratio of correctly predicted positive observations to the total predicted positive observations as given in (2.9). A higher *P* means less false positive and vice versa.

$$P = \frac{t_p}{(t_p + f_p)} \tag{2.9}$$

c) **Recall**: The recall *R* can be defined as in terms of the sensitivity or completeness of the classifier and is given by the ratio of correctly predicted positive observations to the all observations in actual class as given in (2.10). Higher *R* means less false-negative and vice versa.

$$R = \frac{t_p}{(t_p + f_n)} \tag{2.10}$$

d) **F-measure**: F-measure is measured by combining Precision and Recall, which is the weighted harmonic mean of both values, defined as given in (2.11).

$$F = \frac{2PR}{(P + R)} \tag{2.11}$$

2.8 Status of SA Work for Indian Languages

As far as development of SA systems with respect to Indian languages is concerned, most of the research work done in this domain is for English language and European languages. The research work reported in the field of SA for all Indian languages for different language families is presented in this section. The majority of the research work on SA has been performed for Indo-Aryan languages (such as Hindi, Bengali, and Urdu) and Dravidian languages (such as Tamil, Malayalam, and Kannada). The brief description of the research being performed on these languages is given as follows.

2.8.1 Languages with Major Research Work

a) Hindi

Joshi et al., (2010) first attempted to work on SA for Indian languages. The authors proposed a fallback strategy which follows three approaches such as In-language SA, machine translation and resource-based SA by developing own lexical resource Hindi SentiWordNet (HSWN). The authors achieved an accuracy of 78.14%. Balamurali et al., (2012) used WordNet sense-based features and experimented on a dataset travel reviews

to perform SA. Bakliwal et al., (2012) developed Hindi subjective lexicon using bi-lingual dictionary and translation based approach. The authors performed sentiment classification using this lexicon and achieved an accuracy of 79%. Mittal et al., (2013) performed SA of movie reviews and also handled negation as well as discourse relations. The authors used HSWN to perform SA and achieved an accuracy of 80.21%. Arora (2013) performed the sentiment analysis on a corpus of Hindi reviews and blog related to products and movies using subjective lexicon and N-gram approaches. Bansal et al., (2013) used deep learning to perform SA of movie reviews and achieved an accuracy of 64%. Sharma et al., (2014) proposed a SA system using an unsupervised dictionary-based approach and classified movie reviews into three categories, such as positive, negative and neutral. Their proposed methodology also handles negations and the system has achieved an accuracy of 65%. Sharma and Bhattacharyya (2014) proposed a bootstrap approach to extend the HSWN using existing HindiWordNet for all the four parts of speech, i.e., noun, adjective, and verb. The authors used this lexicon to validate the SA system over movie and product reviews domain and achieved an accuracy of 87%. Prasad et al., (2015) performed sentiment classification of tweets using decision tree under a constrained and unconstrained environment. Venugopalan and Gupta (2015) used tweet specific features and performed SA using ML algorithms SVM and DT. Kumar et al., (2015c) performed SA of Hindi tweets using binary and statistical features generated from HSWN. The authors mapped the input features to a random Fourier feature space and performed sentiment classification using a regularized least square method. Kumar et al., (2015a) performed SA of Hindi tweets using SVM and MNB classifier. The authors also constructed their own lexicon, namely DT_COOC lexicon using the distributional thesaurus and sentence co-occurrences. Sarkar and Chakraborty (2015) also performed SA of Hindi tweets using SVM and MNB classifier. Se et al., (2015) reported the work on SA for Hindi tweets using machine learning classifiers and classified them into positive, negative and neutral class. The authors analyzed that machine learning classifiers perform better within a constrained environment, i.e., without the availability of NLP tools like POS tagger, Named Entity Recognition (NER). Jha et al., (2015) proposed a Hindi opinion mining system and used NB classifier and unsupervised approach of POS tagging to perform SA. (Pandey and Govilkar 2015; Sharma et al., 2015) used unsupervised

lexicon based approach to perform SA using HSWN and classified sentences into positive, negative and neutral class. The authors also handled negations and discourse relations. Patra et al., (2015) used RNN and classified tweets into positive, negative and neutral class.

Seshadri et al., (2016) performed SA of Hindi tweets on Sentiment Analysis in Indian Languages (SAIL-2015) dataset. Phani et al., (2016) also used the SAIL-2015 dataset to perform SA for the Hindi language. The authors experimented using six classifiers such as NB, LR, DT, RF, SVM using four categories of features, namely word N-grams, character N-grams, surface, and SWN features. Sharma and Moh (2016) predicted Indian election results of 2016 using lexicon-based and machine learning approaches by collecting tweets in Hindi language and analyzed that ML techniques perform better than lexicon-based. Akhtar et al., (2016b) performed aspect-based SA in Hindi and developed an annotated dataset consisting of Hindi product reviews. The authors used SVM classifier and attained an accuracy of 54.05%. The authors also experimented in four domains ‘electronics’, ‘mobile apps’, ‘travels’ and ‘movies’ using three classifiers such as NB, DT and SMO in MEKA (a Multi-label/Multi-target Extension to WEKA) and reported that NB performs better in electronics and mobile apps domain, while decision tree reports better results for the travels and movies domain (Akhtar et al., 2016a). Akhtar et al., (2016c) first attempted to work on SA for Hindi using deep learning based model such as CNN and performed SA at both aspect and sentence level. Garg and Buttar (2017) performed aspect-based sentiment analysis of Hindi text using dictionary-based approach and classified the text into positive, negative and neutral class. Rai et al., (2017) used machine learning and lexicon-based approach to perform sentiment analysis of political reviews and analyzed that lexicon-based approach with negation handling outperforms the machine learning approach. Nanda et al., (2018) performed sentiment analysis of Hindi movie reviews using RF and SVM. The author classified the reviews into positive and negative class. Hussaini et al., (2018) performed sentiment analysis of book reviews using lexicon-based approach and improved the accuracy of the system using word sense disambiguation and by handling morphological variations. Rani and Kumar (2018) performed sentiment analysis of Hindi movie reviews using deep learning based CNN and classified the reviews into positive, negative and neutral class. Yadav and Bhojane

(2019) used a semi-supervised approach to perform sentiment analysis of Hindi mix-data and analyzed that negation handling and translation of words improve the accuracy of the system.

b) Bengali

Das and Bandyopadhyay (2010a) experimented on Bangla news text to find the polarity of opinions using SVM classifier and classified opinionated phrase as either positive or negative and attained a precision of 70.04% and a recall of 63.02%. Hasan and Rahman (2014) performed SA on Bangla text using contextual valency analysis. The authors used SWN and WordNet to find the prior valence of Bangla words. Kumar et al., (2015c) performed SA of Bengali tweets using binary and statistical features generated from HSWN. The authors mapped the input features to a random Fourier feature space and performed SA using a regularized least square method. Kumar et al., (2015a) performed SA of Bengali tweets using SVM and MNB classifier. The authors also constructed their own lexicon, namely DT_COOC lexicon using the distributional thesaurus and sentence co-occurrences. Sarkar and Chakraborty (2015) also performed SA of Bengali tweets using SVM and MNB classifier. Se et al., (2015) reported the work on SA for Bengali tweets using machine learning classifiers and classified them into positive, negative, and neutral class. The authors analyzed that machine learning classifiers perform better within a constrained environment, i.e., without the availability of NLP tools such as POS tagger, NER. Ghosal et al., (2015) experimented on 6000 sentences of the Bengali horoscope corpus to perform SA. The authors used machine earning techniques such as NB, SVM, k-NN, DT and RF using features such as unigrams, bigrams and trigrams. The authors reported that SVM with 98.7% accuracy outperforms than other techniques without removing stopwords and applying Information Gain (IG) as a feature selection method.

Hassan et al., (2016) experimented using the deep recurrent model, namely LSTM using two types of loss functions such as binary cross-entropy and categorical cross-entropy to perform SA for Bangla and Romanized Bangla text. The authors analyzed that categorical cross-entropy model performs better with 78% accuracy. Phani et al., (2016) used the annotated SAIL-2015 dataset (Patra et al., 2015) to perform SA for the Bengali language.

The authors experimented using six classifiers such as NB, LR, DT, RF, SVM using four categories of features, namely word N-grams, character N-grams, surface, and SWN features. Seshadri et al., (2016) performed SA of Bengali tweets on a dataset of SAIL-2015 using RNN and classified tweets into positive, negative and neutral class. Sarkar and Bhowmick (2017) performed sentiment analysis of Bengali tweets on SAIL 2015 dataset using MNB and SVM with different feature combinations. Al-Amin et al., (2017) performed sentiment analysis of Bengali comments collected from different blogging websites using a combination of *word2vec* and sentiment information of words. Sumit et al., (2018) performed sentiment analysis of Bengali news data using LSTM with different word embedding methods such as *word2vec* skip-gram and Continuous Bag-of-Word (CBOW) along with word to index method. Sarkar (2018) performed sentiment analysis of Bengali tweets using MNB and experimented with character N-gram features. The author analyzed that character N-gram features perform better in comparison to word N-gram features. Tripto and Ali (2018) performed sentiment analysis of Bengali YouTube comments using LSTM and CNN. The authors analyzed that LSTM performs better than CNN. Amin et al., (2019) used lexicon based approach to perform sentiment analysis of Bengali text by developing Bengali VADER using English VADER. Hoque et al., (2019) performed sentiment analysis of Bengali Facebook posts using different ML algorithms such as LR, SVM, Stochastic Gradient Descent (SGD), DT, k-NN, Linear Discriminant Analysis (LDA), GaussianNB, Sequential Model (SM), and Bidirectional LSTM (BLSTM) using *doc2vec* representation. The authors analyzed that BLSTM achieved the highest accuracy of 77.85%. Sarkar (2019) performed sentiment analysis of Bengali tweets using CNN and achieved an accuracy of 46.80% on SAIL-2015 dataset.

c) Tamil

Kumar et al., (2015c) performed SA of Tamil tweets using binary and statistical features generated from HSWN. The authors mapped the input features to a random Fourier feature space and performed SA using a regularized least square method. Se et al., (2015) reported the work on SA for Tamil tweets using machine learning classifiers and classified them into positive, negative and neutral class. The authors analyzed that

machine learning classifiers perform better within a constrained environment, i.e., without the availability of NLP tools such as POS tagger, NER.

Nivedhitha et al., (2016) proposed an unsupervised dictionary-based technique to perform SA of Tamil tweets. The authors used genism *word2vec* topic modeling toolkit to convert the string data into vector form and performed sentiment classification using HSWN. Sharmista and Ramaswami (2016) experimented on 100 Tamil product reviews to perform SA using decision tree classification techniques such as J48, Logistic Model Tree (LMT), BagCart, Recursive, RF, and C50. The authors attained an accuracy of 0.9469 and 0.9457 for LMT and random forest respectively. Se et al., (2016) performed SA of Tamil movie reviews using ML techniques such as NB, J48, SVM and ME. The authors also performed SA on the same dataset by considering SentiWordNet words as features and applied the four ML algorithm and concluded that SVM achieves the best accuracy of 75.9% in comparison to other for SentiWordNet features. Seshadri et al., (2016) performed SA of Tamil tweets on a dataset of SAIL-2015 using RNN and classified tweets into positive, negative and neutral class. Phani et al., (2016) used the annotated SAIL-2015 dataset (Patra et al., 2015) to perform SA for Tamil language and experimented using six ML classifiers using different features, namely word N-grams, character N-grams, surface, and SentiWordNet features. Padmamala and Prema (2017) performed sentiment analysis of Tamil text using RNN and analyzed that RNN performs better than the syntactic approach. Sharmista and Ramaswami (2018) performed sentiment analysis of mobile product reviews using SVM and fuzzy SVM in R language. The authors analyzed that fuzzy SVM outperforms than SVM.

d) Malayalam

Anagha et al., (2014) proposed dictionary-based approach to perform SA and also developed a lexical resource file of 2000 sentiment words for the Malayalam text. The authors classified the Malayalam reviews into positive and negative classes by attaining accuracy of 93.6%. Nair et al., (2014) proposed a rule-based approach to perform SA of Malayalam movie reviews at the sentence-level. The authors used Sandhi splitter for the tokenization of sentences and classified the sentences into three classes positive, negative

and neutral. The authors also handled negations and smileys by building a dictionary pre-tagged with positive and negative sentiment. An accuracy of 85% was achieved. In 2015, the authors proposed a hybrid approach using a combination of rule-based and ML techniques. The authors also computed the ratings of reviews along with sentiment class. It was analyzed that SVM outperforms than Conditional Random Field (CRF) and achieved an accuracy of 91% (Nair et al., 2015). Jayan et al., (2015) proposed a hybrid approach by combining rule-based and ML to perform SA of Malayalam film reviews at aspect, sentence and document level. The authors used CRF for tagging of the dataset and then applied rules to classify the documents into three classes such as positive, negative and neutral. Anagha et al., (2015) proposed a fuzzy-based approach to perform SA of Malayalam movie reviews. The authors used TnT (Trigrams'n'Tags) tagger to tag the input corpus and then fuzzy triangular membership function to extract the sentiment from text. The precision rate of 91.6% was reported while comparing the system's output with manually tagged output.

Thulasi and Usha (2016) performed aspect-based SA on Malayalam movie and product reviews and achieved an accuracy of 84.7%. Ashna and Sunny (2017) performed sentiment analysis of Malayalam movie reviews at sentence and document level using lexicon-based approach. The authors achieved an accuracy of 87.5% and 90% for sentence and document level respectively. Kumar et al., (2017) performed sentiment analysis of Malayalam tweets using LSTM and CNN. The authors experimented with different activation functions such as Exponential Linear Units (ELU), Rectified Linear Units (ReLU) and Scaled Exponential Linear Units (SELU). Kumar et al., (2019) extended the dataset to 13000 tweets and experimented with different machine learning and deep learning techniques to perform sentiment analysis in the Malayalam language.

e) Urdu

Syed et al., (2010) proposed lexicon-based approach to perform SA of Urdu text by manually creating sentiment lexicon. The authors extracted SentiUnits from text using shallow parsing. The authors further extended this work to handle the implicit negation problem and tested their system on the data set of movie reviews (Syed et al., 2011).

Earlier, the authors worked on sentences representing a single target. In this work, the authors extended their model to handle the presence of multiple targets as in the comparative sentences and used dependency parsing algorithm to associate the SentiUnits to their targets. The authors tested their modified approach on a dataset of movie reviews and electronic appliances and achieved an accuracy of 82.5% (Syed et al., 2014).

Rehman and Bajwa (2016) used lexicon-based approach to perform SA of Urdu news articles using Urdu SWN and achieved an accuracy of 66% by classifying the documents into positive and negative class. Mukhtar et al., (2017) validated their SA results using three standard evaluation measures, i.e., McNemar's test, kappa statistic, and root mean squared error. Khan et al., (2017) created a labeled dataset of Urdu tweets and performed sentiment analysis using LR. Mukhtar and Khan (2018) used ML approaches to perform SA of Urdu blogs. Asghar et al., (2019) created their own Urdu sentiment lexicon by using existing lexical resources and performed sentiment classification of product reviews by classifying them into positive and negative class. Mukhtar and Khan (2019) performed sentiment analysis of Urdu blogs using lexicon-based approach by extending the existing Urdu lexicon and classified the reviews into positive, negative and neutral class.

f) Kannada

Deepamala and Kumar (2015) performed SA of Kannada documents. They manually created a polarity lexicon for Kannada language consisting of 5043 words and compared the accuracies of lexicon-based approach with NB and ME. They observed that ME with 93% accuracy outperforms than lexicon-based approach and NB. Kumar et al., (2015b) performed SA on Kannada web documents by exploring the usefulness of semantic and machine learning approaches. They identified that in the case of a semantic approach, baseline method outperforms than other semantic approaches like negation, sentence-based and Turney's methods. In case of ML approaches, NB performs better than other supervised learning methods such as DT, RF, Sequential minimal optimization (SMO), Abstract Data Type (ADT) Tree and Breadth-First. The authors concluded that the precision of ML approaches is 7.22% better than semantic approaches. Hegde and Padma (2015) performed SA of Kannada mobile product reviews extracted from newspaper

‘Prajavani’ using lexicon-based approach for aspect extraction and NB classifier to identify the polarity of reviews. The authors reported an accuracy of 65% but the system lacks in handling the multi-class, comparative, and conditional sentences.

Rohini et al., (2016) performed feature-based SA of Kannada movie reviews using a decision tree. The authors extracted nouns as features and adjectives as sentiment words using Kannada POS Tagger. Hegde and Padma (2017) used Random Forest Ensemble after extending the previous corpus of mobile product reviews and improved accuracy from 65% to 72% in this work.

2.8.2 Languages with Minor Research Work

This sub-section discusses the research work on Indian languages such as Punjabi, Oriya, Telugu, Nepali, Marathi, Konkani, Manipuri, Gujarati, and Sindhi (belonging to Indian language families) which have contributed a little in the field of SA.

a) Punjabi

Kaur and Gupta (2014a) performed SA of Punjabi text using a hybrid approach using N-gram model and NB on a dataset collected from newspapers and blogs. They compared their approach with existing approaches such as Hindi Subjective Lexicon, HSWN, bilingual Dictionary, and Translated Dictionary. Kaur and Gupta (2014b) proposed an algorithm for SA of Punjabi text. They used a bilingual dictionary-based approach to develop a subjective lexicon for Punjabi language using HSWN. The authors validated their approach using a subjective lexicon and analyzed that their approach is better over existing approaches. Arora and Kaur (2015) developed an offline application to perform SA of Punjabi political reviews using scoring approach.

Kaur and Kaur (2017) performed sentiment analysis of Punjabi news using SVM. The authors achieved an accuracy of 90% by performing the sentiment classification into classes such as positive and negative. Kaur and Gupta (2017) performed sentiment analysis of Punjabi newspapers using the hybrid technique by integrating subjective lexicon, N-gram modeling, and SVM. The author achieved an accuracy of 78.02%. Singh et al., (2018) performed sentiment analysis of Punjabi text concerned with farmer suicide

cases of Punjab. The authors used a deep neural network on 275 text documents and achieved an average accuracy of 90.29% for sentiment classification.

b) Oriya

Jena and Chandra (2014) performed opinion mining of Oriya text using SVM. Sahu et al., (2016a) performed SA on a dataset of 1000 sentences of movie reviews in Odia language using NB classifier and achieved an accuracy of 92%. Then the authors applied three supervised classification techniques such as NB, LR, and SVM on a dataset of 6000 sentences and compared the performance of these techniques using evaluation measures precision, recall and accuracy. It was analyzed that LR with accuracy 88% outperformed than NB and SVM (Sahu et al., 2016b).

c) Nepali

Gupta and Bal (2015) performed the first work on sentiment detection of Nepali text on a dataset of 25,435 sentences collected from online Nepali National Dailies, namely 'ekantipur' and 'nagariknews'. The authors developed their own Nepali SWN, namely 'Bhavanakos' and compared NB with resource-based SWN approach. It was concluded that ML approach is better than resource-based approach. Thapa and Bal (2016) reported work on SA for the Nepali language on a dataset of 384 book and movie reviews at the document level. The authors used Bag-of-words and Term Frequency (TF)-Inverse Document Frequency (IDF) features extraction models with and without stop words removal. The authors classified the reviews by applying classifiers such as SVM, Multinomial NB and LR. The authors compared the performance of classifiers with evaluation metrics such as F-measure and accuracy; and concluded that MNB outperforms than SVM and LR with any of feature extraction method.

d) Marathi

Balamurali et al., (2012) used WordNet sense-based features to perform SA of travel reviews. Chaudhari et al., (2017) performed SA of Marathi documents and used Marathi WordNet to compute the sentiment polarity. Deshmukh et al., (2017) performed

sentiment analysis of Marathi language and classified the text into three classes positive, negative and neutral.

e) Telugu

Mukku et al., (2016) performed SA of Telugu sentences collected from Indian Languages Corpora Initiative (ILCI). The authors used *doc2vec* for converting sentences into sentence vectors and performed SA using ML techniques such as NB, LR, SVM, DT, MLP Neural Network and RF. The authors also experimented by an ensemble of all the six ML classifiers. Naidu et al., (2017) proposed a two-phase SA for Telugu news sentences using Telugu SentiWordNet. First, the authors performed subjectivity classification then further classified them into positive and negative sentences. Mukku and Mamidi (2017) performed sentiment analysis of Telugu text and experimented with different machine classification algorithms for binary and ternary sentiment classification. The authors observed that RF performs better for binary classification and LR performs better for ternary sentiment classification. Shalini et al., (2018) performed sentiment analysis of Telugu movie reviews data using CNN and achieved an accuracy of 51%.

f) **Konkani**

Miranda and Mascarenhas (2016) developed an opinion mining system for Konkani language, namely KOP (Konkani OPinion mining system). The authors used Konkani SWN to perform SA and also handled negations, conjunctions as well as sarcasm.

g) Manipuri

Nongmeikapam et al., (2014) performed SA on Manipuri text collected from daily newspapers. They processed the text for POS tagging using CRF then identified the verbs using a modified verb lexicon. After that, they counted the polarity for each class, such as positive, negative and neutral separately. Then highest polarity of the three decided the sentiment polarity of the document.

h) Gujarati

Joshi and Vekaria (2017) performed SA of tweets. The authors used SVM classifier and achieved an accuracy of 92% by classifying the tweets into positive and negative classes. Gohil and Patel (2019) developed Gujarati SWN using Hindi SWN and IndoWordNet. The authors used Gujarati SWN to perform the SA of tweets and achieved an accuracy of 52.72% using unigram presence and 52.95% using the simple scoring based method.

i) Sindhi

Ali and Wagon (2017) performed the first work on SA of Sindhi text using structurization and machine learning supervised model. They used SVM and k-NN classifiers for experimentation and evaluated the performance of the model using precision, recall and F-measure parameters.

Table 2.7 summarizes the different approaches, corpora, corpus sizes, lexical resources/tools/programming languages and evaluation measures used to develop SA systems for all the Indian languages considered in this chapter.

Table 2.7: Summary of approaches and lexical resources for different Indian languages

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Gujarati (Gohil and Patel, 2019)	Scoring based	Tweets	1120 tweets	Hindi SWN	A: 52.83%
Gujarati (Joshi and Vekaria, 2017)	SVM	Twitter	40 tweets	Not Specified	A: 92%
Sindhi (Ali and Wagon, 2017)	SVM, k-NN	General	9779 records	Sindhi NLP tool	Not Specified

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Konkani (Miranda and Mascarenha, 2016)	Lexicon Based	General	Not Specified	SWN	Not Specified
Oriya (Jena and Chandra, 2014)	SVM	General	Not Specified	SVM library	Not Specified
Odia (Sahu et al., 2016a)	NB	Movie	1000 sentences	Natural Language Toolkit (NLTK)	A: 0.92, P: 0.93, R: 0.97
Odia (Sahu et al., 2016b)	NB, SVM, Logistic Regression	Movie Reviews	6000 reviews	Python language	LR- P: 0.75, R:0.797, A: 0.88
Nepali (Gupta and Bal, 2015)	Lexicon based	Nepali National Dailies	25435 sentences	Nepali SentiWordNet	P: 47.2, R:54.8
	NB			Not Specified	P: 23.6, R:70.2
Nepali (Thapa and Bal, 2016)	SVM, MNB, and LR	Book, Movie Reviews	384	Natural Language Toolkit (NLTK) and Python packages	MNB- F: 0.67, A: 0.65
Manipuri (Nongmeikapam et al., 2014)	CRF	Newspapers	550 letters	POS Tagger	R: 72.10%, P: 78.14%, F: 75.00%
Marathi (Balamurali et al., 2012)	SVM	Travel reviews	150 reviews	WordNet, LibSVM package	A: 84%

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Marathi (Chaudhari et al., 2017)	NLP based	General	Not Specified	Marathi WordNet, GATE	Not Specified
Marathi (Deshmukh et al., 2017)	Corpus-based	General	Not Specified	Yandex translator. English SWN	Not Specified
Telugu (Mukku et al., 2016)	NB, SVM, DT, RF, MLP Neural Network, LR	Newspapers	1644 annotated + 7,21,785 raw sentences	Python	Ensemble-A: 73.85% (Binary), A: 60.13% (Ternary)
Telugu (Naidu et al., 2017)	Lexicon based	Newspapers	1400 sentences	SWN	A: 81%, P:0.71, R: 0.77, F:0.74
Telugu (Mukku and Mamidi, 2017)	ML	News, Reviews, Twitter, Facebook	5410 sentences	*doc2vec*, scikit-learn toolkit	Binary: 73.85%, Ternary: 60.13%
Telugu (Shalini et al., 2018)	CNN	Movie Reviews	2000 sentences	Not Specified	Accuracy: 51%
Punjabi (Kaur and Gupta, 2014a)	Hybrid (N-gram, NB)	Newspapers and Blogs	44,200 sentences	WEKA, Java	P:0.78, R:0.6, F: 0.67

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Punjabi (Kaur and Gupta, 2014b)	Lexicon based	General	Not Specified	SWN	A: 78.02
Punjabi (Arora and Kaur, 2015)	Scoring approach	General	Not Specified	SWN	Not Specified
Punjabi (Kaur and Gupta, 2017)	Hybrid	Newspapers and blogs	700 documents	WEKA	A: 78.02%
Punjabi (Kaur and Kaur, 2017)	SVM	News	Not Specified	Not Specified	A: 90%
Punjabi (Singh et al., 2018)	Deep learning	Farmer suicide	275 documents	NLTK toolkit	A: 90.29%
Kannada (Hegde and Padma, 2015)	NB	Mobile product reviews	Not Specified	Python language	A: 65%, P: 62.5%, R: 75%, F: 68.2%
Kannada (Deepamala and Kumar, 2015)	Lexicon based	General	344 documents	Kannada stemmer	ME- A: 0.93, P: 0.9, R:0.89, F:0.89
	NB, ME				
Kannada (Kumar et al., 2015b)	Lexicon based	Products reviews	197 reviews	Kannada POS Tagger	NB-P:0.81
	J48, NB, SVM, Random			WEKA	

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
	Tree, ADT Tree, Breadth-First				
Kannada (Hegde and Padma, 2017)	RF	Mobile Products reviews	Not specified	R Studio	A:72%
Kannada (Rohini et al., 2016)	DT	Movie reviews	100 reviews	Kannada POS Tagger	P: 0.78, R:0.79
Malayalam (Anagha et al., 2014)	Lexicon based	Multi-domain reviews	Not specified	TnT Tagger, Malayalam lexical resource	A: 93.6%
Malayalam (Nair et al., 2014)	Lexicon based	Movie Reviews	Not specified	Sandhi Splitter, Python	A: 85%
Malayalam (Nair et al., 2015)	Hybrid ([SVM, CRF] + rule based)	Movie Reviews	30,000 tokens	SVM and CRF libraries	SVM- P: 0.806, R: 0.951 and F: 0.863
Malayalam (Jayan et al., 2015)	Hybrid (CRF + rule based)	Movie reviews	30,000 tokens	CRF library	A:82%
Malayalam (Anagha et al., 2015)	Fuzzy logic	Movie Reviews	2500 words	TnT Tagger	P: 91.6%

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Malayalam (Thulasi and Usha, 2016)	Lexicon based	Movie and Product reviews	Not specified	Sandhi splitter, TnT Tagger, Malayalam SWN	A: 84.7%
Malayalam (Ashna and Sunny, 2017)	Lexicon based	Movie reviews	Not specified	POS tagger	A: 87.5% sentence, A: 90% document
Malayalam (Kumar et al., 2017)	LSTM, CNN	Tweets	12922 tweets	Not Specified	P: 0.9823, R: 0.9824, F: 0.9823, A: 0.9824
Malayalam (Kumar et al., 2019)	SVM, CNN LSTM, regularized least square classification with random kitchen sink mapping (RKS-RLSC)	Tweets	13000 tweets	TensorFlow, Keras	P: 0.9827, R: 0.9826, F: 0.9828, A: 0.9826 [LSTM with ReLU function]
Urdu (Syed et al., 2010)	Lexicon based	Movie and Product reviews	753 reviews	Shallow parser	Movie- A: 72% Product- A: 78%

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Urdu (Syed et al., 2011)	Lexicon based	Movie reviews	450 reviews	Shallow parser	Set1- P:0.864, R: 0.837, F: 0.850 Set2- P:0.590, R: 0.779, F: 0.677 Set3- P:0.510, R:0.615, F:0.558
Urdu (Syed et al., 2014)	Lexicon based	Movie and electronic appliances reviews	700 movie and 650 electronic appliance reviews	Shallow parser	A: 82.5%
Urdu (Rehman and Bajwa. 2016)	Lexicon based	News	124 documents	SentiWordNet	A: 0.66. R: 0.79, P: 0.69, F: 0.73
Urdu (Mukhtar and Khan, 2017)	SVM, DT, k-NN	Blogs (14 domains)	6025 sentences	WEKA	k-NN- A: 67.0185%, P: 0.674, R: 0.6703, F: 0.6703

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Urdu (Khan et al., 2017)	LR	Tweets	999 tweets	Not Specified	A: 60.54%
Urdu (Asghar et al., 2019)	Lexicon based	Product reviews	1201 reviews	English SWN, English to Urdu bilingual dictionary, Urdu Opinion Lexicon, Bing Liu's List of Opinion Words	P: 95.2, R:88.4, F: 91.3, A: 92.4
Urdu (Mukhtar and Khan, 2019)	Lexicon based	Blogs	6025 sentences	POS tagger	A: 89.03%, P: 0.86, R: 0.90, F: 0.88
Tamil (Nivedhitha et al., 2016)	Lexicon based	Tweets	691 tweets	SWN, Genism Python toolkit	A: 0.7062, P: 0.7065, R:0.6987, F:0.6924
Tamil (Sharmista and Ramaswami, 2016)	J48, LMT, BagCart, Recursive, RF and C50	Product reviews	100 reviews	R	LMT- A: 0.9469
Tamil (Se et al., 2016)	SVM, DT, NB, ME	Movie Reviews	534 reviews	SWN	SVM- A:75.9%

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Tamil (Seshadri et al., 2016)	RNN	Tweets	SAIL-2015	MIKE(Mining Intelligence and Knowledge Exploration)	A: 88.23, F:0.802
Tamil (Phani et al., 2016)	MNB, DT, SVM, RF, LR	Tweets	SAIL-2015	SWN	NB- A: 62.16% (Binary); RF-A: 45.24% (Ternary)
Tamil (Kumar et al., 2015c)	Regularized Least Square	Tweets	SAIL-2015	SWN	A: 32.32%
Tamil (Se et al., 2015)	NB	Tweets	SAIL-2015	SciPy	Constrained - A: 39.28%
Tamil (Padmamala and Prema, 2017)	RNN	General	Not Specified	Not Specified	A: 71.6%
Tamil (Sharmista and Ramaswami, 2018)	SVM and fuzzy SVM	Product reviews dataset	5000 sentences	R language	A: 76.39 (fuzzy SVM)
Bengali (Das and Bandyopadhyay, 2010a)	Hybrid (SVM + rule based)	News	447 sentences	Dependency parser, SWN	P: 70.04%, R: 63.02%

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Bengali (Hasan et al., 2014)	Lexicon based	General	approx. 150 sentences	WordNet, SWN	Percentage of positive, negativity and neutrality
Bengali (Ghosal et al., 2015)	NB, SVM, k-NN, DT, RF	Horoscopes	6000 sentences	WEKA	SVM-A:98.7%
Bengali (Hassan et al., 2016)	LSTM	General	6698 entries	Python's Keras library	A: 78%
Bengali (Phani et al., 2016)	MNB, DT, SVM, RF, Logistic Regression	Tweets	SAIL-2015	SWN	NB- A: 67.83% (Binary) LR-A: 51.25% (Ternary)
Bengali (Seshadri et al., 2016)	RNN	Tweets	SAIL-2015	MIKE	A: 65.16, F:0.644
Bengali (Kumar et al., 2015c)	Regularized Least Square	Tweets	SAIL-2015	SWN	A: 31.4%
Bengali (Kumar et al., 2015a)	SVM	Tweets	SAIL-2015	SWN	Constrained - A: 43.2% , Unconstrained- A: 42%

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Bengali (Sarkar and Chakraborty, 2015)	MNB	Tweets	SAIL-2015	WEKA, SWN	Constrained - A: 41.2%, Unconstrained - A: 40.4%
Bengali (Se et al., 2015)	NB	Tweets	SAIL-2015	SciPy	Constrained -A : 33.6%
Bengali (Sarkar and Bhowmick, 2017)	MNB, SVM	Tweets	SAIL-2015	WEKA	A: 45%
Bengali (Al-Amin et al., 2017)	Hybrid	General	16000 comments	Not Specified	A: 75.5%
Bengali (Sumit et al., 2018)	LSTM	News data	26, 890, 638 sentences	Not Specified	A: 83.79%
Bengali (Sarkar, 2018)	MNB	Tweets	SAIL 2015	WEKA	A: 48.5%
Bengali (Tripto and Ali, 2018)	LSTM, CNN	YouTube comments	8910 sentences	Python Keras framework	LSTM- A: 65.97
Bengali (Amin et al., 2019)	Lexicon based	General	Not Specified	SWN	Not Specified

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Bengali (Hoque et al., 2019)	LR, SVM, SGD DT, k-NN, LDA, GaussianNB, SM, and BLSTM	Facebook Posts	7000 posts	Facebook API	BLSTM-A: 77.85%, P: 78.06%, R: 77.39% and F: 77.72%
Bengali (Sarkar, 2019)	CNN	Tweets	SAIL-2015	Not specified	A: 46.80%
Hindi (Joshi et al., 2010)	Lexicon based	Movie Reviews	250 Hindi reviews + 1000 English Reviews	SWN	A: 60.31%
	Machine learning (SVM)			Rapid Miner 5.0	A: 78.14%
	Translation Based (SVM)			Google Translator	A: 65.96%
Hindi (Pandey and Govilkar, 2015)	Lexicon based	Movie Reviews	Not specified	SWN	Not specified
Hindi (Sharma et al., 2015)	Lexicon based	Tweets	100	SWN	A: 77.75%, P: 0.85. R: 0.88
Hindi (Sharma and Bhattacharyya, 2014)	Lexicon based	Movie and product reviews	900 reviews	SWN	A: 87%

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Hindi (Akhtar et al., 2016b)	SVM	Movie, product, travel, mobile apps reviews	5417 reviews	Shallow Parser	A: 54.05%
Hindi (Akhtar et al., 2016a)	NB, DT, and SMO	Movie, product, travel and mobile apps reviews	5254 reviews	WEKA	DT- A: 54.48% (Products), A: 47.95% (Mobile apps), A: 65.20% (Travels), A: 91.62% (Movies)
Hindi (Seshadri et al., 2016)	RNN	Tweets	SAIL-2015	MIKE	A: 72.01, F:0.714
Hindi (Jha et al., 2015)	NB Lexicon based	Movie Reviews	200 reviews	NLTK TnT POS Tagger	A: 87.1%
Hindi (Sharma et al., 2014)	Lexicon based	Movie Reviews	Not specified	POS tagger	A: 0.65, P: 0.66, R: 0.78
Hindi (Bansal et al., 2013)	Deep belief Network	Movie Reviews	300 reviews	Theano Library	A: 64%

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Hindi (Phani et al., 2016)	MNB, DT, SVM, RF, LR	Tweets	SAIL-2015	SentiWordNet	LR- A: 81.57% (2-class) LR-A: 56.96% (3-class)
Hindi (Mittal et al., 2013)	Lexicon based	Movie Reviews	662 reviews	SWN	A: 80.21%
Hindi (Arora, 2013)	Lexicon based, N-gram Modeling	Products and Movie Reviews	973 reviews	WEKA	A: 61.6% (N-gram + lexical features)
Hindi (Sharma and Moh, 2016)	Lexicon Based	Election tweets	42,235 tweets	SWN	A: 34%
	NB, SVM				NB- A: 62.1%, P:0.71, R:0.61
Hindi (Bakliwal et al., 2012)	Lexicon based	Products reviews	700 reviews	Hindi Shallow parser, Hindi subjective lexicon	A: 79.03%
Hindi (Venugopalan and Gupta, 2015)	SVM, DT	Tweets	SAIL-2015	WEKA	SVM-A:42.83%

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Hindi (Prasad et al., 2015)	DT	Tweets	SAIL-2015	WEKA	Constrained - P: 0.822, R: 0.815, F: 0.804 Unconstrained- P: .735, R: .707, F: 0.680
Hindi (Kumar et al., 2015c)	Regularized Least Square	Tweets	SAIL-2015	SWN	A: 47.96%
Hindi (Kumar et al., 2015a)	SVM	Tweets	SAIL-2015	SWN	Constrained - A: 49.68% Unconstrained- A: 46.25%
Hindi (Sarkar and Chakraborty, 2015)	MNB	Tweets	SAIL-2015	WEKA, SWN	Constrained - A: 50.75 %, Unconstarined- A: 48.82 %
Hindi (Se et al., 2015)	NB	Tweets	SAIL-2015	SciPy	Constarined -A: 55.67%
Hindi (Balamurali et al., 2012)	SVM	Travel reviews	200 reviews	WordNet, LibSVM package	A: 72%

Language (Author)	**Approach**	**Corpus Type**	**Corpus Size**	**Lexical Resource/Tool/ Language used**	**Evaluation Measures**
Hindi (Akhtar et al., 2016c)	Hybrid	Tweets, Products, Movies, Restaurant Reviews	2152 reviews	DL4J, LibSVM package	Tweets- A: 58.62, Laptop reviews- A: 68.04, Restaurant reviews-A: 77.16
Hindi (Garg and Buttar, 2017)	Lexicon based	General	201 reviews	Not Specified	A: 82.3%
Hindi (Rai et al., 2017)	Lexicon based, NB, SVM	Political reviews	2000 sentences	English SWN	NB-A: 48.8%, SVM- A: 78.2%, Lexicon-A: 83.6%
Hindi (Nanda et al., 2018)	RF, SVM	Movie Reviews	Not Specified	Not Specified	RF- A: 91.07 ; SVM- A: 89.73;
Hindi (Hussaini et al., 2018)	Lexicon based	Book reviews	700 sentences	Hindi shallow parser, Python's urllib package, and BeautifulSoup4	A: 82.4%

Language (Author)	Approach	Corpus Type	Corpus Size	Lexical Resource/Tool/ Language used	Evaluation Measures
Hindi (Rani and Kumar, 2018)	CNN	Movie reviews	7354 sentences	Jupyter Notebook, TFLearn library	A: 95%
Hindi (Yadav and Bhojane, 2019)	Lexicon based, Neural Network	Health, Business, Current affairs, Tourism, Movie, Technology and Product	1916 sentences	Shabdkosh, quillpad, HSWN	Approach 1- A: 52%, Approach 2- A: 71.5%, Approach 3- A: 70.27%

** P-Precision, R-Recall, A-Accuracy, F- F-Measure

2.9 Findings of Systematic Survey

This section concludes the results identified while conducting this systematic survey and efforts have been made to answer all the research questions given in Table 2.1.

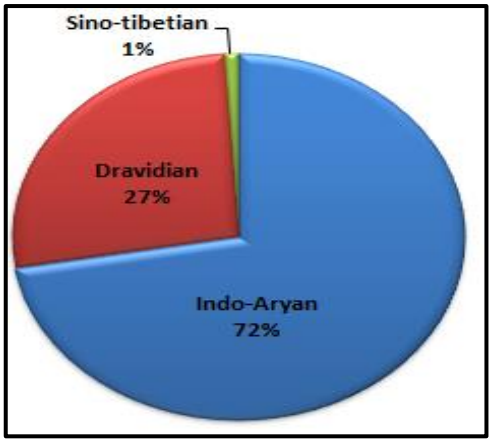


Figure 2.11 Status of SA research work in Indian language families

The answer of research question **RQ3** is reported through Figures 2.11 and 2.12 which show the percentage of research studies covering different Indian languages under Indian language families over a period from 2010 to 2019.

From Figure 2.11, it can be observed that 72% of the research work on SA has been performed on Indo-Aryan languages (out of which major part, i.e., 28% is covered by the Hindi language), followed by 27% on Dravidian and 1% on Sino-Tibetian language families. And the Austroasiatic language family is still unexplored for SA research work.

Figure 2.12 depicts that the majority of research has been done for Hindi language (28%), followed by Bengali (17%), Tamil (9%), Malayalam (9%), Urdu (9%), Punjabi (6%), Kannada (5%), Oriya (3%), Nepali (2%), Telugu (4%), Marathi (3%), Gujarati (2%), Konkani (1%), Sindhi (1%), and Manipuri (1%). The systematic map in Figure 2.12 helps in recognizing the mostly explored Indian languages in the field of sentiment analysis from 2010 to 2019.

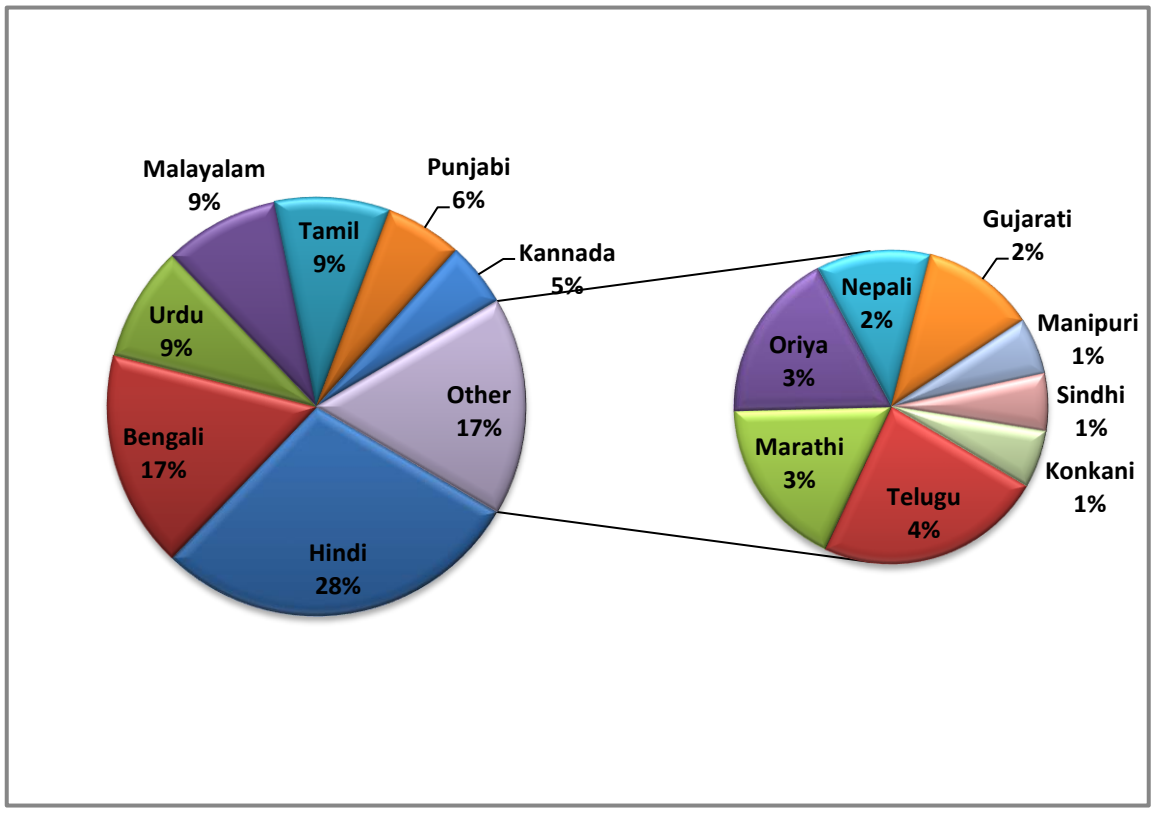


Figure 2.12 Status of SA research work in different Indian languages

From this analysis, it is concluded that one-third of the research work has been done on Hindi language belonging to the Aryan languages family. However, with the introduction of Unicode (UTF-8) standards, web pages in the Hindi language have been increasing rapidly. Hindi is spoken by a total of 422 million people; it's about 41% of the total population of India. The government of India is also promoting the Hindi language by providing online contents of official websites in Hindi. Therefore, researchers are attracting towards performing SA in the Hindi language so that the large volume of opinions shared by people on the web can be effectively leveraged.

The answer for the research question **RQ4** is attained through Table 2.8 which summarizes the SA techniques used by researchers for all Indian languages. It reports the SA research work for Indian languages along with publications count over which different SA techniques have been experimented.

Table 2.8: Indian languages for each SA technique

Techniques (Count)	**Languages (Count)**	**References**
Lexicon based (34)	Hindi (13), Urdu (7), Kannada (2), Bengali (2), Konkani (1), Nepali (1), Punjabi (1), Tamil (1), Telugu (1), Marathi (1), Malayalam (4)	Bakliwal et al., (2012); Miranda and Mascarenhas (2016); Gupta and Bal (2015); Anagha et al., (2014); Nair et al., (2014); Thulasi and Usha (2016); Ashna and Sunny (2017); Deepamala and Kumar (2015); Kaur and Gupta (2014b); Hasan et al., (2014); Syed et al., (2010, 2011); Joshi et al., (2010); Nivedhitha et al., (2016); Pandey and Govilkar (2015); Sharma et al., (2015); Sharma and Bhattacharyya (2014); Naidu et al., (2017); Rehman and Bajwa (2016); Jha et al., (2015); Sharma et al., (2014); Mittal et al., (2013); Sharma and Moh (2016); Deshmukh et al., (2017); Asghar et al., (2019); Mukhtar and Khan (2019); Garg and Buttar (2017); Hussaini et al., (2018); Yadav and Bhojane (2019); Rai et al., (2017); Amin et al., (2019)

Techniques (Count)	Languages (Count)	References
SVM (28)	Hindi (9), Bengali (5), Oriya (2), Tamil (2), Nepali (1), Telugu (1), Kannada (1), Marathi (1), Urdu (1), Sindhi (1), Punjabi (2), Gujarati (1), Malayalam (1)	Jena and Chandra (2014); Sahu et al., (2016b); Thapa and Bal (2016); Mukku et al., (2016); Kumar et al., (2015b); Phani et al., (2016); Ghosal et al., (2015); Joshi et al., (2010); Venugopalan and Gupta (2015); Kumar et al., (2015a); Akhtar et al., (2016b, a); Balamurali et al., (2012); Mukhtar and Khan (2017); Se et al., (2016); Sharma and Moh (2016); Ali and Wagon (2017); Kaur and Kaur (2017); Kaur and Gupta (2017); Joshi and Vekaria (2017); Kumar et al., (2019); Sharmista and Ramaswami (2018); Nanda et al., (2018); Rai et al., (2017); Sarkar and Bhowmick (2017); Hoque et al., (2019)
NB (16)	Hindi (4), Kannada (3), Oriya (2), Bengali (2), Tamil (2), Nepali (1), Telugu (1), Urdu (1)	Sahu et al., (2016a, b); Gupta and Bal (2015); Mukku et al., (2016); Hegde and Padma (2015); Deepamala and Kumar (2015); Kumar et al., (2015b); Ghosal et al., (2015); Se et al., (2015); Akhtar et al., (2016a); Mukhtar and Khan (2017); Se et al., (2016); Jha et al., (2015); Sharma and Moh (2016); Rai et al., (2017)
DT (14)	Hindi (4), Kannada (2), Bengali (3), Tamil (2), Tamil (1), Urdu (1), Telugu (1)	Mukku et al., (2016); Kumar et al., (2015b); Phani et al. (2016); Ghosal et al., (2015); Prasad et al., (2015); Venugopalan and Gupta (2015); Akhtar et al., (2016a); Rohini et al., (2016); Mukhtar and Khan (2017); Sharmista and Ramaswami (2016); Se et al. (2016); Hoque et al., (2019)

Techniques (Count)	Languages (Count)	References
Deep learning (14)	Bengali (5), Hindi (3), Tamil (2), Punjabi (1), Telugu (1), Malayalam (2)	Hassan et al., (2016); Seshadri et al., (2016); Bansal et al., (2013); Singh et al., (2018); Kumar et al., (2017); Shalini et al., (2018); Kumar et al., (2019); Padmamala and Prema (2017); Rani and Kumar (2018); Sumit et al., (2018); Tripto and Ali (2018); Sarkar (2019)
Logistic Regression (9)	Oriya (1), Nepali (1), Telugu (2), Bengali (2), Hindi (1), Tamil (1), Urdu (1)	Sahu et al., (2016b); Thapa and Bal (2016); Mukku et al., (2016); Mukku and Mamidi (2017); Phani et al., (2016); Khan et al., (2017); Hoque et al., (2019)
Random Forest (9)	Bengali (2), Tamil (2), Telugu (2), Hindi (2), Kannada (1)	Mukku et al., (2016); Phani et al., (2016); Ghosal et al., (2015); Hegde and Padma (2017); Sharmista and Ramaswami (2016); Mukku and Mamidi (2017); Nanda et al., (2018)
MNB (8)	Hindi (2), Bengali (4), Tamil (1), Nepali (1)	Thapa and Bal (2016); Phani et al., (2016); Sarkar and Chakraborty (2015); Sarkar (2018); Sarkar and Bhowmick (2017)
Hybrid (6)	Hindi (1), Punjabi (1), Bengali (2), Malayalam (2)	Akhtar et al., (2016c); Nair et al., (2015); Jayan et al., (2015); Kaur and Gupta (2014a); Das and Bandyopadhyay (2010a); Al-Amin et al., (2017)

Techniques (Count)	Languages (Count)	References
k-NN (3)	Bengali (2), Sindhi (1)	Ghosal et al., (2015); Ali and Wagon (2017); Hoque et al., (2019)
Regularized Least Square (3)	Tamil (1), Bengali (1), Malayalam (1)	Kumar et al., (2015c); Kumar et al., (2019)
ME (2)	Kannada (1), Tamil (1)	Deepamala and Kumar (2015); Se et al., (2016)
Neural Network (2)	Telugu (1), Hindi (1)	Mukku et al., (2016); Yadav and Bhojane (2019)
CRF (2)	Manipuri (2)	Nongmeikapam et al., (2014)
Scoring based (2)	Punjabi (1), Gujarati(1)	Arora and Kaur (2015); Gohil and Patel (2019)
NLP based (1)	Marathi (1)	Chaudhari et al., (2017)
Fuzzy logic (1)	Malayalam (1)	Anagha et al., (2015)

From Table 2.8, it can be stated that lexicon-based approach has been experimented over almost all Indian languages which conclude that researchers first experimented with SA in their own language by constructing polarity lexicons, while the majority of the researchers have used ML approaches followed by lexicon-based, deep learning and hybrid as shown in Figure 2.13.

It has been observed that the majority of the research studies (i.e., 60%) opted for ML to perform SA. However, in recent times, researchers are also attracting towards experimenting with deep learning techniques due to improvement in accuracy irrespective of time constraint that is needed to train the data.

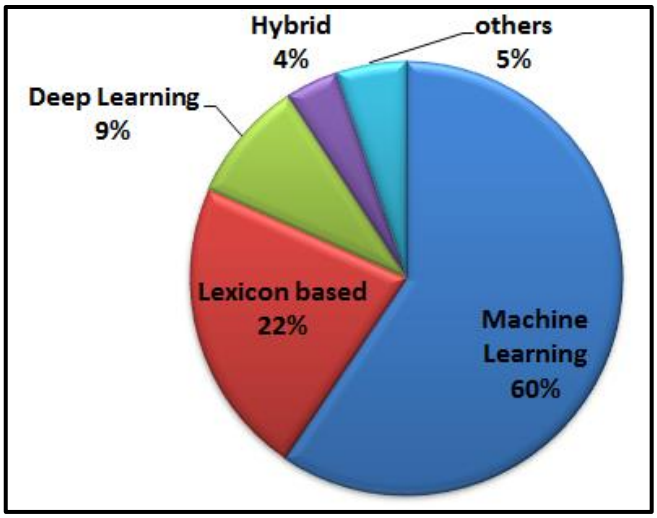


Figure 2.13 Percentage of research work using different SA techniques

Out of ML approaches researchers mostly used SVM, NB and DT as shown in Figure 2.14 covering approximately 70% of the research studies for the development of SA systems.

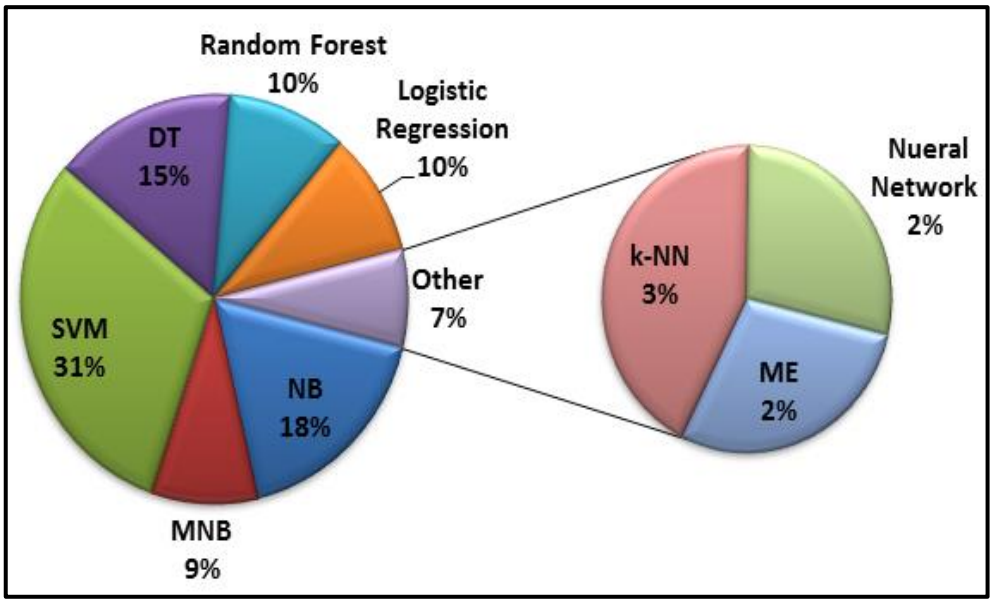


Figure 2.14 Percentage of research work using different ML techniques

Figure 2.15 shows that in case of lexicon-based techniques, 60% of the researchers have created their own SWN using bi-lingual dictionary-based approach, i.e., the authors created lexicon by performing translation of entries of English SWN to their own

language and 25% of the researchers started with some seed list of polarity words and used WordNet based approach to extend the sentiment lexicon while remaining 15% of the researchers manually developed SWN for their own language means that they manually assigned sentiment polarity to words.

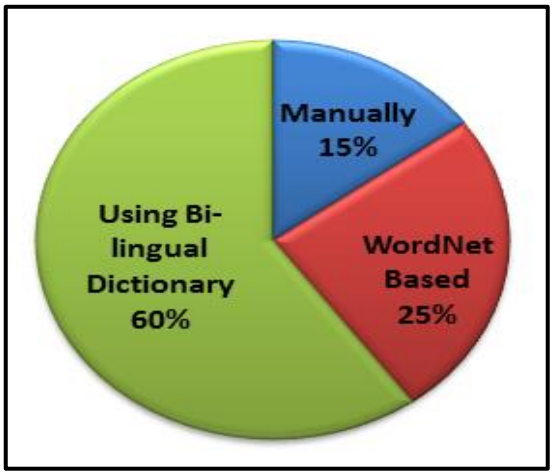


Figure 2.15 Percentage of research work using Lexicon-based techniques

The answer to research question **RQ5** has been addressed through Figure 2.16 which depicts the domains considered by the researchers to perform SA for different Indian languages.

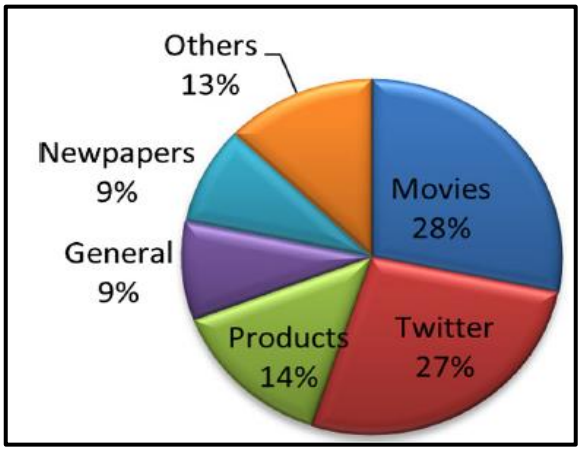


Figure 2.16 Percentage of SA work for different domains

It has been observed that mainly the research work has been performed on movie reviews (i.e., 28%) and tweets (i.e., 27%). The reason behind the majority of research work on movie reviews is due to the availability of annotated dataset. The 14% of the research

work has been performed on products reviews dataset, 13% of the research work covers blogs, websites, etc., 9% of the research work includes general text and remaining 9% of the research work covers newspapers.

Table 2.9 summarizes the research studies according to the different sentiment levels such as aspect, sentence, and document; classes namely positive, negative and neutral; and whether the SA system handles negations or not. It is analyzed that approximate 30% of the research studies have handled negations which also plays a major role in improving sentiment classification.

Table 2.9: Classification of Indian languages according to different parameters

Language[Ref]	**A**	**S**	**D**	**P**	**N**	**O**	**Handling of Negations**
Odia (Sahu et al., 2016a; Akhtar et al., 2016c), Nepali (Gupta and Bal 2015; Mukku et al., 2016), Marathi (Balamurali et al., 2012), Telugu (Naidu et al., 2017; Shalini et al., 2018), Urdu (Syed et al., 2014; Khan et al., 2017; Asghar et al., 2019), Tamil (Se et al., 2016), Bengali (Ghosal et al., 2015; Hassan et al., 2016; Al-Amin et al., 2017; Hoque et al., 2019), Hindi (Joshi et al., 2010; Balamurali et al.. 2012; Bansal et al., 2013; Nanda et al., 2018; Hussaini et al., 2018), Punjabi (Kaur and Kaur, 2017), Gujarati (Joshi and Vekaria, 2017; Gohil and Patel, 2019)	-	+	-	+	+	-	-
Marathi (Chaudhari et al., 2017), Malayalam (Anagha et al., 2014), Bengali (Hasan et al., 2014)	-	+	+	+	+	+	-
Malayalam (Jayan et al., 2015)	+	+	+	+	+	+	-

Language[Ref]	A	S	D	P	N	O	Handling of Negations
Telugu (Mukku et al., 2016; Mukku and Mamidi, 2017), Malayalam (Anagha et al., 2015; Kumar et al., 2017; Kumar et al., 2019), Urdu (Rehman and Bajwa 2016; Mukhtar and Khan, 2017), Tamil (Kumar et al., 2015c; Nivedhitha et al., 2016; Sharmista and Ramaswami, 2016; Se et al., 2016; Sharmista and Ramaswami 2018), Bengali (Phani et al., 2016; Kumar et al., 2015a, c; Sarkar and Chakraborty 2015; Se et al., 2015; Seshadri et al., 2016; Sarkar and Bhowmick, 2017; Sumit et al., 2018; Sarkar, 2018; Tripto and Ali, 2018; Sarkar, 2019), Hindi (Phani et al., 2016; Prasad et al., 2015; Venugopalan and Gupta 2015; Kumar et al., 2015c, a; Sarkar and Chakraborty 2015; Se et al., 2015; Seshadri et al., 2016; Rani and Kumar, 2018), Sindhi (Ali and Wagon, 2017), Marathi (Deshmukh et al., 2017)	-	+	-	+	+	+	-
Punjabi (Kaur and Gupta, 2014b; Kaur and Gupta, 2017), Malayalam (Nair et al., 2014), Hindi (Bakliwal et al., 2012; Pandey and Govilkar, 2015; Sharma et al., 2015, 2014; Arora, 2013; Sharma and Moh, 2016; Rai et al., 2017), Urdu (Mukhtar and Khan, 2019), Bengali (Amin et al., 2019)	-	+	-	+	+	+	+
Malayalam (Nair et al., 2015), Hindi	-	+	-	+	+	-	+

Language[Ref]	**A**	**S**	**D**	**P**	**N**	**O**	**Handling of Negations**
(Sharma and Bhattacharyya, 2014), Tamil (Padmamala and Prema 2017)							
Oriya (Jena and Chandra, 2014), Manipuri (Nongmeikapam et al., 2014)	-	-	+	+	+	+	-
Kannada (Deepamala and Kumar, 2015), Urdu (Syed et al., 2010)	-	+	+	+	+	-	+
Punjabi (Kaur and Gupta, 2014a), Kannada (Kumar et al., 2015b), Malayalam (Ashna and Sunny, 2017)	-	+	+	+	+	+	+
Malayalam (Thulasi and Usha, 2016), Hindi (Akhtar et al., 2016a, b)	+	-	-	+	+	+	-
Kannada (Hegde and Padma, 2015, 2017; Rohini et al., 2016)	+	+	-	+	+	-	-
Konkani (Miranda and Mascarenhas, 2016), Hindi (Yadav and Bhojane, 2019)	+	+	+	+	+	+	+
Punjabi (Arora and Kaur, 2015)	-	+	+	+	+	-	-
Bengali (Das and Bandyopadhyay, 2010a)	+	-	-	+	+	-	+
Hindi (Mittal et al., 2013)	-	-	+	+	+	-	+
Hindi (Jha et al., 2015)	-	-	+	+	+	+	+
Urdu (Syed et al., 2011)	+	+	-	+	+	-	+
Hindi (Akhtar et al., 2016c)	+	+	-	+	+	+	-
Hindi (Garg and Buttar, 2017)	+	-	-	+	+	+	+

**A: Aspect, S: Sentence, D: Document, P: Positive, N: Negative, O: Neutral, '+' indicates that corresponding language supports the corresponding factor that is used for SA

Table 2.9 also assists in concluding that mostly the researchers (i.e., 72%) have worked at the sentence level, 18% have worked on document level and the remaining 10% of the

researchers cover the SA work on aspect-level as shown in Figure 2.17(a) which states that the research work on aspect level is still in the growing phase. As the SA at aspect-level helps in performing fine-grained analysis, therefore, the researchers are attracting towards it. As shown in Figure 2.17(b), majority of researchers (i.e., 64%) have considered have only 2 classes, i.e., positive and negative for SA of Indian languages and the remaining 36% of the researchers have worked on all the three classes, i.e., positive, negative and neutral. The reason behind the consideration of 2 classes for the SA process is that better accuracy can be achieved for 2 classes in comparison to three classes. Figure 2.17(a) and Figure 2.17(b) help to find the answer for research question **RQ6.**

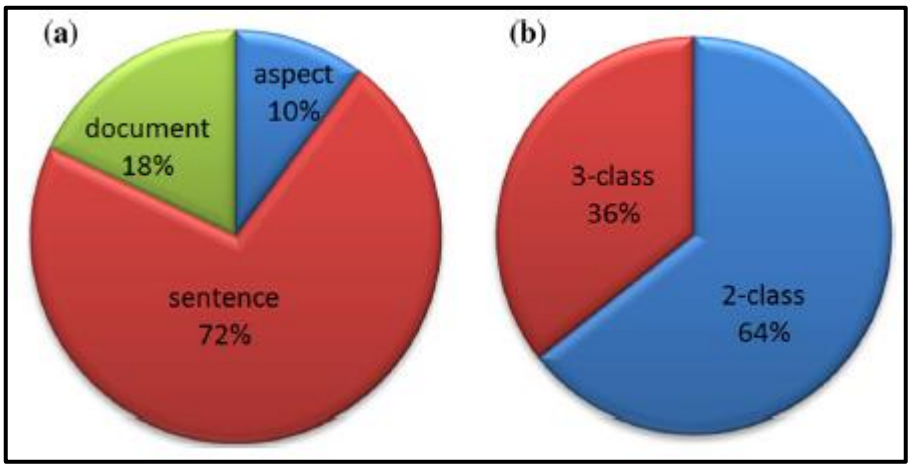


Figure 2.17 Percentage of work a) at different sentiment levels, b) for different sentiment classes

To address the research question **RQ7**, the online available SA tools such as Alchemy API, Semantria, Trackur, Sentigem, etc. have been explored, it has been identified that no such system is available online till now which can perform SA of Indian languages. The answer to research question **RQ8** is addressed as follows.

The major findings of the research questions from this systematic survey can be summarized as follows.

- The research work in the field of SA for Indian languages started in 2010 when Joshi et al. (2010) first set the benchmark by performing SA for the Hindi

language. This research study has the highest impact in the field of SA over Indian languages till now as it has more than a hundred citations.

- After 2010, it has been observed that SA research work has been performed in 15 Indian languages such as Hindi, Bengali, Tamil, Malayalam, Kannada, Urdu, Punjabi, Oriya, Nepali, Telugu, Konkani, Manipuri, Marathi, Gujarati and Sindhi out of 22 languages.
- This research study provides a brief description of the different SA techniques such as ML, lexicon-based, and deep learning.
- This systematic review helps in providing knowledge about the online availability of annotated datasets, linguistics resources and polarity lexicons for different Indian languages.
- The annotated datasets are available for Hindi, Bengali, Tamil, and Marathi. Similarly, linguistic resources such as Morph analyzer, POS tagger, dependency parser are available for different Indian languages. Researchers can easily use these resources as a description along with online availability is provided in this systematic review.
- From this survey, it has been observed that majority of research work in the field of SA has been performed for Indo-Aryan language, i.e., Hindi that covers approximately one-third of the research work performed for Indian languages.
- It has been analyzed that mostly the researchers have used ML techniques, however, the researchers are also attracting towards deep learning techniques due to better accuracy achieved by these techniques.
- From this systematic survey, it can be observed that researchers have performed mostly SA work on sentence-level for positive and negative sentiment classes. Also, mainly the authors have experimented on movie reviews dataset and tweets.

Thus, this systematic and comprehensive survey provides a detailed description about online available annotated datasets, pre-processing linguistic resources, SA techniques for different Indian languages which can be used by researchers to further explore the area of sentiment analysis in their own native languages.

Chapter Summary

The growth of research work in the field of SA for Indian language content motivated us to conduct this systematic survey. In India, there are 22 official languages and due to availability of data from multiple sources for each language, it is easy to gather data and analyze them. The research work on SA in context to Indian languages was first commenced by Joshi et al., (2010), the highest cited research study till now. Afterward it, the research work in this field is continuously growing from the last couple of years as Indian language content on the web is also increasing. Till now, no research study is available which covers an in-depth analysis for Indian languages in the field of SA. Therefore, this chapter is a significant contribution in the literature of SA for Indian languages which includes the systematic survey over 90 research studies published on SA for all Indian language families from 2010 to 2019 (till end of July). The 90 research studies considered in this systematic survey have been decided by developing a review protocol which includes the research questions, sources of information, inclusion and exclusion criteria. The different findings of this survey have been analyzed to get the answers of the targeted research questions framed in this chapter.

The summary about the different SA approaches, type, and size of corpora, lexical resources/tools and evaluation measures for each Indian language is given in this chapter. From this summary, it has been analyzed that SA work has been reported on 15 Indian languages and the majority of the work in this field has been published in conferences followed by journals. It has been observed from the comprehensive analysis that 70% of the research work has been done for Indo-Aryan language family in which major part is covered by Hindi and Bengali language (i.e., 45%). It has also been noticed that the researchers have mainly used ML (i.e., 60%) approach in comparison to other lexicon-based, deep learning and hybrid approaches. Also, the researchers have performed mainly SA work at the sentence level and considered two sentiment classes, i.e., positive and negative in the majority of research studies using different domains like tweets, movies, and products reviews, etc. This chapter also gives the details about online available annotated datasets, pre-processing linguistic resources available for different Indian languages which can help the researchers to perform SA in other Indian languages. The

online available SWN(s) for various Indian languages and the approaches to develop them are also discussed in this chapter.

Chapter 3

Implementation of Sentence Based Sentiment Analysis System

This chapter presents the implementation of SA system for Hindi language which performs the SA at sentence level. The sentences are classified into three classes, i.e., positive, negative and neutral. For experimentation, corpus of Hindi sentences of movie reviews and tweets has been collected from reviews websites and Twitter respectively. After annotation of corpus, experments have been perfoemed using ML techniques. The next section presents the architecture of ML based SA system in detail.

3.1 Architecture of the Sentence based SA system

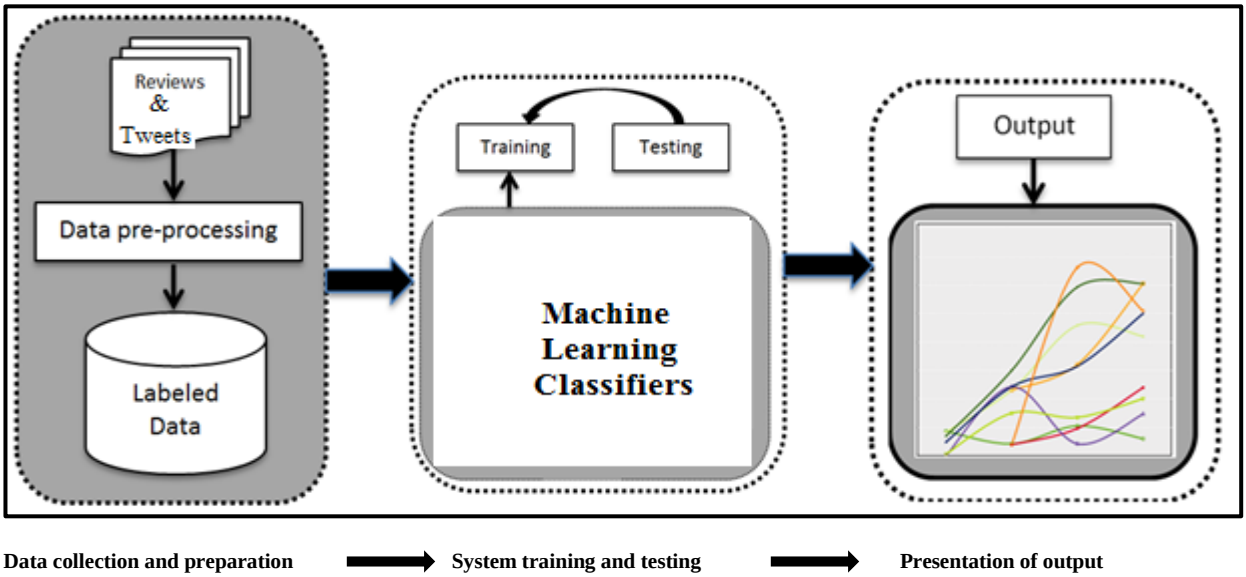


Figure 3.1 Architecture of the sentence based SA system

The architecture of the sentence based sentiment analysis system consists of three phases, i.e., Corpus Collection and Preparation, System Training and Testing, and Presentation of Output. The detailed description about these phases is given as follows.

3.1.1 Corpus Collection and Preparation

First of all, corpus of movie reviews and tweets is collected. The movie reviews have been extracted from फिल्म-समीक्षा *philm-sameeksha* '*filmreview*' section of aajtak (http://aajtak.intoday.in/film-review.html) and jagran

(http://www.jagran.com/entertainment/reviews-news-hindi.html) online newspapers. To extract the reviews, a Graphical User Interface (GUI) has been developed in python language using java-based '*boilerpipe*' library. The library extracts the required text from a web page using '*ArticleExtractor*' parameter and saves it in '*UTF-8'* format with '*.txt'* extension. The corpus of tweets has been collected by extracting the Twitter posts using Twitter API. Some sample sentences of the corpus are given in Table 3.1.

Table 3.1: Example sentences of corpus

Sr. No.	**Sentence**
1.	Negative Hindi Sentence: कहानी को ठीक से समेटा नहीं गया है । Transliteration: *kahaanee ko theek se sameta nahin gaya hai.* English Translation: '*The story has not been properly compiled.*'
2.	Positive Hindi Sentence: आमिर बशीर स्टेज के बेहतरीन कलाकार हैं । Transliteration: *aamir basheer stej ke behatareen kalaakaar hain.* English Translation: '*Amir is the best actor of Bashir Stage*.'
3.	Neutral Hindi Sentence: उन्हें रास्ते में सैमुअल जैक्सन मिलते हैं । Transliteration: *unhen raaste mein saimual jaiksan milate hain.* English Translation: '*They get Samuel Jackson on the way*.'
4.	Positive Hindi Sentence: फिल्म का इशारा बिलकुल सही है । Transliteration: *philm ka ishaara bilakul sahee hai.* English Translation: '*The gesture of the film is absolutely correct*.'
5.	Neutral Hindi Sentence: अजय देवगन एक सीनियर ऐक्टर और स्टार हैं । Transliteration: *ajay devagan ek seeniyar aiktar aur staar hain.* English Translation: '*Ajay Devgan is a senior actor and star*.'
6.	Negative Hindi Sentence: फिल्म में कई जगह डबिंग सही नहीं है । Transliteration: *philm mein kaee jagah dabing sahee nahin hai.* English Translation: *'Dubbing is not correct in many places in the film.*'
7.	Positive Hindi Sentence: तब्बू ने इस फिल्म में बेहतरीन अभिनय किया है । Translitertaion: *tabboo ne is philm mein behatareen abhinay kiya hai.* English Translation: '*Tabu has done an excellent performance in this film*.'
8.	Neutral Hindi Sentence: फिल्म गौरी के जेल से छूटने के समय आरंभ होती है। Transliteration: *philm gauree ke jel se chhootane ke samay aarambh hotee hai.* English Translation: '*The film begins at the time Gauri is released from prison*.'
9.	Negative Hindi Sentence: यह बेस्ट फिल्म नहीं है । Transliteration: *yah best philm nahin hai.* English Translation: '*This is not the best movie*.'

10.	Positive Hindi Sentence: फुवाद खान का यह प्रयास सराहनीय है । Transliteration: *phuvaad khaan ka yah prayaas saraahaneey hai.* English Translation: '*Fuwad Khan's effort is commendable*.'

Pre-processing of Corpus: The collected corpus is preprocessed by removing irrelevant data in order to generate a rich set of features. Transliteration of romanized text into Hindi text is performed using Google API. The punctuations, URLs, etc., have been removed using regular expressions of Python. To handle the issue of abbreviations, a mapping dictionary consisting of approximately 100 abbreviations has been prepared containing a mapping of abbreviations to its full form as given in Tale 3.2.

Table 3.2: Mapping dictionary of abbreviations

Abbreviation	**Full Form**
यूपी	उत्तर प्रदेश
पाक	पाकिस्तान
इसरो	भारतीय अंतरिक्ष अनुसंधान संगठन
डा	डॉक्टर
भाजपा	भारतीय जनता पार्टी
किमी	किलोमीटर
राजग	राष्ट्रीय जनतांत्रिक गठबंधन
आईआईटी	भारतीय प्रौद्योगिकी संस्थान
एमटीएनएल	महानगर टेलीफोन निगम लिमिटेड
बीएसएनएल	भारत संचार निगम लिमिटेड
बुध	बुधवार
आईटी	सूचान प्रौद्योगिकी
वीएसएनएल	विदेश संचार निगम लिमिटेड
आईसीयू	गहन ईकाई कक्ष
एसएमएस	लघु संदेश सेवा
आईएसओ	अंतर्राष्ट्रीय मानक संगठन
आईबीएम	अंतर्राष्ट्रीय व्यवसाय तंत्र
एलआईसी	जीवन बीमा निगम

Further, the pre-processed corpus is manually annotated by three native speakers of Hindi into three classes such as positive, negative and neutral. This corpus consists of 7354 movie reviews which include 2341 positive, 2037 negative and 2976 neutral sentences. The final polarity of the sentence is decided on the basis of maximum number of votes. The summary of corpus is given in Table 3.3.

Table 3.3: Corpus summary

Domain	**Polarity**	**Number of sentences**
Movie Reviews Corpus	Positive (P)	2,341
	Negative (N)	2,037
	Neutral(O)	2,976
Total		7,354
Tweets	Positive (P)	782
	Negative (N)	1,229
	Neutral(O)	1,262
Total		3,273

The corpus is evaluated by using Kappa (κ) statistical measure. Kappa (κ) is defined by two measures, *i.e.,* observed agreement (A_o) and chance agreement (A_e). A_o is the percentage of annotations on which both annotators agree. A_e is the percentage of chance of agreement by annotators according to their individual class distribution and is given by the formula shown in (3.1). In this formula, c_A and c_B represent the chances of agreement of annotators A and B on class k mean that these describe the number of agreements of independent annotators A and B that would have been expected by chance for class k.

$$A_e = \sum_{k \in K} P(c_A \mid k).(c_B \mid k) \tag{3.1}$$

The confusion matrices for the inter-annotator agreement between each pair of the three annotators A1, A2 and A3 are given in Table 3.4(a), 3.4(b) and 3.4(c) respectively, and kappa scores for inter-annotator agreement of the three annotators is given in Table 3.5.

Table 3.4(a): Confusion matrix for annotators A1 and A2

A1/A2	P	N	O	Total
P	0.296	0.003	0.025	0.324
N	0.010	0.303	0.018	0.331
O	0.028	0.019	0.298	0.345
Total	0.334	0.325	0.341	1

Table 3.4(b): Confusion matrix for annotators A2 and A3

A2/A3	P	N	O	Total
P	0.280	0.008	0.032	0.320
N	0.003	0.295	0.026	0.324
O	0.040	0.012	0.304	0.356
Total	0.323	0.315	0.362	1.0

Table 3.4(c): Confusion matrix for annotators A1 and A3

	P	N	O	Total
P	0.275	0.003	0.017	0.295
N	0.003	0.293	0.075	0.371
O	0.013	0.034	0.287	0.334
Total	0.291	0.330	0.379	1.0

The formula of statistical measure kappa (κ) is given in (3.2).

$$\kappa = \frac{(A_o - A_e)}{(1 - A_e)} \tag{3.2}$$

Table 3.5: kappa scores for inter-annotator agreement

Kappa (κ) score between annotator 'i' and annotator 'j'	**Observed Agreement**	**Chance Agreement**	**kappa score**
κ_{12}	0.897	0.333	0.84
κ_{13}	0.855	0.334	0.82
κ_{23}	0.879	0.334	0.82
Average κ	0.877	0.334	0.83

In general, kappa values between 0.6 and 0.8 are considered a substantial agreement. In our case, we got the kappa value to be 0.83, which is in good agreement and is an indication of the reliability of the annotations.

3.1.2 System Training and Testing

In this phase, different machine learning algorithms are trained on the annotated corpus using different parameter setting and classified into positive, negative and neutral class. The detailed description about the parameter settings and results has been explained in the next sections.

3.1.3 Presentation of Output

The output results are presented by drawing line charts, and these charts help in performing the comparative analysis between different models of ML. The different parameter settings for conducting experimentation are given in the next section.

3.2 Tools Used

Jupyter Notebook (an open-source Web application) (Jup, 2017) has been used as development environment for performing experiments. It helps in creating and sharing documents consisting of code, text, equations and visualizations. It also includes machine learning, data cleaning, transformation and statistical modeling, etc. The CNN models have been developed using Python package TFLearn (a deep learning library) (TFL, 2017) built on the top of the TensorFlow (an open-source software library for numerical computations) (Ten, 2017) that speeds up the development of deep learning models. All strings in the sentences are transformed to list of sequences using vocabulary processor of TFLearn as neural networks do not handle strings.

3.3 Experimentation using Traditional ML Algorithms

The experiment using different machine learning algorithms has been performed on a corpus of 3273 tweets and 7,354 movie reviews. The corpus of tweets includes 782 positive, 1229 negative and 1262 neutral tweets. The corpus of movie reviews includes 2341 positive, 2037 negative and 2976 neutral sentences. The experimental results of both

the corpra have been analyzed using the performance measures precision, recall, F-measure and accuracy. Table 3.6 and Table 3.7 depict the experimental results of ML algorithms on corpus of tweets and movie reviews, respectively.

Table 3.6: Overall accuracy statistics of experimental results for tweets

Algorithm	**Precision**	**Recall**	**F-measure**	**Accuracy**
Naïve Bayes	0.89469	0.88948	0.88916	0.88948
Multinomial Naïve Bayes	0.74646	0.74590	0.74569	0.74590
Bernoulli Naïve Bayes	0.74894	0.74783	0.74803	0.74783
k-Nearest Neighbor	0.65883	0.61523	0.60850	0.61523
Support Vector Machines	0.64638	0.42912	0.36171	0.42912
Decision Tree	0.61776	0.50309	0.48439	0.50309
Random forest	0.71167	0.64451	0.63852	0.64451
AdaBoost	0.61053	0.58630	0.58718	0.61213
Gradient Boosting	0.72852	0.70154	0.70307	0.70154

Figure 3.2 represents the comparison of precision, recall and F-measure of ML algorithms on corpus of tweets.

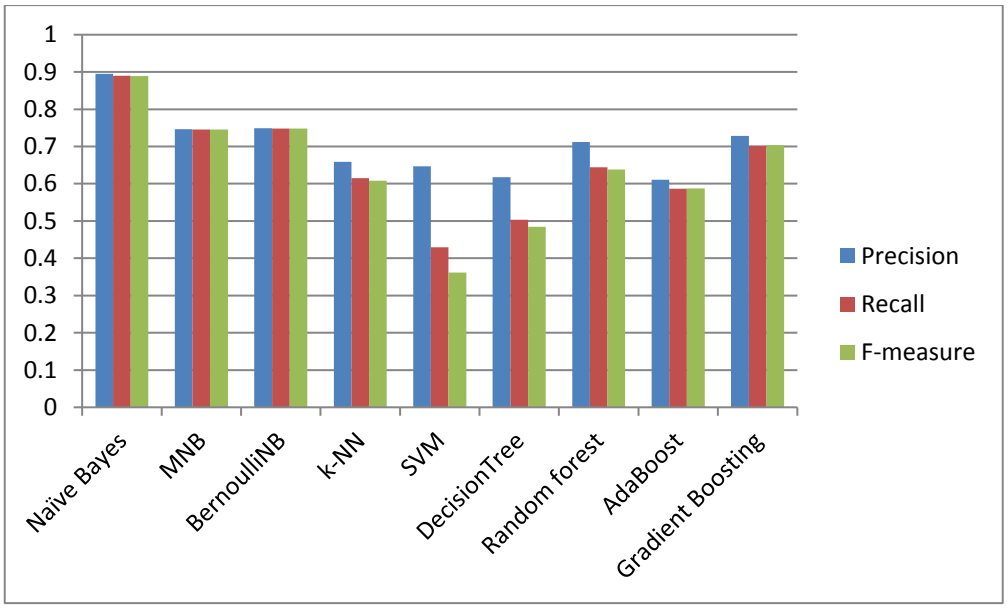


Figure 3.2 Comparison of performance measures of ML algorithms for tweets

Figure 3.3 comparison of accuracy of all the ML algorithms and it shows that NB performs the best and Bernouli NB is the second best out of all other traditional ML algorithms for a corpus of tweets in Hindi language.

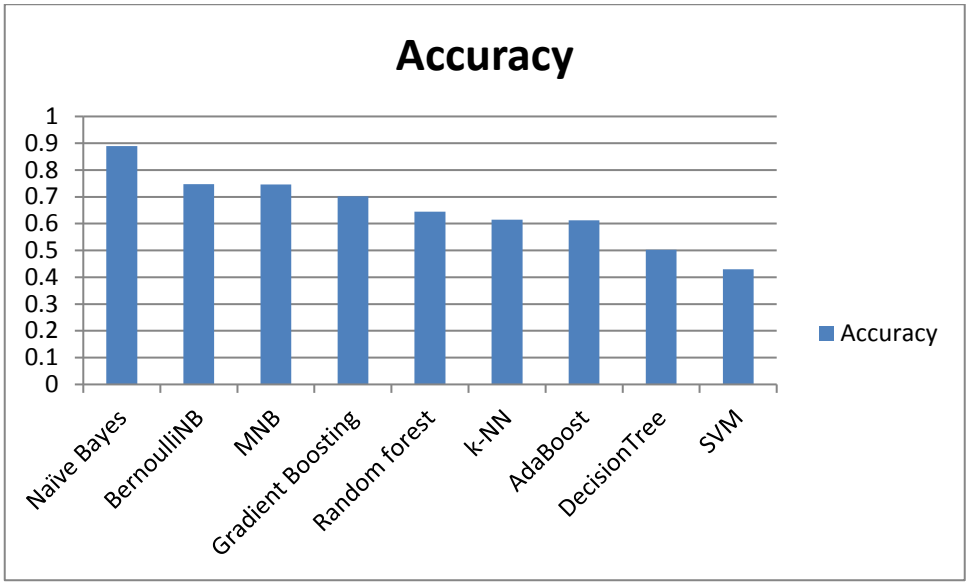


Figure 3.3 Comparison of accuracy of ML algorithms for tweets

Table 3.7: Overall accuracy statistics of experimental results for movie reviews

Algorithm	**Precision**	**Recall**	**F-measure**	**Accuracy**
Naïve Bayes	0.88491	0.88095	0.87809	0.88095
Multinomial Naïve Bayes	0.69036	0.69047	0.68997	0.69047
Bernoulli Naïve Bayes	0.68531	0.67687	0.67921	0.67687
k-Nearest Neighbor	0.74028	0.68027	0.66794	0.68027
Support Vector Machines	0.67831	0.33673	0.25285	0.33673
Decision Tree	0.70331	0.68547	0.68660	0.68547
Random forests	0.72788	0.72797	0.72765	0.72797
AdaBoost	0.62381	0.56462	0.57639	0.56462
Gradient Boosting	0.71778	0.65306	0.66197	0.65306

Similarly, Figure 3.4 and Figure 3.5 represent the comparison of precision, recall, F-measure and accuracy, respectively given by ML algorithms on corpus of movie reviews in Hindi language.

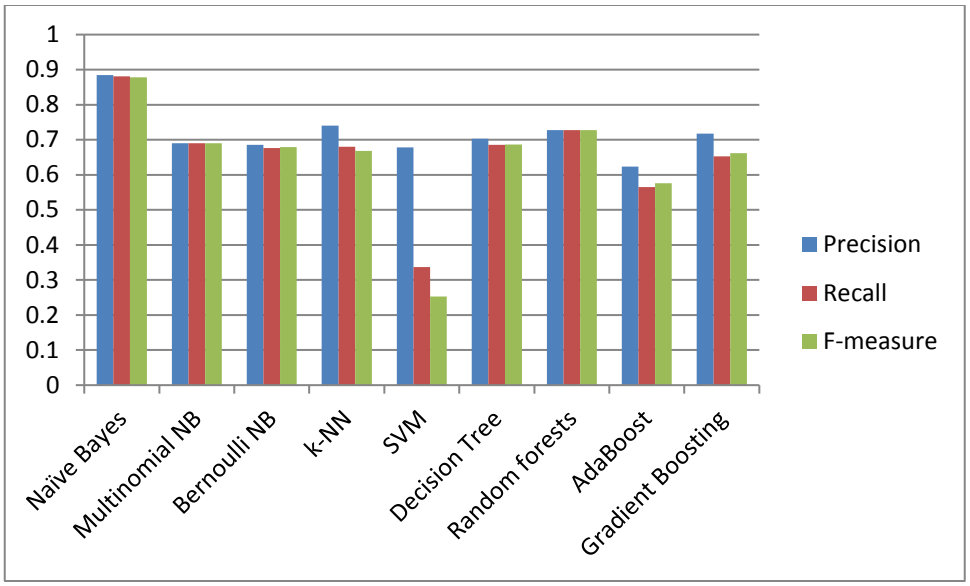


Figure 3.4 Comparison of performance measures of ML algorithms for movie reviews

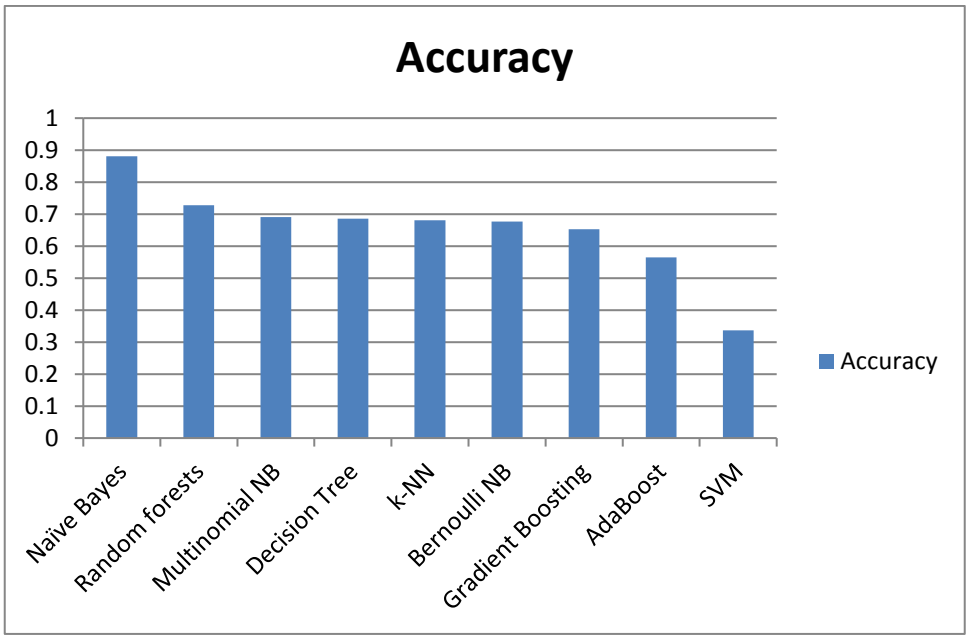


Figure 3.5 Comparison of accuracy of ML algorithms for movie reviews

From the results, it has been analyzed that NB performs the best and Random Forest is the second best out of all traditional ML algorithms for the corpus of movie reviews. In the next section, the deep learning based technique has been experimented using different parameters on the same corpus of Hindi movie reviews.

3.4 Introduction to CNN

CNN is a class of deep, feed-forward artificial neural networks and use a variation of multilayer perceptrons designed to require minimal preprocessing. CNNs are generally used in computer vision; however they have recently been applied to various NLP tasks such as sentiment analysis, machine translation, question answering systems etc. The CNN model consists of four layers, i.e., input layer, convolution layer, global max pool layer and fully connected layer. These layers are briefly described as follows.

a) Input Layer

A neural network requires word embedding as an input to the CNN. For example, in the case of movie reviews, the length of the sentence varies. Before feeding a review into a neural network as input, each word of the sentence is converted into a numerical value. This process is called word encoding or tokenization. In this encoding process, each unique word of the corpus is recorded as vocabulary of the model. The reviews are considered as a sequence of words where each word is signified by a vector $v \in \mathbb{R}^{1\times d}$, known as word embedding. Here, d is the dimension and $d \approx |V|$, the vocabulary size V. Each vocabulary word is encoded as a unique integer, called a token. These tokens are assigned based on the frequency of occurrence of a word in the corpus. The word that appears most frequently throughout the corpus, will have the associated token: 0. For example, if the most common word is "the", it would have the associated token value of 0. Then the next most common word will be tokenized as 1, and that process continues and a form of token dictionary is generated as shown in Figure 3.6.

```
{'the': 0, 'of': 1, 'so': 2, 'then': 3, 'you': 4, … }
```

Figure 3.6 Token dictionary

A common encoding step is to one-hot encode each token; representing each word as a vector that has as many values in it as there are words in the vocabulary. That is, each column in a vector represents one possible word in a vocabulary. The vector is filled with

0's except for the index at that word's token value, say index 0 for "the" as shown in Figure 3.7.

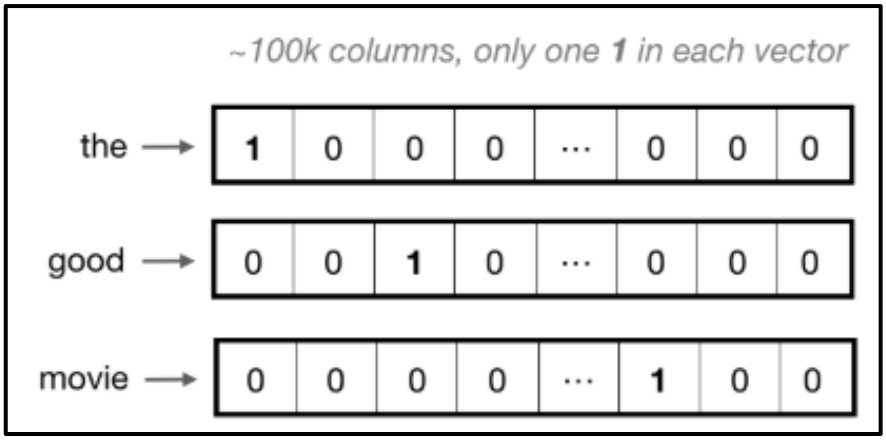


Figure 3.7 One-hot encoding

Another vector representeation is known as embedding. Word embeddings are vectors of a specified length, typically on the order of 100, and each vector of 100 or so values, represents one word. The values in each column represent the features of a word, rather than any specific word as shown in Figure 3.8.

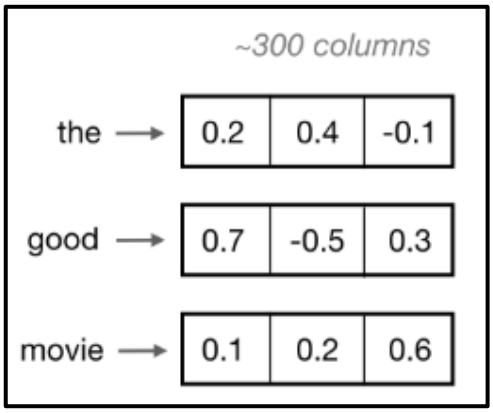


Figure 3.8 *Word2Vec* embedding

b) Convolution Layer

Convolutional layers are designed to find spatial patterns in an image by sliding a small kernel window over an image. In the case of text classification, a convolutional kernel is sliding window to look at multiple word embeddings in a sequence. The height of the kernel is the number of embeddings it will see at once, similar to representing an *N*-gram

in a word model. The width of the kernel should span the length of an entire word embedding as shown in following Figure 3.9.

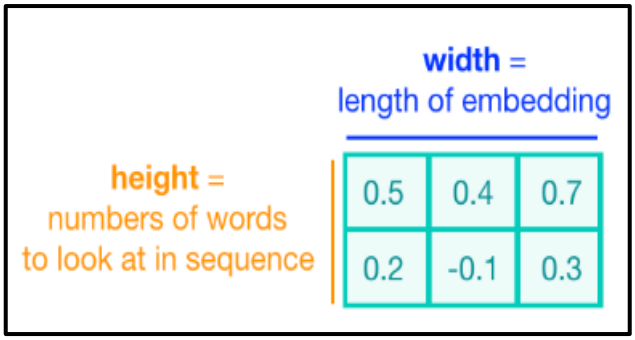


Figure 3.9 Example of convolutional kernel

Thus, in this layer, a set of *m* filters are applied to a sliding window of length *h* over each sentence. These filters are applied to every possible window of words in the sentence, and a feature c_i is generated. Each filter has its own separate bias. These *m* filters working in parallel generate multiple feature maps.

This convolutional operation has a property that similar words will have similar embeddings. So, when a convolutional kernel is applied to different sets of similar words, it will produce a similar output value. For example, the convolutional output value for the input 2-grams "good movie" and "fantastic song" are about the same because the word embeddings for those pairs of words are also very similar as shown in Figure 3.10.

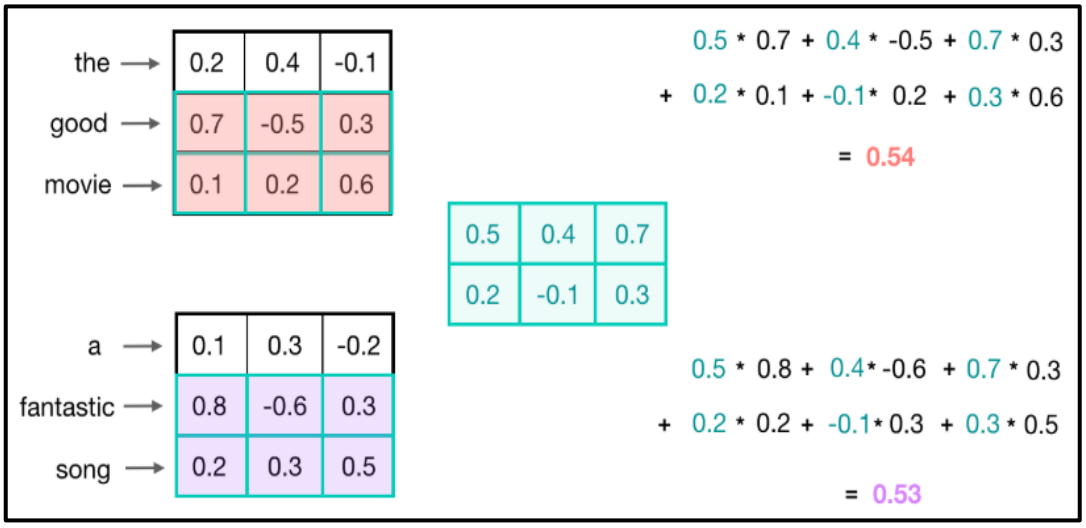


Figure 3.10 Example of convolution operation

In this example, the convolutional kernel has learned to capture a more general feature; not just a good movie or song, but a positive thing. A model can use those general features to classify entire texts.

c) Pooling Layer

The pooling layer samples the feature map generated by the convolution layer and the local optimum features. It consists of applying some operation over regions/patches in the input feature map and extracting some representative value for each of the analysed regions/patches. This layer aggregates the information and reduces the representation. Two of the most common pooling operations are max-pooling and average-pooling. Max-pooling selects the maximum of the values in the input feature map region of each step and average-pooling the average value of the values in the region. The output in each step is therefore a single scalar, resulting in significant size reduction in output size. Figure 3.11 shows the example of pooling operation with stride length of 2.

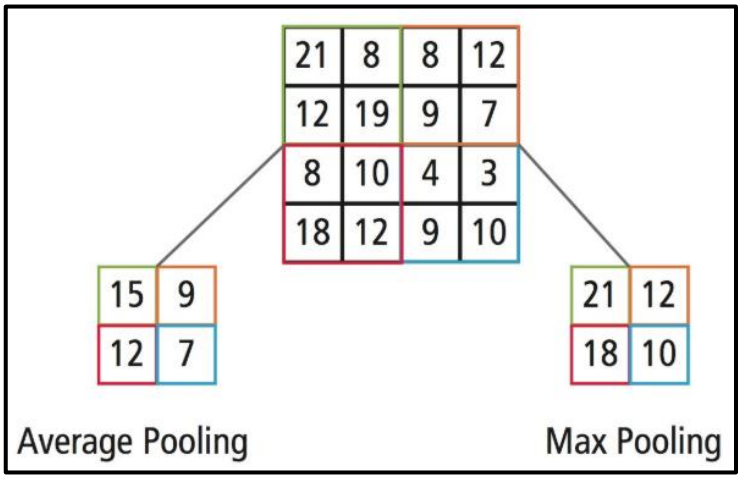


Figure 3.11 Example of pooling operation with stride length of 2

d) Fully Connected Layer

The fully-connected layer computes the transformation using the equation given in (3.3) where α is the Rectified Linear Unit (ReLU) activation function, $W \in \square^{m \times m}$ is the weight matrix, $b \in \square^{m}$ is the bias and C_{pool} is the feature map matrix generated by pooling layer.

$$x = \alpha(W * C_{pool} + b) \tag{3.3}$$

The output vector of this layer corresponds to the sentence embedding for each review. Finally the output of the previous layer is passed to a fully connected *softmax* layer. It returns the class *K* with the largest probability. The *softmax* layer returns the classification result then, the model parameters are updated by the back-propagation algorithm according to the actual classification label of the training data. Finally, each sentence gets three labels with values where one value represents the real label. For example, 'positive' = [0, 0, 1], 'negative' = [1, 0, 0] and 'neutral' = [0, 1, 0]. The different CNN models are built by experimenting with varying parameter settings such as number of convolution layers, filter size and number of filters.

3.4.1 Experiment Setup

For the proposed system, *word2vec* tool has been used which is able to capture the semantic properties of words in the corpus. The model has been trained on 50% of the corpus and this trained model is used for mapping a word into its respective vector representation. The high dimensional vectors are calculated for every word by calculating *softmax* probability for every word. The '*categorical_crossentropy'* loss function has been used in *softmax* layer as there are three classes to measure the error probability between the network prediction and real output label. The dimension of vector corresponds to the number of neurons in the hidden layer. The vector dimension of a word has been set to 100. Each sentence is padded with zero vectors in order to make its length uniform throughout the corpus. All the vectors are subsets of the word embedding matrix M consisting of all words in V. These words are mapped into indices $1 \ldots |V|$ to quickly lookup the vector of the word in M. Then for each review *x*, a sentence vector $X = \{w_1, w_2, \ldots, w_i, \ldots, w_{|x|}\}$,, has been built and $X \in \mathbb{R}^{d \times |x|}$, where w_i represents the word embedding at the corresponding position i in a sentence. Then X is fed to the convolutional neural network.

In this work, different CNN models have been built using several parameters for each layer. For the experimentation, the number of convolution layers has been taken either 2 or 3 and number of filters has been varied from 10 to 256. Also, the experiments have been conducted by varying filter sizes such as 3 × 3, 4 × 4, 5 × 5 and 7 × 7 to capture the different patterns in sequential group of 3, 4, 5 and 7 words that help in leraning different

relationships between words. The values of these parameters have been set by analyzing the studies conducted by other authors in this area (Svensson, 2017). The parameter settings used for CNN model such as vocabulary size, vector size, number of convolutional layers, hidden layers, fully-connected layers, number of filters, filter-size, activation function, regularizer, dropout, number of epochs and batch size are specified in Table 3.8.

The other parameters such as output dimension, regularizer, drop out, number of epochs and batch size has been fixed as change in these parameters have not shown in any improvement in accuracy of the model.

Table 3.8: Parameters settings of proposed CNN

Parameter	**Value**
Vocabulary size	13, 398
Input Vector Size	100
Number of convolutional layers	2, 3
Number of hidden layers	6,7
Activation Function	ReLU
Number of Filters	10, 50, 60, 100, 128, 256
Filter size	3, 4, 5, 7
Number of Fully connected layers	1
Output dimension	128
Regularizer	L2
Dropout	0.5
Number of epochs	5
Batch Size	64

In all models, the number of convolutional layers has been varied along with other parameters such as number of filters and size of filters. The configuration settings of all the 12 CNN models are described in Table 3.9.

Table 3.9: Parameters settings of different CNN models

Model Name	Convolution Layers	Hidden Layers	Number of Filters	Filter Size
CNN1	2	6	10	3.4
CNN2	2	7	10	3.5
CNN3	2	6	50	3.4
CNN4	2	7	50	3.5
CNN5	2	6	60	3.4
CNN6	2	7	60	3.5
CNN7	3	6	100	3.4.5
CNN8	3	7	100	7.4.3
CNN9	3	6	128	3.4.5
CNN10	3	7	128	7.4.3
CNN11	3	6	256	3.4.5
CNN12	3	7	256	7.4.3

The results given by CNN models after experimentation with different parameter settings of CNN are discussed in the next section.

3.4.2 Results and Discussions

The trained CNN model has been run on the PC hardware specifications given in Table 3.10.

Table 3.10: Hardware specifications

RAM	16 GB
Processor	Intel(R) Core(TM) i7-7600U CPU@ 2.80GHz 2.90GHz
System Type	64-bit Windows OS, x64-based processor
GPU	Nvidia GeForce

The validation accuracy and loss score of all CNN models are listed in Table 3.11 along with its training time (in seconds). After performing several experiments with different parameters, it has been observed that CNN model with 2 convolution layers and filter size

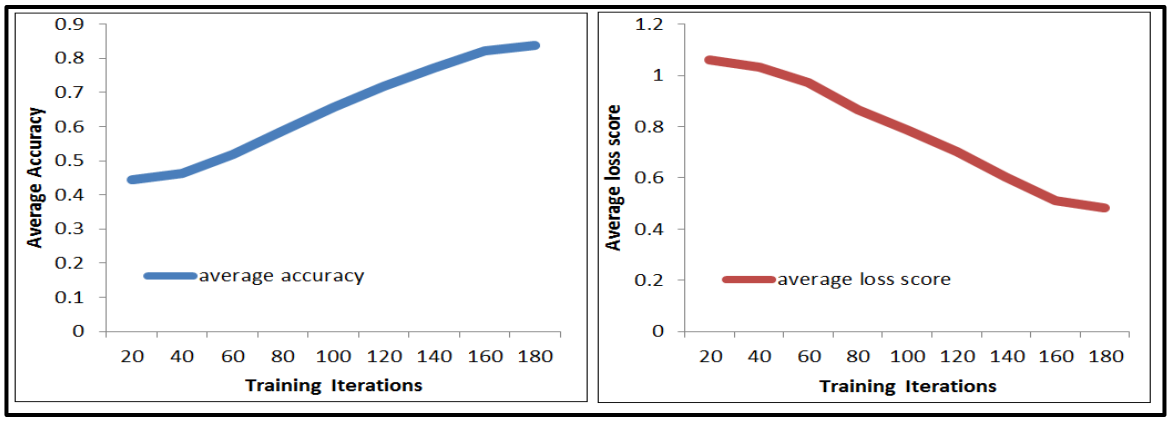
3, 4 performs better and achieved an accuracy of 95.4%. In case of CNN model with 3 convolutional layers, maximum achieved accuracy is 93.44%. It has been analyzed that by increasing the number of convolutional layers and filters, increases the training time of the model.

Table 3.11: Accuracy and loss of CNN models

Model Name	**Model Deatils**				**Validation Accuracy**	**Validation Loss**	**Training Time (seconds)**
	CL	HL	NF	FS			
CNN1	2	6	10	3.4	0.926	0.221	16.73
CNN2	2	7	10	3.5	0.953	0.142	16.87
CNN3	2	6	50	3.4	0.954	0.155	24.02
CNN4	2	7	50	3.5	0.884	0.298	25.07
CNN5	2	6	60	3.4	0.938	0.165	25.53
CNN6	2	7	60	3.5	0.941	0.183	26.12
CNN7	3	6	100	3.4.5	0.737	0.673	46.4
CNN8	3	7	100	7.4.3	0.859	0.331	53.43
CNN9	3	6	128	3.4.5	0.92	0.235	61.22
CNN10	3	7	128	7.4.3	0.934	0.165	67.11
CNN11	3	6	256	3.4.5	0.927	0.235	166.84
CNN12	3	7	256	7.4.3	0.934	0.205	169.83

*CL-Convolutional Layers, HL-Hidden Layers, NF- Number of Filters, FS-Filter Size

Figure 3.12 shows the average validation accuracy and loss score of all CNN models. The X-axis specifies the number of training iterations and Y-axis specifies the percentage of accuracy and loss score in Figure 3.12(a) and 3.12(b), respectively. The average learning curve in the Figure 3.12(a) shows that there is gradual increase in the accuracy percentage with the increase in training, it means that models are learning from data. The loss score is the total number of errors that the model predicted. Figure 3.12(b) shows that at the start, there were a lot of errors and as the number of training steps increased, the errors decreased. As most of the models had a good non-linear learning curve, dropping rapidly in the beginning and mostly have reached below 0.6.



(a) (b)

Figure 3.12 Average learning curve of (a) accuracy, (b) loss score for all CNN models

The other performance parameters such as precision, recall, F-measure, Kappa score, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for each of the CNN model are listed in Table 3.12.

Table 3.12: Other performance measures

Model Name	**Model Deatils**				**Precision**	**Recall**	**F-measure**	**Kappa**	**MAE**	**RMSE**
	CL	**HL**	**NF**	**FS**						
CNN1	2	6	10	3.4	0.931	0.926	0.926	0.889	0.08	0.304
CNN2	2	7	10	3.5	0.954	0.953	0.953	0.93	0.05	0.239
CNN3	2	6	50	3.4	0.954	0.954	0.954	0.93	0.06	0.297
CNN4	2	7	50	3.5	0.885	0.884	0.883	0.826	0.117	0.347
CNN5	2	6	60	3.4	0.939	0.938	0.938	0.907	0.076	0.322
CNN6	2	7	60	3.5	0.942	0.941	0.941	0.911	0.065	0.277
CNN7	3	6	100	3.4.5	0.753	0.736	0.716	0.604	0.365	0.754
CNN8	3	7	100	7.4.3	0.863	0.859	0.855	0.788	0.143	0.383
CNN9	3	6	128	3.4.5	0.921	0.92	0.92	0.881	0.11	0.413
CNN10	3	7	128	7.4.3	0.934	0.934	0.934	0.901	0.073	0.298
CNN11	3	6	256	3.4.5	0.93	0.927	0.927	0.89	0.076	0.286
CNN12	3	7	256	7.4.3	0.935	0.934	0.934	0.901	0.068	0.265

Figure 3.13 represents the comparison of precision, recall and F-measure of all 12 CNN models. The X-axis of bar chart specifies the CNN models with different parameter settings and Y-axis specifies the value of precision, recall and F-measure for each of the CNN model. Figure 3.14 compares the error rates in terms of MAE and RMSE given by CNN models.

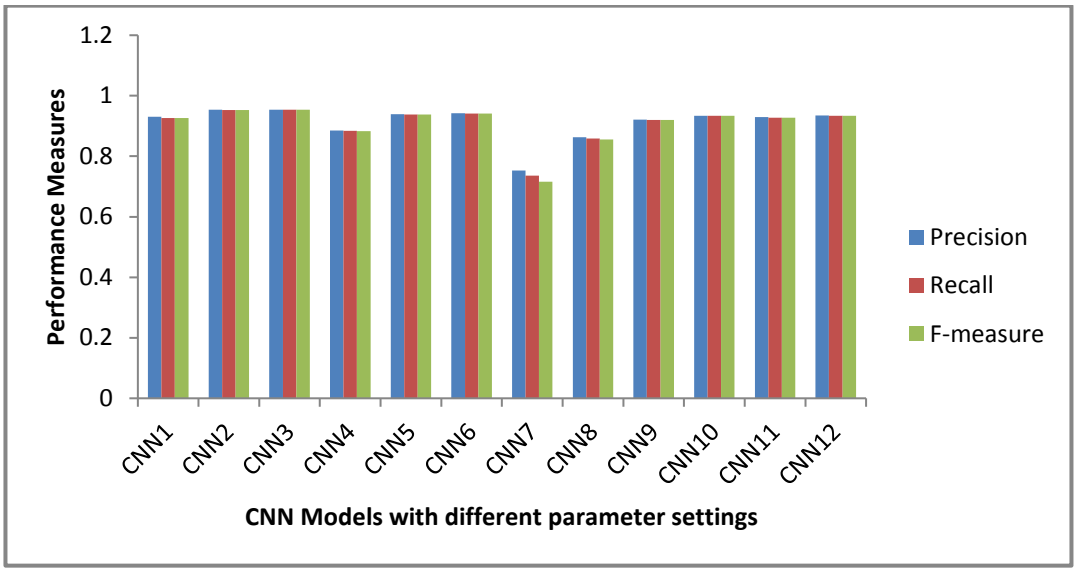


Figure 3.13 Comparison of precision, recall and F-measure for all CNN models

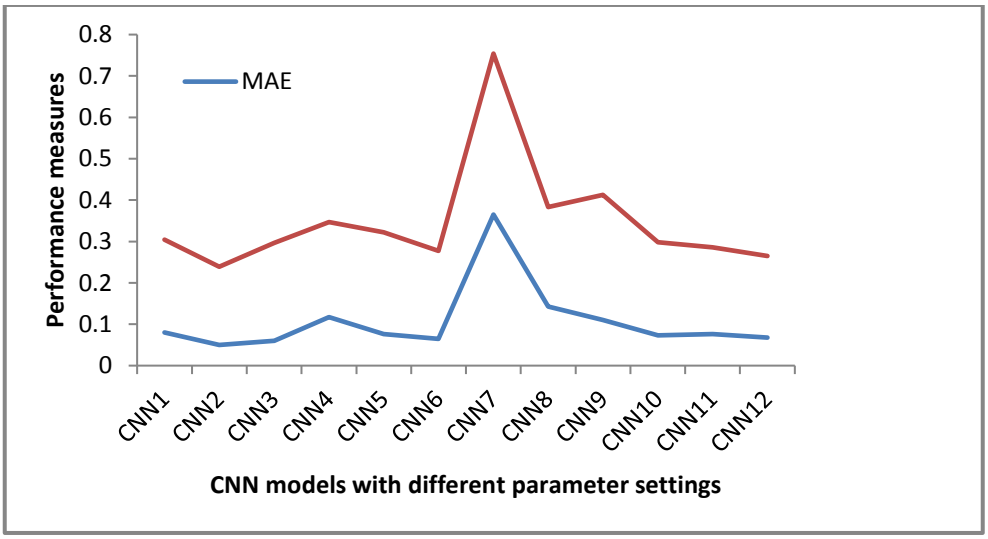


Figure 3.14 Comparison of error rates of all CNN models

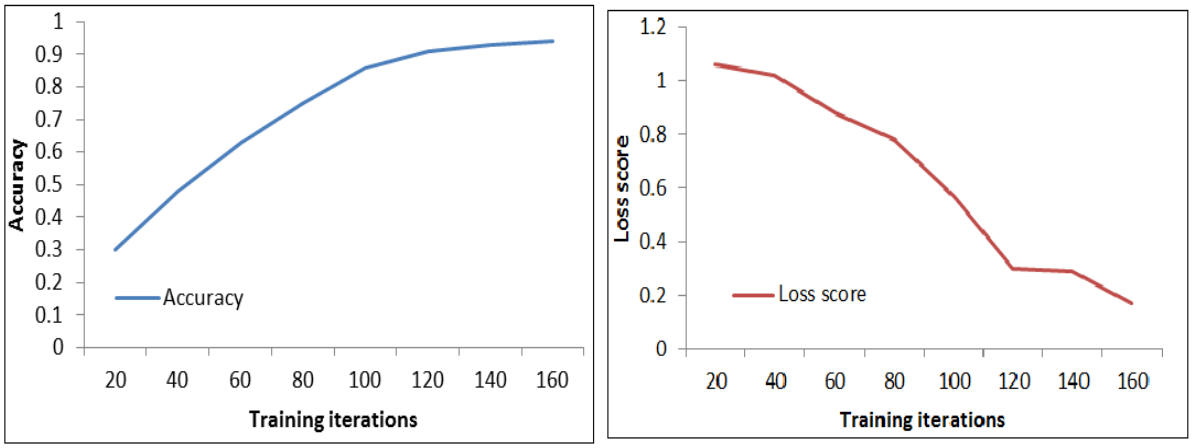
From Figure 3.13 and Figure 3.14, it has been observed that CNN3 is the best model having 2 convolutional layers, 6 hidden layers, 50 number of filter and two filters with

size is 3 and 4. CNN7 model performs worst out of all models having 3 convolutional layers, 6 hidden layers, 100 number of filters and three filters with size 3, 4 and 5 as its accuracy is less than 80%. The confusion matrix of the best model CNN3 is given in Table 3.13.

Table 3.13: Confusion matrix

		Predicted		
		A	**B**	**C**
Actual	A = Positive	948	20	28
	B = Negative	25	1505	24
	C = Neutral	7	16	1104

Figure 3.15(a) and 3.15(b) represent the learning curve of accuracy and loss score for the best performing model CNN3 having 2 convolutional layers, 6 hidden layers, 50 number of filter and two filters with size is 3 and 4.

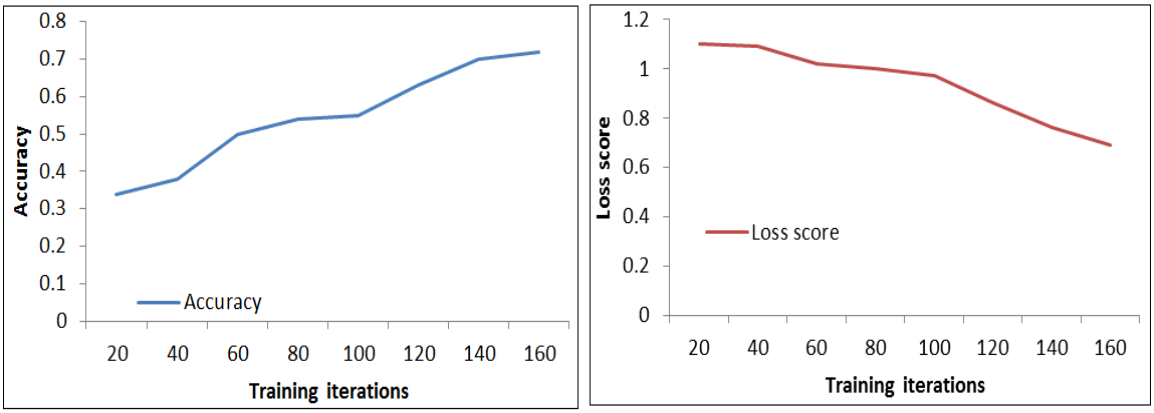


(a) (b)

Figure 3.15 Learning curve of (a) accuracy, (b) loss score for model CNN3

From the Figure 3.15(a), it has been observed that the accuracy of the model increases with the number of training iterations. But after 130 iterations, the accuracy gets stabled. Figure 3.15(b) represents that the learning curve of loss score for model CNN3 drops rapidly and reaches below 0.2 which means that this models has least number of errors.

Similarly, Figure 3.16(a) and 3.16(b) represent the learning curve of accuracy and loss score for the worst performing model CNN7 having 3 convolutional layers, 6 hidden layers, 100 number of filters and three filters with size 3, 4 and 5.



(a) (b)

Figure 3.16 Learning curve of (a) accuracy, (b) loss score for model CNN7

3.5 Comparison with Traditional ML Algorithms

In this section, results given by CNN based system are compared with traditional ML algorithms to analyze the improvement in accuracy. The traditional ML algorithms such as NB, k-Nearest Neighbor (k-NN), Maximum Entropy (ME) and Support Vector Machines (SVM) have been applied on the same corpus. The system learns on the basis of bag-of-words unigram feature model using different classifiers such as NB, k-NN, ME and SVM. The system uses 50% of the corpus for training and 50% of the corpus for testing purpose.

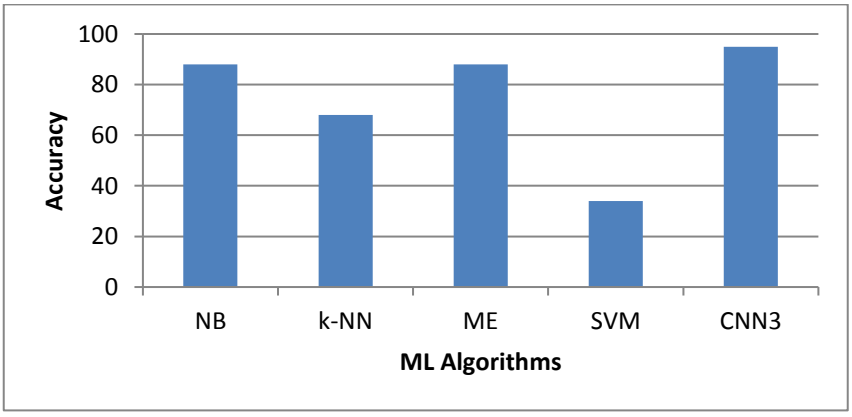


Figure 3.17 Comparative analysis of accuracy with traditional ML algorithms with CNN

The accuracy achieved using ML approaches on the tested corpus is compared with our CNN model and is represented in Figure 3.17. After performing experiments with different parameters variations of CNN models, it has been analyzed that 9 out of 12 CNN models are able to achieve an accuracy of above than 90%. This accuracy is comparatively better than traditional ML algorithms. Out of traditional ML algorithms, maximum accuracy (*i.e.*, 88%) is achieved by NB.

Also, statistical t-test has been performed which measures the significance of proposed approach. For this, the proposed approach with different parameter settings have been compared with existing models. The results signify the improvements of proposed over existing methods.

3.6 Comparison with Existing Works on Hindi Language

The results given by the proposed system using CNN are compared with existing works on SA for Hindi language using traditional ML algorithms. Table 3.14 presents the comparison of existing works on SA for Hindi language with proposed system on the basis of algorithm used, corpus type, size and accuracy.

Table 3.14: Comparison of proposed system with existing works on SA for Hindi language

Author(Year)	**ML algorithm**	**Corpus Type**	**Corpus Size**	**Accuracy**
Joshi et al., (2010)	SVM	Movie Reviews	250 reviews	78.14%
Balamurali et al., (2012)	SVM	Travel Reviews	200 reviews	72%
Bansal et al., (2013)	Deep Belief Network	Movie Reviews	300 reviews	64%
Jha et al., (2015)	NB	Movie reviews	200 reviews	87.1%
Se et al., (2015)	NB	Tweets	1,673 tweets	55.67%
Sheshadri et al., (2016)	RNN	Tweets	1,673 tweets	72.01%
Phani et al., (2016)	LR	Tweets	1,673 tweets	56.96%
Proposed Work	**CNN**	**Movie Reviews**	**7,354 reviews**	**95%**

Figure 3.18 depicts the comparison of accuracy given by the existing works on SA for Hindi language with the proposed Hindi SA system using CNN.

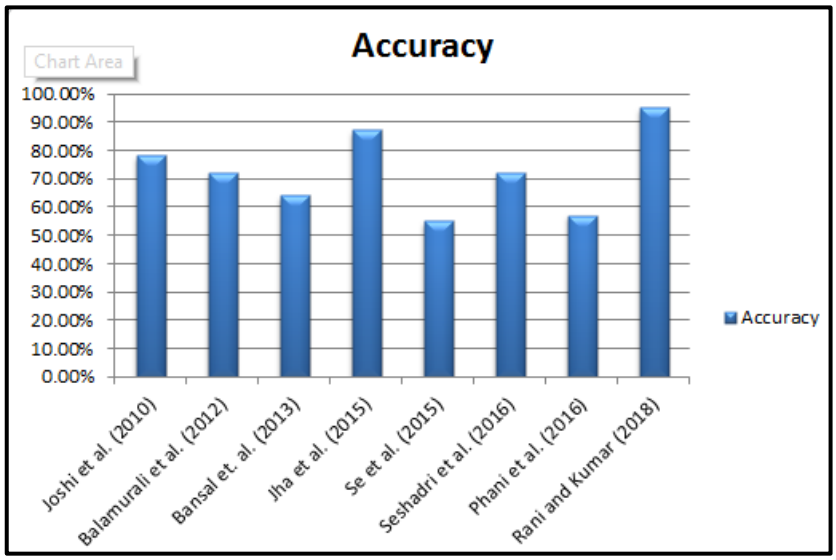


Figure 3.18 Comparison of accuracy of CNN based SA system with existing works

From Figure 3.18, it has been observed that CNN model is able to achieve better results on the newly constructed corpus in comparison to the existing works on SA for Hindi language.

3.7 Error Analysis

In this section, the errorneous cases have been discussed where the proposed CNN model fails to predict accurately. The proposed system classifies wrongly if explicity any sentiment bearing word is not present in the sentence. For example, consider the sentence given in (3.4) and its equivalent English translation is given in (3.5).

Hindi Sentence: इंटरवल के बाद फिल्म को थोड़ा जल्दबाजी में निपटाया गया है। (3.4)

Transliteration: *intaraval ke baad philm ko thoda jaldabaajee mein nipataaya gaya hai.*

Equivalent English Translation: *'After the interval, the film is dealt in a little hurry.'* (3.5)

In sentence (3.4), any sentiment bearing word is not present therefore, the system classifies it as neutral although the sentence consists of negative sentiment. The system

sometimes also miss-classifies some of the sentences on the basis of occurrence of sentiment bearing words present in the sentence.

For example, consider the sentence given in (3.6) and its corresponding English translation is given in (3.7).

Hindi Sentence: फ़िल्म का संगीत अच्छा है पर तेज़ है। (3.6)

Transliteration: *philm ka sangeet achchha hai par tez hai.*

Equivalent English Translation: *'The film's music is good but fast.'* (3.7)

The sentence (3.6) consists of both positive अच्छा *achchha 'good'* and negative तेज़ *tez 'fast'* sentiment words for the movies domain. However, the system predicts it as positive on the basis of frequency of positive word अच्छा *achchha 'good'* that is more in comparsion to negative sentiment word तेज़ *tez 'fast'*.

Chapter Summary

This chapter presents the implementation of sentiment analysis system for Hindi language using different ML algorithms. To train the ML algorithms, corpus of reviews and tweets has been collected from online websites and Twitter, respectively. The corpus has been annotated by three Hindi native speakers and has been validated using the statistic kappa measure. The experimental results given by different ML algorithms have been measured using performance measures precision, recall and F-measure. To further improve the accuracy of the system, deep learning based CNN has been applied on the corpus of Hindi reviews. The experimental results suggest that properly trained CNNs can outperform the traditional ML algorithms for sentiment classification. In CNN model, the sentences of reviews are labeled into three classes such as positive, negative and neutral. All the experiments are performed using different parameter settings for all CNN models and it has been observed that CNN model having 2 convolutional layers with filter size 3 and 4 performs the best all over models and is able to achieve an accuracy of 95%.

Chapter 4

Implementation of Aspect-based Sentiment Analysis System

This chapter presents the implementation of sentiment analysis system at aspect level. The lexical resources Hindi SentiWordNet (HSWN) and Hindi Dependency Parser (HDP) have been discussed in detail in this chapter which help in performing sentiment analysis at aspect level. This chapter also includes the example sentences of aspect-based sentiment analysis system along with its dependency graphs and sentiment polarity assigned to its corresponding aspects identified in a sentence. The working principle behind the implementation of aspect-based sentiment analysis system is explained with the help of procedures. The next section presents the introduction and need of aspect-based sentiment analysis system.

4.1 Introduction

Earlier the research work on SA mainly focuses on classifying the overall polarity of a review or a sentence into positive, negative or neutral. However, from the past few years, researchers are working on aspect-based (also known as feature-specific) sentiment analysis because sometimes reviews consist of different sentiments about different aspects/features and overall polarity does not help to identify the exact sentiments of people. Both the terms aspect-based and feature-specific have been used interchangebely throughout the thesis.

The aspect-based sentiment analysis helps to capture nuances about the objects of interest. For example, “battery life”, “screen” and “price” represent different aspects/features of a phone. The aspect-based sentiment analysis is very favorable because it assigns a separate sentiment expressed towards different aspects/features of an entity in a review and also helps in evaluating the overall sentiment expressed in a review or a sentence.

Consider a movie review “The story of the movie is awesome, but the songs are not good”. This review conveys a mixed opinion. Here, the sentiment about “story” is positive, whereas the sentiment about “songs” is negative. So, it is important to extract

only the relevant sentiments expressed about a particular feature from a sentence and categorize them, instead of extracting the overall sentiment.

In case of aspect-based/feature-specific SA, it has been observed that nouns, verbs, and adverbs also play an important role along with adjectives to find the sentiment about certain features in Hindi language. Some example sentences in which nouns, verbs, adverbs and adjectives act as sentiment-bearing words are given in Table 4.1.

Table 4.1: Sentences representing nouns, verbs, adverbs and adjectives as sentiment nodes

Sr. No.	**Input Sentence**	**Description**
1.	राम सीता को प्यार करता है। *raam seeta ko pyaar karata hai.* *'Ram loves Seeta'.*	The word प्यार *pyaar 'love'* acts as a noun in the sentence and it represents the positive sentiment about राम *raam 'Ram'.*
2.	वो रो रहा था । *vo ro raha tha.* *'He was crying'.*	The word रो *ro 'cry'* acts as a verb in the sentence and it represents the negative sentiment about वो *vo 'He'.*
3.	लता मंगेषकर अच्छा गाती है। *lata mangeshakar achchha gaatee hai.* *'Lata Mangeshkar sings well'.*	The word अच्छा *achchha 'well'* acts as an adverb in the sentence and it represents the positive sentiment about singing of लता मंगेषकर *lata mangeshakar 'Lata Mangeshkar'.*
4.	सीता की आँखें बहुत सुंदर है। *seeta kee aankhen bahut sundar hai.* *'Seeta has beautiful eyes'.*	The word सुंदर *sundar 'beautiful'* acts as an adjective in the sentence and it represents the positive sentiment about आँखें *aankhen 'eyes'* in the sentence.

Sr. No.	Input Sentence	Description
5.	इस फ़िल्म की कहानी कमज़ोर है, गीत मधुर हैं पर निर्देशन बहुत बुरा है। *is philm kee kahaanee kamajor hai, geet madhur hain par nirdeshan bahut bura hai.* *'The story of this film is weak, songs are melodious but the direction is very poor'.*	The words कमज़ोर *kamajor 'weak',* मधुर *madhur 'melodious'* and बुरा *bura 'poor'* in the sentence are acting as adjectives and representing the opinion about aspects कहानी *kahaanee 'story'*, गीत *geet 'song'* and निर्देशन *nirdeshan 'direction'*, respectively.

4.2 Related Work on Aspect-based Sentiment Analysis

From past many years, researchers from all over the world are working in the area of feature-specific SA over different languages and the brief literature of their work is presented as follows.

Wu et al., (2009) and Zhang et al., (2009) used dependency parsing to perform opinion mining of product reviews for the English language. They extracted product features, sentiment expressions and relations between them to perform a feature-specific SA. Mosha and Tianfang (2010) also used dependency parsing to extract opinion-element relation and semantic information to perform SA for Chinese language. They categorized relations on the basis of location of a topic and sentiment present in the structure of sentence. Thet et al., (2010) performed feature-based SA of movie reviews using lexicon based approach. They used SentiWordNet for determining the score of the sentiment words. Bora (2011) performed feature-based sentiment analysis of Twitter. Author used a combination of minimum word frequency threshold and Categorical Proportional Difference (CPD) as feature selection method and trained Naïve Bayes (NB) classifier using a training set of 1.5 million tweets on manually labelled data set.

Mukherjee and Bhattacharyya (2012) performed feature-specific SA of product reviews using dependency parser for English language. They extracted the potential features from

a review and partitioned the sentiment expressions into clusters where each cluster described each feature. They merged closely connected sentiment expressions on the basis of threshold parameter. Di Caro and Grella (2013) proposed a context-based model in which SA was performed by syntactic-based propagation rules using dependency parsing. They used five dependencies, *i.e.*, modifiers, tuners, inverters, prepositions and verbs for propagation process. Singh et al., (2013) performed aspect-based SA of movie reviews for English language using SentiWordNet and different linguistic feature selections including combinations of adjectives, adverbs and verbs. They compared their results with results obtained by Alchemy API and showed that their approach is more accurate. Robaldo and Di Caro (2013) proposed an XML-based methodology for annotation of affective sentiments in domain-independent textual expressions. They evaluated their approach by performing fine-grained analysis of the disagreement between different annotators.

Erdmann et al., (2014) performed feature-specific SA of tweets for English and Japanese language. They extracted features from online review articles of products and used these features to perform SA for tweets. Their method improved the feature extraction process in comparison to features extracted directly from tweets. Poria et al., (2014) used dependency parse tree-based rules to identify the associated sentiment from text by extracting concepts and aspects. Jiménez-Zafra et al., (2015) also performed feature-based SA using unsupervised lexicon based approach by combining different linguistic resources. Araque et al., (2015) performed SA of tweets at global and aspect level for Spanish language. They used graph-based algorithm to extract the features and polarity lexicons to determine the sentiment words. Vilares et al., (2015) proposed syntactic-based approach of SA for Spanish reviews. They handled some linguistic features such as intensifiers, negations *etc*. and concluded that their proposed approach is better over ML and lexicon-based approaches. Rana and Cheah (2016) proposed rule based hybrid approach for aspect extraction and categorization from customer reviews. They used sequential patterns and Normalized Google Distance (NGD) to extract explicit as well as implicit aspects. Salas-Zárate et al., (2017) performed feature-based SA on financial domain. They proposed an ontology based approach for feature and news polarity classification using the linguistic expressions of the feature. Dehkharghani et al., (2018)

performed SA on Turkish movie reviews using dependency parsing approach at different granularity levels, such as aspect, sentence and document. They used polarity lexicon SentiTurkNet to perform SA.

There are relatively few works on sentiment analysis of Hindi language text as comparison to other languages. Mittal et al., (2013) and Arora (2013) performed SA of Hindi reviews using unsupervised lexicon based approach. They also handled negations and discourse relations to improve the accuracy of the proposed sentiment analysis system for Hindi language. Sharma et al., (2014) performed sentiment analysis of movie reviews using dictionary based approach. They also handled negations and achieved an accuracy of 65%. Pandey and Govilkar (2015) performed SA of Hindi movie reviews using HSWN and also handled negations and discourse relations. Akhtar et al., (2016b) performed aspect based SA for Hindi language on a dataset of product reviews using Conditional Random Field (CRF) and Support Vector Machines (SVM). They achieved an accuracy of 54.05% for sentiment classification. Garg and Buttar (2017) performed aspect-based sentiment analysis of Hindi text using dictionary-based approach and classified the text into positive, negative and neutral class. Rai et al., (2017) used machine learning and lexicon-based approach to perform sentiment analysis of political reviews and analyzed that lexicon-based approach with negation handling outperforms the machine learning approach. Hussaini et al., (2018) performed sentiment analysis of book reviews using lexicon-based approach and improved the accuracy of the system using word sense disambiguation and by handling morphological variations.

From the literature review, it has been analyzed that the most of the research work has been performed for English language and there exists limited research work for Hindi on aspect-based/feature-specific SA. For Hindi, mainly the research work on SA has been reported at sentence level using different techniques such as lexicon based and ML. In case of lexicon based techniques, the authors used simple baseline approaches of sentiment word count and sentiment prior score based approach to perform SA, which do not provide productive results. Till now, researchers have not proposed any effective approach to perform a aspect-based/feature-specific SA for domain-independent datasets. Furthermore, researchers have evaluated their approaches on their own data sets for

different languages, which make it difficult to compare various approaches with each other.

In this chapter, an aspect-based sentiment analysis system is presented. The proposed system benefits from the existing techniques and lexical resources available in the literature for Hindi language such as Hindi SentiWordNet (HSWN) and Hindi Dependency Parser (HDP) to deal with sentiment analysis issues. It also proposes new solutions for SA linguistic issues like handling of transliteration, negations and intensifiers.

4.3 Lexical Resources

The brief description about the lexical resources used to perform the aspect-based SA for Hindi language is given as follows.

4.3.1 Hindi SentiWordNet (HSWN)

Hindi SentiWordNet[1] is developed by IIT Bombay. Joshi et al., (2010) created this lexical resource using two lexical resources Hindi WordNet and English SentiWordNet. The format of HSWN is given in Table 4.2.

Table 4.2: Format of HSWN

Field 1	**Field 2**	**Field 3**	**Field 4**	**Field 5**
POS Tag	Synset ID (Hindi WN)	Positive score	Negative score	Related terms {separated by comma}

For example, consider the entry of HSWN given in (4.1).

a 1831 0.75 0.0 अच्छा *achchha 'good'* (4.1)

Here, 'a' represents the *POS* tag, *i.e.*, adjective, '1831' is Synset ID, '0.75' is positive score, '0.0' is negative score and अच्छा *achchha 'good'* represents sentiment word. HSWN consists of 2995 Synset IDs. In order to extend the coverage of existing HSWN,

[1] http://www.cfilt.iitb.ac.in/resources/senti/HSWN_downloaderInfo.php

some of the sentiment words are added manually into HSWN. The manually added sentiment words are annotated by three Hindi native speakers. Each word is assigned a positive and negative score (between 0-5) based on the polarity. The final sentiment score of the manually added sentiment words is decided by taking the average of sentiment scores assigned by the three Hindi native speakers and then the sentiment score is normalized into 0-1 to maintain the consistency with the existing HSWN lexicon.

4.3.2 Hindi Dependency Parser (HDP)

Hindi Dependency Parser[2] is developed by IIIT Hyderabad. The output format of HDP consists of six columns as given in Table 4.3, where *Parent id* represents the ID of dependent word and *Dependency label* specifies the relation between the current word and the dependent word. Some of the important dependency relations considered by HDP are *k1* (agent), *jk1* (causee), *k1s* (noun complement of agent), *k2* (object), *k3* (instrument), *k5* (source), *jjmod* (adjective modifier), *nmod* (noun modifier), *vmod* (verb modifier), *r6* (possessive relation between two nouns) and *k7* (location other than time and place) etc.

Table 4.3: Format of HDP

Field 1	**Field 2**	**Field 3**	**Field 4**	**Field 5**	**Field 6**
Word id	Word	lemma	POS tag	Parent id	Dependency label

For example, consider the Hindi sentence given in (4.2) and its equivalent English translation is given in (4.3).

Hindi Sentence: इस फ़िल्म की कहानी कमज़ोर है, गीत मधुर हैं पर निर्देशन बहुत बुरा है। (4.2)

Transliteration: *Is philm kee kahaanee kamajor hai, geet madhur hain par nirdeshan bahut bura hai.*

Equivalent English Translation: *'The story of this film is weak, songs are melodious but the direction is very poor.'* (4.3)

The HDP output of the sentence given in (4.2) is shown in Table 4.4. In this output, the first column represents the *word id*, i.e., a number is assigned to each word or token of the

[2] https://bitbucket.org/sivareddyg/hindi-dependency-parser/downloads

sentence. The sentence (4.2) consists of total 16 words/tokens. The second column is *word* which represents the actual word contained by the sentence. The third column *lemma* represents the root word of each word consisted by sentence. For example, for *word id* 3, i.e., की *kee 'of'*, the root word is का *ka 'of'*. The fourth column *POS Tag* represents the part-of-speech of the word, for example, 'NN' represents noun for word फ़िल्म *philm 'film'*. The fifth column *parent id* represnts the dependency of a word over another word of the sentence. In the given output, *parent id* of word इस *Is 'this'* is 2 means it depends on *word id* 2 which is फ़िल्म *philm 'film'*. The last column *dependency label* represents the dependency relation between the dependent words. For example, the relation between words बहुत *bahut 'bahut'* and बुरा *bura 'bura'* is '*jjmod__intf'* which means adjective modifier or intensifier.

Table 4.4: Output of the HDP for the sentence (4.2)

Word id	Word	Lemma	POS tag	Parent id	Dependency label
1	इस	इस	DEM	2	nmod__adj
2	फ़िल्म	फ़िल्म	NN	4	r6
3	की	का	PSP:का	2	lwg__psp
4	कहानी	कहानी	NN	6	k1
5	कमज़ोर	कमज़ोर	JJ	6	k1s
6	है	है	VM	0	main
7	,	,	,	6	rsym
8	गीत	गीत	NN	10	k1
9	मधुर	मधुर	JJ	10	k1s
10	हैं	हैं	VM	15	k7
11	पर	पर	PSP: पर	10	lwg__psp
12	निर्देशन	निर्देशन	NN	15	k1
13	बहुत	बहुत	INTF	14	jjmod__intf
14	बुरा	बुरा	JJ	15	k1s
15	है	है	VM	6	vmod
16	.	.	.	15	rsym

4.3.3 Other Resources Used

As Hindi is morphologically rich and free order language, HDP effectively handles all these issues. Sometimes, people use romanized and Hinglish text (*i.e.*, English words blending with Hindi) while posting their reviews on the web and direct processing of this type of text is not possible. Therefore, to handle the issue of romanized text and abbreviations, the Google API and mapping dictionary have also been used. Consider the sentence given in (4.4) and its equivalent English translation is given in (4.5).

Hindi Sentence: Coolie फ़िल्म अच्छी और रोमांचक है। (4.4)

Transliteration: *chooliai philm achchhee aur romaanchak hai.*
Equivalent English Translation: *'The film Coolie is good and thriller.'* (4.5)

The sentence (4.4) consists of romanized (*e.g.*, Coolie) text. In this case, Therefore, Google API helps in handling the issue of romanized text by performing the transliteration from Coolie to कुली *chooliai 'Coolie'*

In the next section, the proposed aspect-based sentiment analysis system for Hindi is presented.

4.4 Architecture of Proposed System

The proposed system takes a document (or a review) as input which is segmented into sentences. The system pre-processes the input sentence and parses them using HDP that provides a dependency structure of a sentence and morphological analysis of each word of the sentence. The system extracts the relevant features and sentiment words using Part-of-Speech (POS) information and polarity lexicon HSWN. Then, the system generates a dependency graph of the sentence which helps in performing aspect-based SA. Based on the idea that closely connected words come together to express a sentiment about a certain aspect, the sentiment word is assigned to the particular aspect having the least distance from the aspect word. The proposed system consists of following eight phases as illustrated in Figure 4.1.

- data extraction phase;
- pre-processing phase;
- extraction of sentiment nodes;
- aspect extraction phase;
- creation of aspect vector;
- dependency graph generation phase;
- negation and intensifiers handling phase;
- polarity assignment phase.

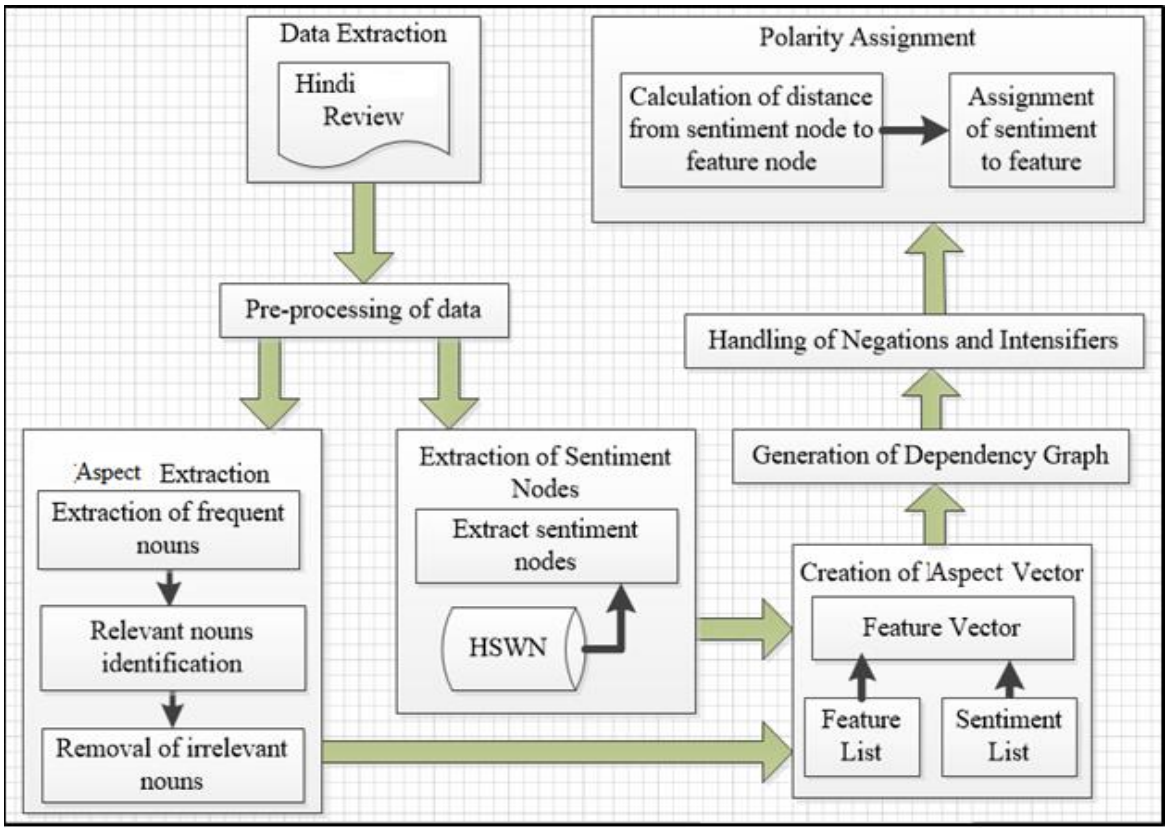


Figure 4.1 Architecture of proposed aspect-based sentiment analysis system

The detailed description about each phase of the system is given as follows. The Algorithm 4.1 ABSA(*s*) describes the overall process of aspect-based sentiment analysis system.

ALGORITHM 4.1: ABSA(*s*)
Input: Sentence(*s*) **Output:** $a\_id \rightarrow s\_id \rightarrow p\_SCR \rightarrow n\_SCR \rightarrow sent \rightarrow overall\_sent$

▷ Call to procedure pre-process(*s*)

1. Pre-processing of sentence

▷ Call to procedure SAE(*s'*)

2. Extraction of sentiment and aspect nodes

▷ Call to procedure ALSL(*s', HDP_{output}*)

3. Creation of aspect list and sentiment list

▷ Call to procedure DGOL(*AL, SL, HDP'_{output}*)

4. Generation of dependency graph and output list

▷ Call to procedure PAOS(*OL,HDP'_{output}*)

5. Final assignment and calculation of overall score

Description: The Algorithm 4.1 describes the overall process of aspect-based sentiment analysis system using dependency parsing. This algorithm calls five different procedures such as pre-process(*s*) for pre-processing of input sentence, SAE(*s'*) to extract the sentiment and aspect nodes by using HSWN and output of HDP for the input sentence, ALSL(*s', HDP_{output}*) to create an aspect list and sentiment list from the input sentence, DGOL(*AL*, *SL*, *HDP'_{output}*) to generate the dependency graph from the output of HDP and PAOS(*OL,HDP'_{output}*) to assign polarity and overall score of the sentence for performing sentiment analysis at aspect level.

4.4.1 Data Extraction Phase

For the proposed work, the corpus of reviews in Hindi language that has been extrcated in previous chapter is considered to perform aspect-based SA. These movie reviews are extracted from फ़िल्म-समीक्षा *philm-sameeksha* '*film-review'* section of *aajtak (http://aajtak.intoday.in/film-review.html)* and *jagran (http://www.jagran.com/entertainment/reviews-news-hindi.html)* online newspapers. The products reviews have been collected from http://hindi.mymobileindia.com/. To extract reviews in Hindi language, Hindi Web Text Crawler is developed using a java-based

"*boilerpipe*" library. This library extracts the required text from a web page using *ArticleExtractor* parameter and saves it in '*UTF-8'* format with '*.txt'* extension. The corpus consists of 2247 movie reviews in Hindi language.

4.4.2 Pre-Processing Phase

A pre-processing phase that is used in the previous chapter has been followed for the same corpus to prepare the data for further processing. After a transliteration and mapping, each sentence is processed by HDP. The Procedure 4.1 to illustrate the pre-processing of input dataset has been given as follows.

PROCEDURE 4.1 Pre-process(*s*)

Input: Sentence
Output: s' and
$HDP_{output} = (node\_id, node, lemma, POS, dep\_node\_id, dep\_rel)$;
// Transliteration of sentences
1. **for** $\forall(s \in D)$ **do** // D is dataset.
   perform transliteration using $Google\ API$ to obtain s'; // s' is a new sentence.
   $s' \leftarrow s$;
   end for

// Mapping of Hinglish words
2. **for** $\forall(Hinglish\_words \in s')$ **do**
   replace $Hinglish\_words$ by $Hindi\_words$ using $mapping$ dictionary;
   end for

// Processing using HDP
3. **for** $\forall s'$ **do**
   process s' using HDP to obtain
   $HDP_{output} = (node\_id, node, lemma, POS, dep\_node\_id, dep\_rel)$;
   end for

Description: The Procedure 4.1 performs the pre-processing of input corpus. It includes the transliteration of input sentences using Google API, mapping of Hinglish words to Hindi words using mapping dictionary and processing of sentences using Hindi Dependency Parser (HDP). The output of HDP is stored in the form of vector which

consists of six parameters, i.e., *node id*, *node*, *lemma*, *part-of-speech*, *dependent node id* and *dependency relation* between node and dependent node.

4.4.3 Extraction of Sentiment Words

In this phase, sentiment words are identified using HSWN. Each token of the sentence is matched with entries of the HSWN. If the token is found in the HSWN then information about sentiment-bearing token is stored in a list according to the format given in (4.6). The list represents three attributes of the node, *i.e.*, POS information of sentiment node, positive score and negative score.

{Node: [POS, p_SCR, n_SCR]} (4.6)

For example, sentiment words extracted from the sentence (4.2) are {कमज़ोर *kamajor 'weak'*: [JJ, 0.0, .50]}, {मधुर *madhur 'melodious'*: [JJ, 0.50, 0.0]} and {बुरा *bura 'poor'*: [JJ, 0.0, 0.50]}.

4.4.4 Aspect Extraction Phase

An efficient aspect representation process has been used in this work to extract the relevant aspects from the corpus. For aspect extraction process, POS tags of HDP output are used and the following steps are performed to extract the relevant aspects.

a) **Frequent Nouns Identification:** According to the comprehensive sentiment analysis and conclusion given by Nakagawa and Mori (2002), it has been observed that the frequent nouns in the text generally relate to the relevant aspects of the domain. The output of HDP helps in extracting all the frequent nouns. A noun is considered as frequent if its frequency is greater than a threshold which is determined by analyzing the corpus.

b) **Relevant Nouns Identification:** From the corpus, it has been analyzed that there are still some important aspects that may not be frequent in the corpus. These aspects are extracted by finding the nouns adjacent to adjectives. For example, consider the Hindi sentence given in (4.7) and its equivalent English translation is given in (4.8). Here, the word नज़म *nazam 'poem'* is a non-frequent noun in the

corpus, but it has been identified as relevant noun because it occurs adjacent to adjective अच्छी *achchhee 'good'.*

Hindi Sentence: फ़िल्म में अच्छी नज़में हैं। (4.7)

Transliteration: *philm mein achchhee nazamen hain.*

Equivalent English Translation: *'The film has good poems'.* (4.8)

c) **Unrelated Nouns Removal:** To further improve the extraction process, all the irrelevant nouns are removed which have *Pointwise Mutual Information-Information Retrieval (PMI-IR)* measure less than the threshold. *PMI-IR* measure is used to determine the association between two or more words. Turney (2002) used this measure for sentiment classification which calculates *PMI-IR* by querying a web search engine according to the formula given in (4.9).

$$PMI_{IR}\left(word_1, word_2\right) = \log_2 \frac{(Hits(word_1 \cap word_2))}{(Hits(word_1) * Hits(word_2))} \quad (4.9)$$

Here, $Hits(word_1)$ represents the number of pages containing $word_1$, $Hits(word_2)$ specifies the number of pages containing $word_2$ and $Hits(word_1 \cap word_2)$ describes the number of pages containing both the words. The Google search engine has been used to calculate the number of hits for every feature word in isolation, number of hits for a word demonstrating the domain, *e.g.,* '*movie*' domain and the number of hits containing both domain word and aspect word.

For example, consider the Hindi sentence given in (4.10) and its equivalent English translation is given in (4.11).

Hindi Sentence: "तारे ज़मीन पर" फ़िल्म में दर्शकों को स्कूल अच्छा लगा । (4.10)

Transliteration: *"taare zameen par" philm mein darshakon ko skool achchha laga.*

Equivalent English Translation: *'The audience liked the school in the movie "Taare Zameen Par"'.* (4.11)

Here, स्कूल *skool 'school'* is a noun and it is added as an aspect. *PMI-IR* measure of this noun is less than the threshold value in case of '*movie'* domain. So, it is considered as an unrelated noun and it is removed in this phase.

This phase is an important phase of aspect-based sentiment analysis system. Because, if the POS of the word is not identified correctly by the HDP in this phase, then it will affect the performance of the proposed system. There can be some words in Hindi that can behave as nouns as well as adjectives in different sentences. In that case, the proposed system will not be able to identify the sentiment of that sentence correctly. For example, the word प्रेम *prem 'love'* can be a noun as it may be the name of a person or can behaves as an adjective also in some sentences. Therefore, if a sentence consisting of the word प्रेम *prem 'love'* is processed by the HDP then it identifies it as noun. In this case, the proposed sytem will not classify this sentence as positive although it should be positive which will affect the accuracy of the system.

Instead of nouns only, proper nouns and pronouns have also been considered as aspects and these are added to the final aspect list. The Procedure 4.2 to illustrate the process of Sentiment and Aspect Extraction (SAE) is given as follows.

PROCEDURE 4.2: SAE (s')

Input: s' and HDP_{output}

Output: FNL

// Tokenization of sentences

1. **tokenize** s' where $s' = node_1, node_2, ..., node_n; n \in N$

// Extraction of sentiment nodes

2. **for** $\forall(senti\_POS \in POS)$ of HDP_{output} where
   $senti\_POS = \{NN, JJ, ADV, VB\}$ **do**
   search in $HSWN$ to obtain $HSWN_{output} = \{node : [POS, p\_SCR, n\_SCR]\}$
   where $node \in s'$; // p_SCR and n_SCR are positive and negative score of node respectively.
   end for

// Extraction of frequent nouns if it's count is greater than threshold

3. **for** $\forall(node \in s')$ **do**
   if $(POS(node) == NN)$ of HDP_{output} **then**
   if $count > threshold$ **then**
   add $node$ to $noun\_list$ NL;

PROCEDURE 4.2: SAE (s')

else

add *node* to *irrelevant_noun_list* $INL \mid node' \leftarrow node$;

// INL is an irrelevant noun list.

end if

end if

end for

// Identification of relevant nouns which are adjacent to adjectives

4. **for** $\forall(node' \in INL)$ **do**

if $\exists(POS(node) == JJ)$ of HDP_{output} **then**

if *node'* is adjacent to *node* **then**

add *node'* to NL;

end if

end if

end for

// Removal of irrelevant nouns if PMI_{IR} is less than threshold

5. **for** $\forall(node \in NL)$ **do**

$PMI_{IR}(node, domain) = log_2 \frac{(Hits(node \cap domain))}{(Hits(node) * Hits(domain))}$ // domain= movie

if $(PMI_{IR}(node) > threshold)$

add *node* to *final noun_list* FNL;

end if

end for

6. **Extract** proper nouns as well as pronouns and **add** them to FNL.

Description: The Procedure 4.2 named as Sentiment and Aspect Extraction (SAE) takes the output of pre-processed sentence(s') as input. According to this procedure, each sentence s' is tokenized into nodes, *i.e.*, $node_1$, $node_2$... $node_n$ etc. The sentiment nodes are extracted using HSWN. To extract the sentiment nodes, four POS tags, *i.e.*, Noun (NN), Adjective (JJ), Adverb (ADV) and Verb (VB) are considered. The POS tag, positive and negative score of corresponding node extracted from HSWN are stored in a list $HSWN_{output}$. Then, the frequent Nouns List (NL) is computed using a threshold from HDP_{output} and adds the irrelevant nouns to Irrelevant Nouns List (INL) if its count is greater than threshold. Then, nouns which are adjacent to adjectives are extracted from

INL by considering it as relevant and added to the NL. The last step of this procedure prepares a Final Noun List (FNL) by calculating the PMI_{IR} for each node of NL for '*movie*' domain. Also, proper nouns and pronouns have been added into FNL in this step.

4.4.5 Creation of Aspect Vector

In this phase, an aspect vector is created for each node of the sentence and finally a dictionary is prepared consisting of the same. The structure of the feature vector is given by (4.12).

{Node: [node_id, dep_node_id, dep_rel, POS, lem, a_flag, s_flag, p_SCR, n_SCR]}(4.12)

Here, *Node* represents a word in a sentence. The description about feature vector attributes of the node is presented in Table 4.5.

Table 4.5: Description of aspect vector attributes

Sr. No.	**Attribute**	**Description**	**Remarks**
1.	*node_id*	node ID	The position of the head node in a sentence.
2.	*dep_node_id*	dependent node ID	The position of the dependent node.
3.	*dep_rel*	dependency relation	The relation between head node and dependency node.
4.	*POS*	part-of-Speech	POS information about head node.
5.	*Lem*	lemma	Root word of the node.
6.	*a_flag*	aspect flag	In case of aspect node, set *a_flag* = 1 and in case of sentiment node, set *s_flag* = 1 else set both to 0.
7.	*s_flag*	sentiment flag	
8.	*p_SCR*	positive score	In case of aspect node, set both *p_SCR* and *n_SCR* to 0 else set according to the polarity of sentiment node.
9.	*n_SCR*	negative score	

The aspect vector dictionary consisting of aspect vector for each node of the sentence (4.2) is represented in Figure 4.2.

{‘इस’: [‘N1’, ‘N2’, ‘nmod__adj’, ‘DEM’, ‘इस’, 0, 0, 0.0, 0.0], ‘फ़िल्म’: [‘N2’,‘N4’, ‘r6’, ‘NN’, ‘फ़िल्म’, 1, 0, 0.0, 0.0], ‘की’:[‘N3’, ‘N2’, ‘lwg__psp’, ‘PSP’, ‘की’, 0, 0, 0.0, 0.0], ‘कहानी’:[‘N4’, ‘N6’, ‘k1’, ‘NN’, ‘कहानी’, 1, 0, 0.0, 0.0], ‘कमज़ोर’:[‘N5’, ‘N6’, ‘k1s’, ‘JJ’, ‘कमज़ोर’, 0, 1, 0.0, 0.50], ‘है’:[‘N6’, ‘N0’, ‘main’, ‘VM’, ‘है’, 0, 0, 0.0, 0.0], ‘,’:[‘N7’, ‘N6’, ‘rsym’, ‘,’, ‘,’, 0, 0, 0.0, 0.0], ‘गीत’:[‘N8’, ‘N10’, ‘k1’, ‘NN’, ‘गीत’, 1, 0, 0.0,0.0], ‘मधुर’:[‘N9’, ‘N10’, ‘k1s’, ‘JJ’, ‘मधुर’, 0, 1, 0.5, 0.0], ‘हैं’:[‘N10’, ‘N15’, ‘k7’, ‘VM’, ‘हैं’, 0, 0, 0.0, 0.0], ‘पर’:[‘N11’, ‘N10’, ‘lwg__psp’, ‘PSP’, ‘पर’, 0, 0, 0.0, 0.0], ‘निर्देशन’:[‘N12’, ‘N15’, ‘k1’, ‘NN’, ‘निर्देशन’, 1, 0, 0.0, 0.0], ‘बहुत’:[‘N13’, ‘N14’, ‘jjmod__intf’, ‘INTF’, ‘बहुत’, 0, 0, 0.0, 0.0], ‘बुरा’:[‘N14’, ‘N15’, ‘k1s’, ‘JJ’, ‘बुरा’, 0, 1, 0.0, 0.50], ‘है’:[‘N15’, ‘N6’, ‘vmod’, ‘VM’, ‘है’, 0, 0, 0.0, 0.0], ‘.’:[‘N16’, ‘N15’, ‘rsym’, ‘.’, ‘.’, 0, 0, 0.0, 0.0]}

Figure 4.2 Aspect vector dictionary of sentence (4.2)

For example, the aspect vector created for node फ़िल्म *philm* ‘*film*’ has *node id* ‘N2’, its *dependent node id* is ‘N4’, the *dependency relation* between ‘N2’ and ‘N4’ is ‘r6’, its *POS tag* is ‘NN’, i.e., noun, *lemma* is फ़िल्म *philm* ‘*film*’, the aspect flag *a_flag* has value 1 because the given node is acting as an aspect node for the given sentence, the sentiment flag *s_flag* has value equal to 0 as the given node is not sentiment bearing node, the value of *p_SCR* and *n_SCR* are also set to 0 as the given node is acting as aspect node and does not have any positive and negative score. In the aspect vector dictionary, the nodes having *a_flag* = 1 are added into aspect list and nodes having *s_flag* = 1 are added into sentiment list.

The Procedure 4.3 to illustrate the creation of Aspect List and Sentiment List (ALSL) from the aspect vector dictionary of the sentence has been given as follows.

PROCEDURE 4.3: ALSL (s', $\boldsymbol{HDP_{output}}$)

Input : s' and HDP_{output}

Output: AL and SL

// Updating HDP_{output} to HDP'_{output} by adding new parameters

1. **for** $\forall(node \in s')$ **do** // Amendment of each node of HDP_{output}

$HDP'_{output} = HDP_{output} \cup \{f\_flag, s\_flag, p\_SCR, n\_SCR\}$ Where

$$a\_flag(p) = \begin{Bmatrix} node \in FNL & 1 \\ node \notin FNL & 0 \end{Bmatrix}, \quad s\_flag(p) = \begin{Bmatrix} node \in HSWN_{output} & 1 \\ node \notin HSWN_{output} & 0 \end{Bmatrix}$$

and $m = \begin{Bmatrix} a\_flag = 1 & 0 \\ a\_flag = 0 & [0-1] \end{Bmatrix}$; a_flag is aspect flag, s_flag is sentiment flag,

$p : a\_flag, s\_flag \in \{0,1\}$ and $m : p\_SCR, n\_SCR \in [0-1]$

end for

// Creation of aspect list and sentiment list from aspect and sentiment flag bits of HDP'_{output}

2. **for** $\forall(node \in s')$ **do**

if $(a\_flag(node) == 1)$ of HDP'_{output} **then**

add $node\_id$ to $feature\_list\ FL$;

else if $(s\_flag(node) == 1)$ of HDP'_{output} **then**

add $node$ to $sentiment\_list\ SL$ & set $node\_id' \leftarrow node\_id$;

end if

end for

Description: The Procedure 4.3 helps in preparing Aspect List and Sentiment List (ALSL). This procedure adds four more parameters such as feature flag (a_flag), sentiment flag (s_flag), positive score (p_SCR) and negative score (n_SCR) to HDP_{output} and updates it as HDP'_{output}. If a node belongs to FNL, then a_flag is set to 1 otherwise 0 and adds the nodes to Aspect List (AL) having a_flag equal to 1. Similarly, if a node belongs to $HSWN_{output}$, then set s_flag to 1 else 0 and add corresponding nodes to Sentiment List (SL) having s_flag equal to 1. Both positive and negative score of the nodes (having a_flag = 0) are set according to its polarity otherwise both set to 0. For example, the aspect list and the sentiment list created for the sentence (4.2) are given in (4.13) and (4.14).

Aspect List (AL) = [N2: फ़िल्म *philm 'film', N4:* कहानी *kahaanee 'story', N8:* गीत *geet 'song', N12:* निर्देशन *nirdeshan 'direction']* (4.13)

Sentiment List (SL) = [N5: कमज़ोर *kamajor 'weak', N9:* मधुर *madhur 'melody', N14:* बुरा *bura 'poor']* (4.14)

4.4.6 Dependency Graph Generation Phase

In this phase, the dependency graph generated for sentence (4.2) along with the assignment of sentiment nodes to its aspect nodes is shown in Figure 4.3.

Here, red and green vertices of the graph represent aspect nodes and sentiment nodes, respectively with their node *IDs* and names. The edges of the graph represent the relation among the nodes. For example, 'N2' is connected to 'N4' by relation 'r6' and 'N4' is further connected to 'N5' through an intermediate node 'N6'. Then distance from each node of sentiment list for each node of the aspect list is calculated. Lesser the distance, the greater is the probability that node represents the sentiment about aspect node. The sentiment nodes assigned to aspect nodes based on computation of minimum distance are represented with dotted circles in Figure 4.3. For example, distance of sentiment node 'N9' from aspect nodes 'N2', 'N4', 'N8' and 'N12' is calculated. On the basis of minimum distance, sentiment node 'N9' is assigned to aspect node 'N8'.

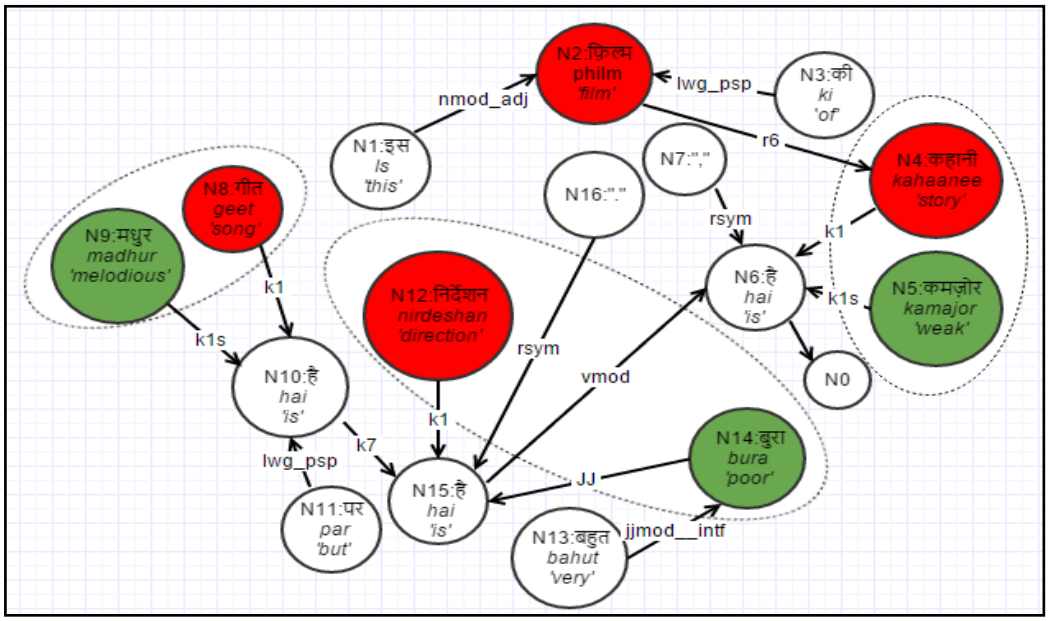


Figure 4.3 Dependency graph for sentence (4.2)

The Procedure 4.4 to illustrate the concept of generation of Dependency Graph and Output List (DGOL) has been given as follows.

PROCEDURE 4.4: DGOL (AL, SL, HDP'_{output})

Input: AL, SL, HDP'_{output}

Output: $DG(v,e), OL$

//Creation of Dependency Graph

$DG(v,e)$ *where* $v \in nodes, e \in edge(node\_id, dep\_node\_id)$

1. **for** $\forall(node \in HDP'_{output})$ **do**

   Create an edge from $node\_id \xrightarrow{dep\_rel} dep\_node\_id$ to obtain DG;

   end for

// Creation of Output list which stores $s\_id's$ corresponding to $a\_id's$

2. $\min = \infty$
3. $for(j = 1; j \le n; j++)$ **do**
4. $for(i = 1; i \le m; i++)$ **do**

   $D = dist(node\_id'_j, node\_id_i)$ *from DG*; // $node\_id_i \in AL, node\_id'_j \in SL$;

   if $(D < \min)$ **then**

   $\min = D$;
   $a\_id = node\_id_j$;
   $s\_id = node\_id_i$;

   end if

   end for

   Add a_id and s_id to $output\_list\ OL$;

   end for

Description: The Procedure 4.4 helps in creation of Dependency Graph (DG) and Output List (OL) by taking *AL*, *SL* and HDP'_{output} as input. A graph *DG* is generated by creating an edge between node and its dependent node with representation of dependency relation on its edge. In next steps, distance of each sentiment node is calculated from the aspect node. The nodes having minimum distance between the sentiment node and aspect node are added to *OL*.

4.4.7 Negations and Intensifiers Handling Phase

The proposed system is also able to handle linguistic features such as intensifiers and negations. These are handled on the basis of work described by Kar and Mandal (2011). To handle the sentences having negation, the polarity of the adjacent sentiment word is negated, if both the negation and sentiment word are connected through the main verb, then its corresponding sentiment score is calculated using equation (4.15).

$$(sw)(neg) = (-1) * scr(sw) \tag{4.15}$$

Here, *sw* represents the sentiment word, *neg* represents the negation and *scr* specifies the score of sentiment word. For example, अच्छा *achchha 'good'* is a positive sentiment word and its score is 0.75 (taken from HSWN) and नहीं *nahin 'not'* is negation, then a score of (अच्छा *achchha 'good'*) (नहीं *nahin 'not'*) will be -0.75 with negative polarity calculated as follows according to the equation (4.15).

(अच्छा *achchha 'good'*) (नहीं *nahin 'not'*) = -1* *scr*(अच्छा *achchha 'good'*)

= -1 * 0.75

= -0.75

Intensifiers (*intf)* in the form of adjectives, verbs and nouns also play an important role as they can increase or decrease the polarity of sentiment words. If there exists any *(jj/v/n)mod_intf* relation, (here, *jj, v and n* represent adjective, verb and noun, respectively) between the sentiment word and intensifier in the output of HDP, then its corresponding score is modified using equation (4.16). The equation (4.16) is based on general property of mathematics, *i.e.*, the square root (square) of a number n increases (decreases) with increase in n when $n \in [0,1]$. Intensifiers increase the score of sentiment words when these appear with positive words and decrease the sentiment score when these are used with negative words. Therefore, to increase the value of positive sentiment words occurring adjacent to intensifiers, the square root of the value of the positive sentiment word is taken and to decrease the value of negative sentiment words, square is taken according to the formula given in (4.16).

$$(intf)(sw) = \begin{cases} \sqrt{scr(sw)} & if\ sw\ is\ positive \\ (scr(sw))^2 & if\ sw\ is\ negative \end{cases} \quad (4.16)$$

According to this equation, if there is any positive sentiment word preceded by an intensifier, then the score of the combination of intensifier and sentiment word will be the square of the score of sentiment word, otherwise it will be the square root of the sentiment word.

For example, in case of (बहुत *bahut 'very')* (अच्छा *achchha 'good'*), there is an intensifier बहुत *bahut 'very'* followed by a positive sentiment word अच्छा *achchha 'good'*, therefore, score of (बहुत *bahut 'very')* (अच्छा *achchha 'good'*) will be the square root of the sentiment score of the word अच्छा *achchha 'good', i.e.,* $\sqrt{0.75} = 0.87$ according to equation (4.16).

4.4.8 Polarity Assignment Phase

In this phase, sentiment polarity is assigned to each feature word on the basis of the sentiment score and final polarity of each sentence is computed. The Procedure 4.5 to describe the process for computation of Polarity Assignment and Overall Score (PAOS) of the sentence is given as follows.

PROCEDURE 4.5: PAOS (OL, HDP'_{output})

Input: OL, HDP'_{output}

Output: $f\_id, s\_id, p\_SCR, n\_SCR, sent, overall\_sent$

// Creation of Output List OL and sentiment polarity of each node

1. **for** $\forall(s\_id \in OL)$ **do**

   $Obtain\ (p\_SCR, n\_SCR) \in s\_id$ of HDP'_{output};

   if $(p\_SCR > n\_SCR)$ **then**

   $sent = pos$;

   else if $(p\_SCR < n\_SCR)$

   $sent = neg$;

   end if

   add $(p\_SCR, n\_SCR, sent)$ to OL;

end for

// Calculation of overall sentiment of sentence

PROCEDURE 4.5: PAOS (OL, HDP'_{output})

2. **for** $\forall(s\_id \in OL)$ **do**

$$p = \sum(p\_scr) \text{ and } n = \sum(n\_scr)$$

end for
if *(p > n)* **then**
overall_sent =pos;
else if *(p < n)* **then**
overall_sent =neg;
else
overall_sent =neu;
end if

Description: The Procedure 4.5 computes the sentiment polarity of each node as well as the overall sentiment of the sentence. If the p_SCR of the node is greater than the n_SCR then polarity of the node is positive otherwise polarity is negative. Similarly, the overall polarity of the sentence is computed by taking summarization of scores of positive as well as negative nodes.

The detailed description about the assignment of sentiment nodes to aspect nodes for example sentence is given in Table 4.6.

Table 4.6: Assignment of sentiment node to aspect node for example sentence

Hindi Sentence: इस फ़िल्म की कहानी कमज़ोर है, गीत मधुर हैं पर निर्देशन बहुत बुरा है। Transliteration: *is philm kee kahaanee kamajor hai, geet madhur hain par nirdeshan bahut bura hai.* Equivalent English Sentence*: 'The story of this film is weak, songs are melodious but the direction is very poor'.*			
Selection of Aspect node close to Sentiment Node *N5:* कमज़ोर *kamajor 'weak'*			
Path Length from Sentiment Node *S* to Aspect Node *F*	**Distance**	**Description**	**Sentiment assigned to aspect**
The path length from node *N5* to *N2:* *फ़िल्म philm 'film'*	3	*N5* is assigned to *N4* as the distance between them is minimum.	Negative
The path length from node N5 to N4: कहानी *kahaanee 'story'*	**2**		
The path length from node *N5* to *N8:* गीत *geet* '*song'*	4		
The path length from node *N5* to *N12:* निर्देशन *nirdeshan* '*direction*'	3		

<table>
<tr><td colspan="4">Hindi Sentence: इस फ़िल्म की कहानी कमज़ोर है, गीत मधुर हैं पर निर्देशन बहुत बुरा है।
Transliteration: is philm kee kahaanee kamajor hai, geet madhur hain par nirdeshan bahut bura hai.
Equivalent English Sentence: 'The story of this film is weak, songs are melodious but the direction is very poor'.</td></tr>
<tr><td colspan="4">Selection of Aspect node close to Sentiment Node N9: मधुर madhur 'melodious'</td></tr>
<tr><td>The path length from node N9 to N2: फ़िल्म philm 'film'</td><td>5</td><td rowspan="4">N9 is assigned to N8 as the distance between them is minimum.</td><td rowspan="4">Positive</td></tr>
<tr><td>The path length from node N9 to N4: कहानी kahaanee 'story'</td><td>4</td></tr>
<tr><td>The path length from node N9 to N8: गीत geet 'song'</td><td>2</td></tr>
<tr><td>The path length from node N9 to N12: निर्देशन nirdeshan 'direction'.</td><td>3</td></tr>
<tr><td colspan="4">Selection of Aspect node close to Sentiment Node N14: बुरा bura 'poor'</td></tr>
<tr><td>The path length from node N14 to N2: फ़िल्म philm 'film'</td><td>4</td><td rowspan="4">N14 is assigned to N12 as the distance between them is minimum.</td><td rowspan="4">Negative</td></tr>
<tr><td>The path length from node N14 to N4: कहानी kahaanee 'story'</td><td>3</td></tr>
<tr><td>The path length from node N14 to N8: गीत geet 'song'</td><td>3</td></tr>
<tr><td>The path length from node N14 to N12: निर्देशन nirdeshan 'direction'</td><td>2</td></tr>
</table>

If there exists any sentiment node in the sentence that does not express any aspect node, then an average sentiment score of positive and negative sentiment nodes is computed and corresponding polarity is assigned to that sentence.

Some example sentences on which aspect-based sentiment analysis has been performed are given as follows.

Example 4.1 : Consider the Hindi sentence given in (4.17) and its equivalent English translation is given in (4.18).

For sentence (4.17), the dependency graph generated is given in Figure 4.4. In this sentence, the sentiment word सुंदर *sundar* '*beautiful*' is assigned to aspect यूआई *yooaee* '*UI*' with positive polarity score 0.75.

Hindi Sentence : इसमें एक सुंदर यूआई भी है। (4.17)

Transliteration : *isamen ek sundar yooaee bhee hai.*

English Translation : '*It also has a beautiful UI.*' (4.18)

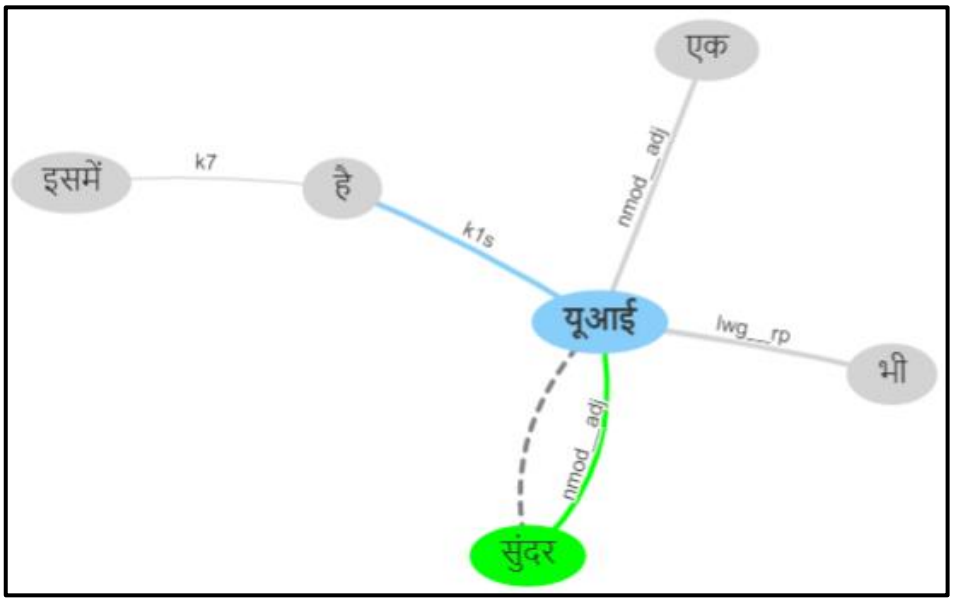


Figure 4.4 Dependency graph of sentence (4.17)

Output : यूआई--- सुंदर (Polarity: Positive, Score: 0.75)

Example 4.2 : Consider the Hindi sentence given in (4.19) and its equivalent English translation is given in (4.20).

For sentence (4.19), the dependency graph generated is given in Figure 4.5. In this sentence, the sentiment word खूबसूरत *khoobasoorat* '*beautiful*' is assigned to aspect कैमरा *kaimara* '*Camera*' with positive polarity score 0.625.

Hindi Sentence: बहुत खूबसूरत दिखने वाला यह कैमरा जल्द ही बाजार में आने वाला है।

(4.19)

Transliteration : *bahut khoobasoorat dikhane vaala yah kaimara jald hee baajaar mein aane vaala hai.*

English Translation: '*This beautiful looking camera is coming to the market soon.*'

(4.20)

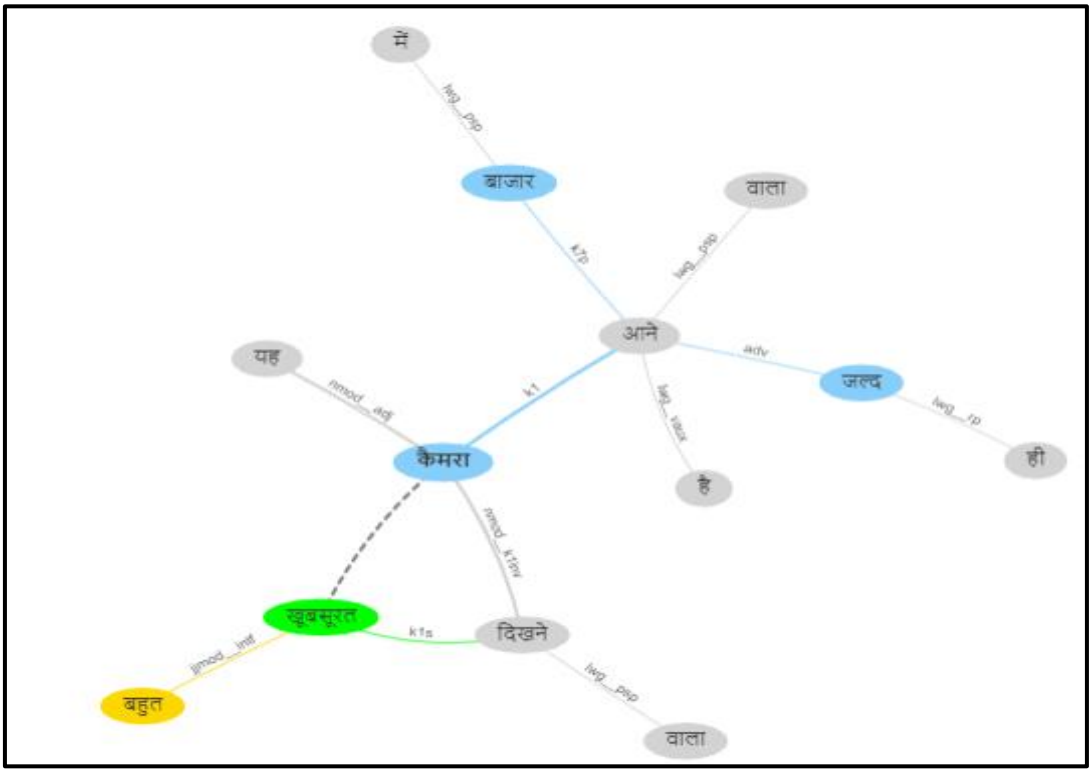


Figure 4.5 Dependency graph of sentence (4.19)

Output : कैमरा--- बहुत--- खूबसूरत (Polarity: Positive, Score: 0.625)

Example 4.3 *:* Consider the Hindi sentence given in (4.21) and its equivalent English translation is given in (4.22).

For sentence (4.21), the dependency graph generated is given in Figure 4.6. In this sentence, the sentiment word अच्छा *achchha* '*good*' is assigned to aspect कीबोर्ड *keebord* '*keyboard*' with positive polarity score 0.625.

Hindi Sentence: इसमें एक बहुत अच्छा कीबोर्ड भी दिया गया है।	(4.21)
Transliteration : *isamen ek bahut achchha keebord bhee diya gaya hai.*	
English Translation : '*It also has a very good keyboard.*'	(4.22)

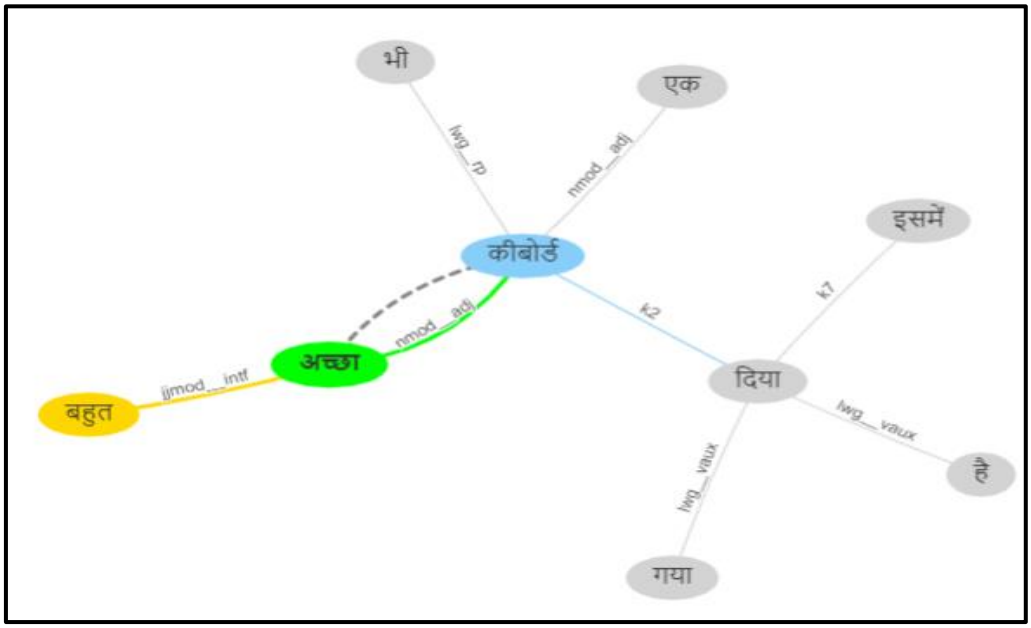


Figure 4.6 Dependency graph of sentence (4.21)

Output : कीबोर्ड--- बहुत--- अच्छा (Polarity: Positive, Score: 0.625)

Example 4.4: Consider the Hindi sentence given in (4.23) and its equivalent English translation is given in (4.24).

For sentence (4.23), the dependency graph generated is given in Figure 4.7. In this sentence, the sentiment words दमदार *damadaar* '*powerful*' and सर्वोत्तम *sarvottam* '*best*' are assigned to aspects नोटबुक *notabuk* '*notebook'* and परफोर्मेश *paraphormesh* '*performers*' having positive polarity scores 0.375 and 0.75, respectively.

Hindi Sentence: डेल ने भारत में दो दमदार नोटबुक लॉन्च किए हैं और वो सर्वोत्तम परफोर्मेश वाले हैं। (4.23)

Transliteration : *del ne bhaarat mein do damadaar notabuk lonch kie hain aur vo sarvottam paraphormesh vaale hain.*

English Translation : '*Dell has launched two powerful notebooks in India and is one of the best performers.*' (4.24)

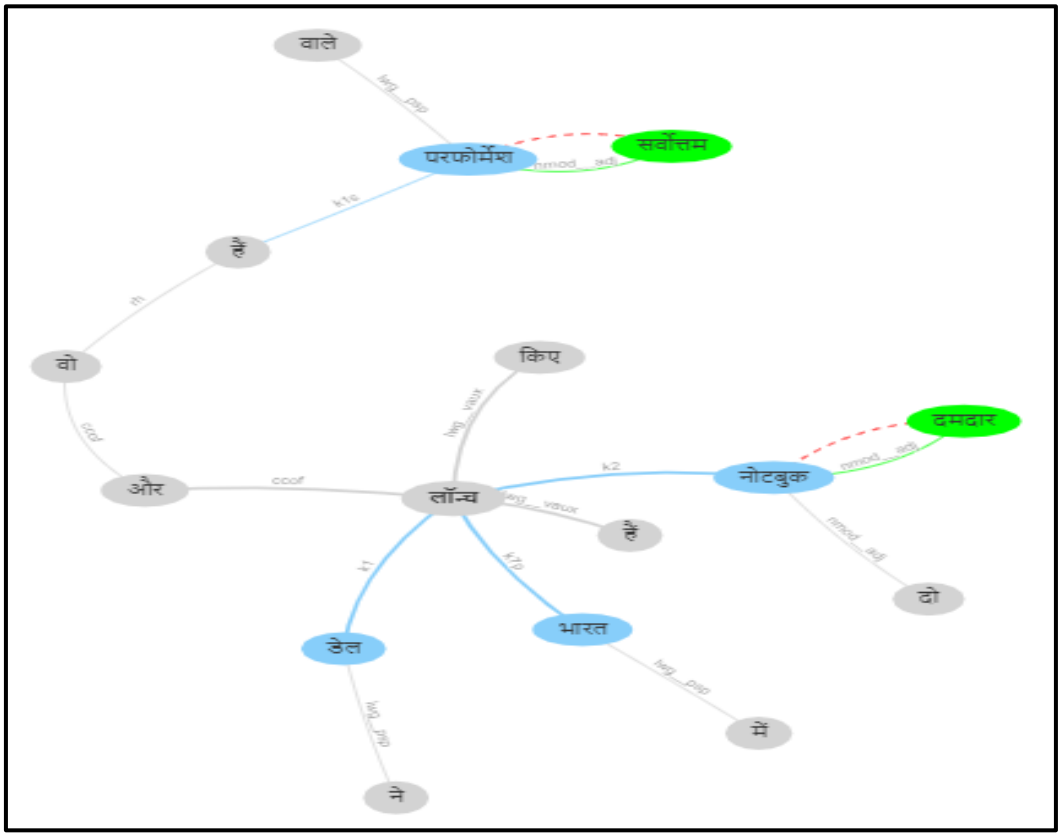


Figure 4.7 Dependency graph of sentence (4.23)

Output : नोटबुक--- दमदार (Polarity: Positive, Score 0.375)

परफोर्मेंश--- सर्वोत्तम (Polarity: Positive, Score: 0.75)

Example 4.5: Consider the Hindi sentence given in (4.25) and its equivalent English translation is given in (4.26).

For sentence (4.25), the dependency graph generated is given in Figure 4.8. In this sentence, the sentiment word आकर्षक aakarshak 'attractive' is assigned to aspect डिस्प्ले *disple* '*display'* with positive polarity score 0.75.

Hindi Sentence: आकर्षक डिस्प्ले है।	(4.25)
Transliteration : *aakarshak disple hai.*	
English Translation : '*Has an attractive display.*'	(4.26)

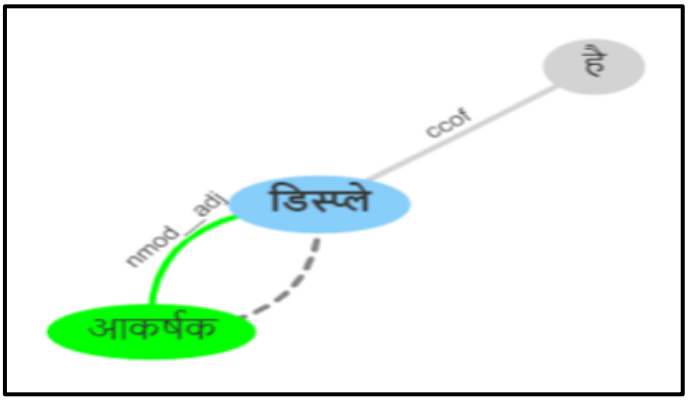


Figure 4.8 Dependency graph of sentence (4.25)

Output : डिस्प्ले--- आकर्षक (Polarity: Positive, Score: 0.75)

Example 4.6: Consider the Hindi sentence given in (4.27) and its equivalent English translation is given in (4.38).

Hindi Sentence: इसके अलावा टच अहसास भी बहुत अच्छा है। (4.27)

Transliteration : *isake alaava tach ahasaas bhee bahut achchha hai.*

English Translation : '*Apart from this, the touch feeling is also very good.*' (4.28)

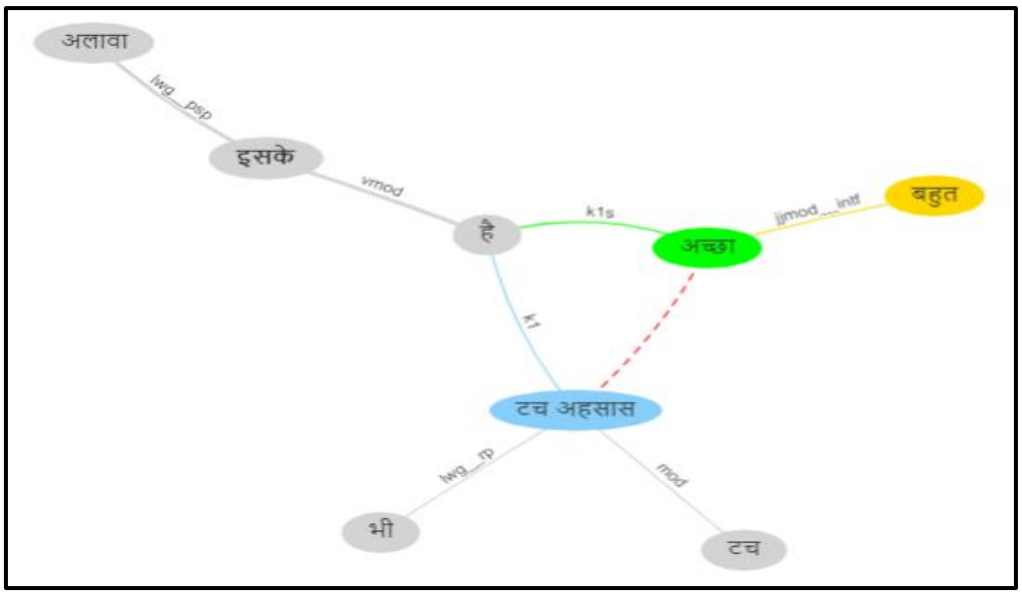


Figure 4.9 Dependency graph of sentence (4.27)

Output : टच अहसास--- बहुत--- अच्छा (Polarity: Positive, Score: 0.625)

For sentence (4.27), the dependency graph generated is given in Figure 4.9. In this sentence, the sentiment word अच्छा *achchha* '*good*' is assigned to aspect टच अहसास *tach ahasaas* '*touch feeling'* with positive polarity score 0.625.

Example 4.7: Consider the Hindi sentence given in (4.29) and its equivalent English translation is given in (4.30).

For sentence (4.29), the dependency graph generated is given in Figure 4.10. In this sentence, the sentiment word अच्छा *achchha* '*good*' is assigned to aspect व्यूइंग एंगल *vyooing engal* '*viewing angle'* and then negated due to presence of negation नहीं *nahin* 'not' in the sentence therefore, negative polarity score -0.75 is assigned to aspect word.

Hindi Sentence: स्क्रीन का व्यूइंग एंगल कहीं से अच्छा नहीं माना जा सकता। (4.29)

Transliteration : *skreen ka vyooing engal kaheen se achchha nahin maana ja sakata.*

English Translation : '*The viewing angle of the screen cannot be considered good from anywhere.*' (4.30)

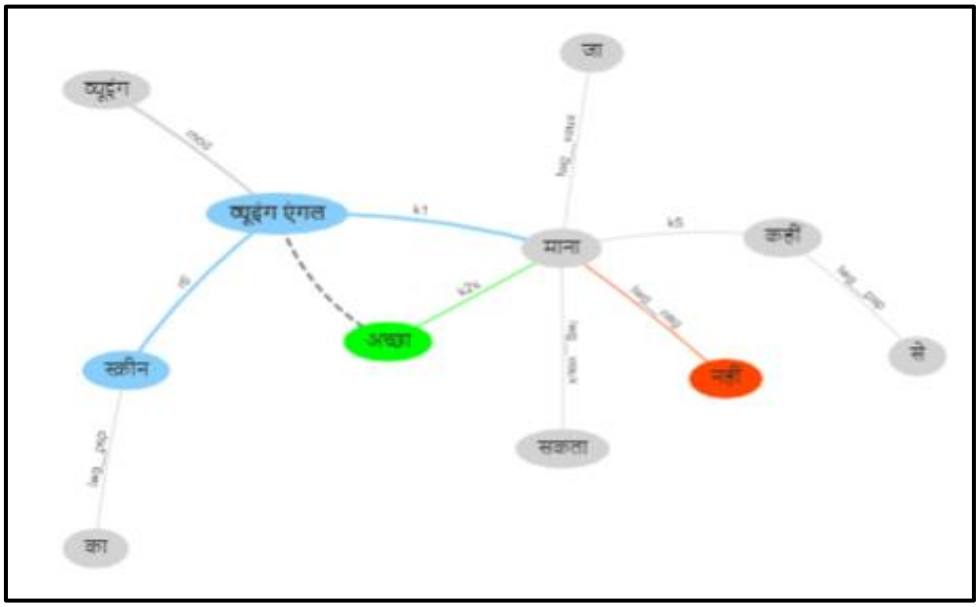


Figure 4.10 Dependency graph of sentence (4.29)

Output : व्यूइंग एंगल--- अच्छा ---नहीं (Polarity: Negative, Score: -0.75)

Example 4.8: Consider the Hindi sentence given in (4.31) and its equivalent English translation is given in (4.32).

For sentence (4.31), the dependency graph generated is given in Figure 4.11. In this sentence, the sentiment word दमदार *damadaar* '*strong*' is assigned to aspect परफॉर्मेंस *paraphormens* '*performance'* with positive polarity score 0.375.

Hindi Sentence: इसके फंक्शंस की परफॉर्मेंस दमदार है। (4.31)

Transliteration : *isake phankshans kee paraphormens damadaar hai.*

English Translation : '*The performance of its functions is strong.*' (4.32)

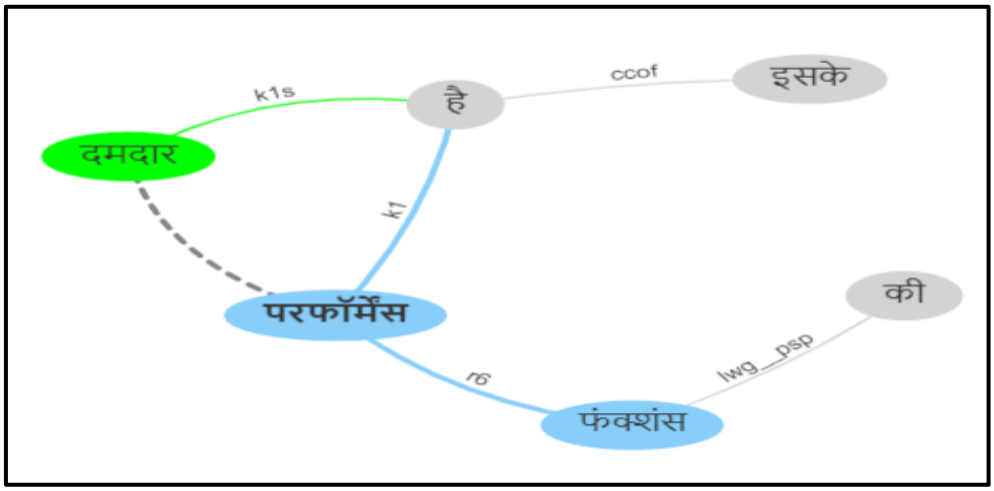


Figure 4.11 Dependency graph of sentence (4.31)

Output: परफॉर्मेंस--- दमदार (Polarity: Positive, Score: 0.375)

Example 4.9: Consider the Hindi sentence given in (4.33) and its equivalent English translation is given in (4.34).

For sentence (4.33), the dependency graph generated is given in Figure 4.12. In this sentence, the sentiment word जानदार *jaanadaar* '*lively*' is assigned to aspect कैमरा *kaimara* '*camera*' and then negated due to presence of negation नहीं *nahin* '*not*' in the sentence therefore, negative polarity score -0.625 is assigned to aspect word.

Hindi Sentence: कैमरा भी बहुत जानदार नहीं है। (4.33)

Transliteration : *kaimara bhee bahut jaanadaar nahin hai.*

English Translation : '*The camera is also not very lively.*' (4.34)

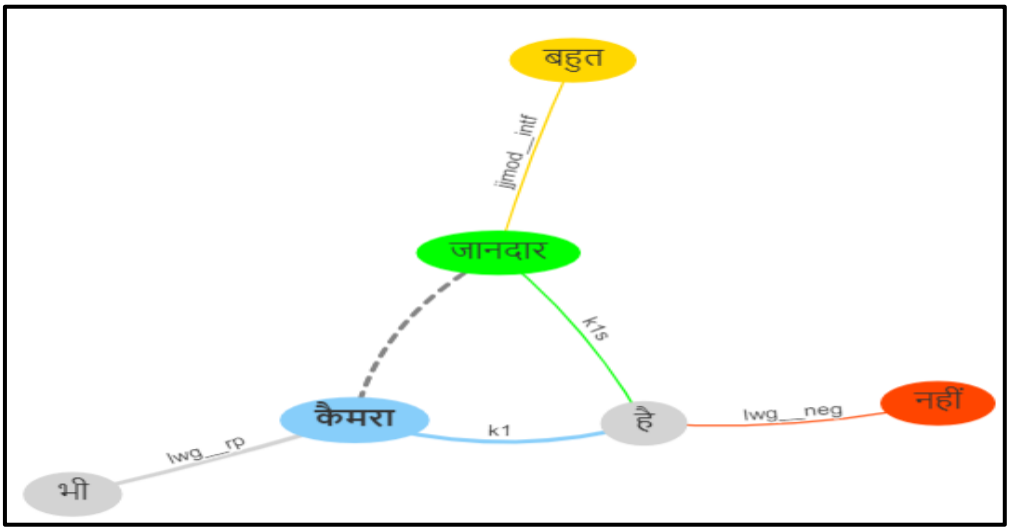


Figure 4.12 Dependency graph of sentence (4.33)

Output : कैमरा--- बहुत--- जानदार ---नहीं (Polarity: Negative, Score: -0.625)

Example 4.10: Consider the Hindi sentence given in (4.35) and its equivalent English translation is given in (4.36).

For sentence (4.35), the dependency graph generated is given in Figure 4.13. In this sentence, the sentiment word अच्छा *achchha* '*good*' is assigned to aspect कैमरा *kaimara* '*camera*' with positive polarity and the sentiment word खराब *kharaab* '*bad*' is assigned to aspect बैटरी *baitaree* '*battery*' with negative polarity.

Hindi Sentence: फ़ोन का कैमरा अच्छा है पर बैटरी खराब है । (4.35)

Transliteration: *fon ka kaimara achchha hai par baitaree kharaab hai.*

English Translation: '*The camera of the phone is good but the battery is bad.* (4.36)

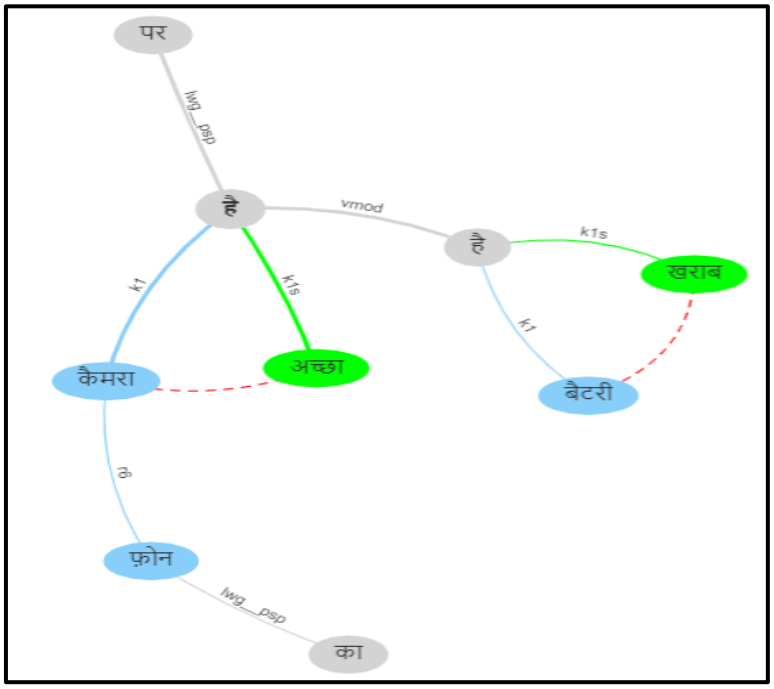


Figure 4.13 Dependency graph of sentence (4.35)

Output : कैमरा--- अच्छा (Polarity: Positive, Score : 0.75)

बैटरी--- खराब (Polarity: Negative, Score : -0.625)

Example 4.11: Consider the Hindi sentence given in (4.37) and its equivalent English translation is given in (4.38).

For sentence (4.37), the dependency graph generated is given in Figure 4.14. In this sentence, the sentiment word खराब *kharaab* '*poor*' is assigned to aspect सर्विस *sarvis* '*service'* with negative polarity score -0.625.

Hindi Sentence: एचपी/कॉम्पैक का सर्विस भी बेहद खराब है। (4.37)

Transliteration : *echape/kompaik ka sarvis bhee behad kharaab hai.*

English Translation : '*HP/Compaq's service is also very poor.*' (4.38)

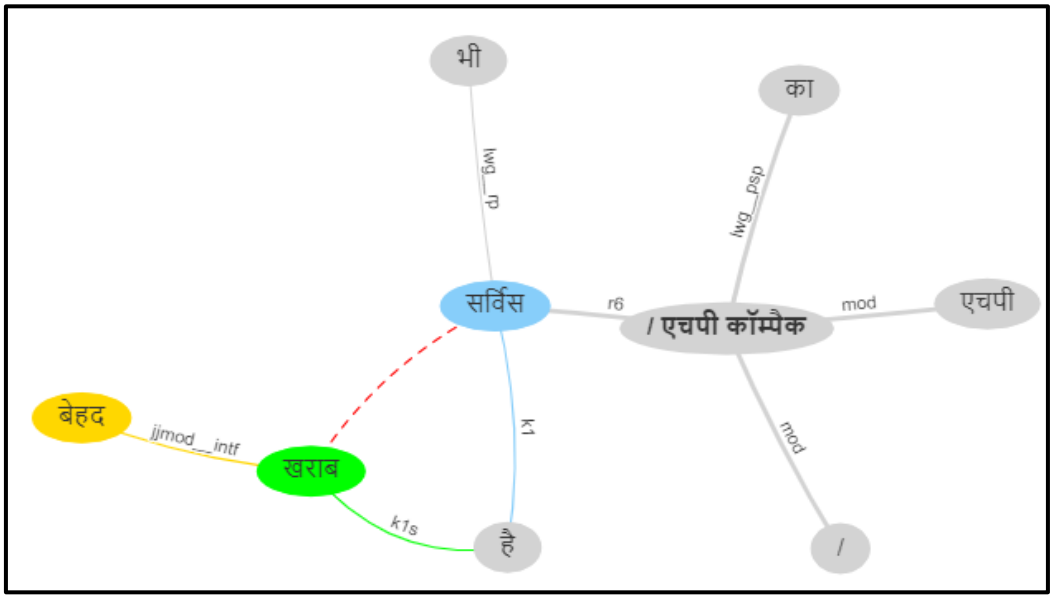


Figure 4.14 Dependency graph of sentence (4.37)

Output : सर्विस--- बेहद--- खराब (Polarity: Negative, Score: -0.625

4.5 Testing

The testing of the proposed aspect-based sentiment analysis system has been performed manually at sentence level. For testing, first of all, the corpus of movie reviews has been manually tagged into positive, negative and neutral class. For the task of tagging, the sentences having either one positive or one negative sentiment word are tagged depending upon the polarity score of that positive or negative sentiment word, while the sentences consisting of more than one positive or negative sentiment words in a sentence are also tagged depending upon the polarity of that sentiment word. However, the sentences consisting of both positive and negative sentiment words in a sentence are tagged by calculating the average of summation of polarity scores of both positive and negative sentiment words and polarity is assigned to that sentence depending upon the final polarity score.

The results of the sentiments assigned by the proposed system are compared with manual testing to measure the performance of the system. Out of 2247 sentences, the system has identified 602 positive, 558 negative and 1087 neutral sentences. It has been observed that

out of 2247 sentences, the system has correctly identified the sentiments of 1871 sentences and has achieved an accuracy of 83.2%. The results of the proposed system are compared with manual testing as given in Table 4.7 and its confusion matrix is represented in Table 4.8.

Table 4.7: Comparison of proposed system with manual testing

Polarity of Sentence	**Number of sentences identified while manual tagging**	**Number of sentences identified by the proposed system**	**Number of sentences in which the results agree with each other**
Positive	676	602	516
Negative	584	558	501
Neutral	987	1087	854
Total	2247	2247	1871

Table 4.8: Confusion matrix

	Predicted				
		A	**B**	**C**	**Total**
Actual	A = Positive	516	8	152	676
	B = Negative	2	501	81	584
	C = Neutral	84	49	854	987

Table 4.9 represents the polarity wise evaluation of the proposed system using performance measures like recall, precision and F-measure, and its bar chart is represented in Figure 4.15.

Table 4.9: Polarity wise evaluation of the proposed system

Polarity	**Recall**	**Precision**	**F-measure**
Positive	0.763	0.857	0.807
Negative	0.858	0.898	0.878
Neutral	0.865	0.786	0.824
Average	0.829	0.847	0.836

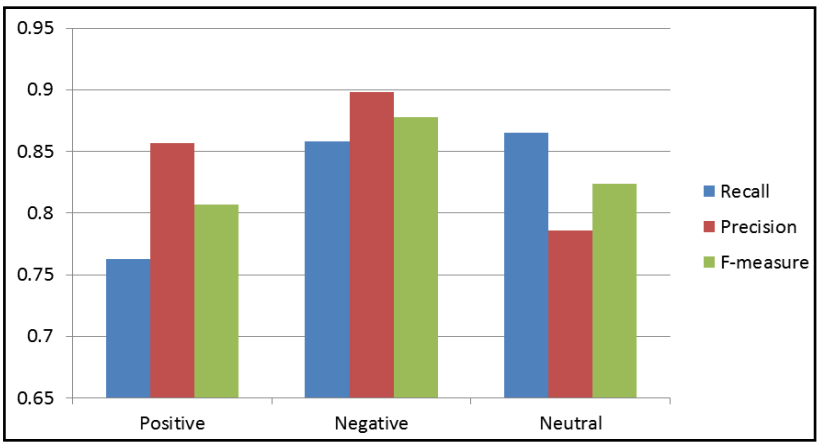


Figure 4.15 Polarity wise evaluation of proposed system

4.6 Comparison

As reported earlier, a little research work has been done on aspect-based SA for Hindi language. Therefore, the evaluation of the proposed approach has been performed by comparing it with the traditional lexicon based as well as with existing works on aspect-based SA for Hindi language.

4.6.1 Comparison with Traditional Lexicon based Approaches

In this section, the proposed aspect-based approach is compared with traditional lexicon based approaches. The traditional lexicon based approaches that are used for comparison are Sentiment Word Count (SWC), Prior Sentiment Score (PSS) and a lexicon based approach which uses a library of National Research Council Canada (NRC) emotion lexicon in "*R*" language. The presented traditional lexicon based approaches work at sentence level and do not use any parsing. This is to determine whether the parsing used in proposed approach provides any improvement in accuracy. The detailed analysis of this comparison is given as follows.

Comparsion with traditional SWC approach: In traditional SWC approach, each word of tokenized sentence is matched with entries of lexicon and the count of positive and negative sentiment words is taken and final sentiment classification is done on the basis of the count of sentiment words. If the count of positive words is greater (less) than that of negative words, it is considered as positive (negative).

For example, consider the Hindi sentence given in (4.39) and its English translation is given in (4.40).

Hindi Sentence: इस फ़िल्म की कहानी कमज़ोर है, गीत मधुर हैं पर निर्देशन बहुत बुरा है। (4.39)

Transliteration: *is philm kee kahaanee kamajor hai, geet madhur hain par nirdeshan bahut bura hai.*

Equivalent English Sentence*: 'The story of this film is weak, songs are melodious but the direction is very poor'.* (4.40)

In this case, SWC approach results the sentiment "negative" for sentence (4.39) as it consists of two negative words कमज़ोर *kamajor 'weak'* and बुरा *bura 'poor'* whose count is greater than positive word मधुर *madhur 'melodious'*. The major drawback of this approach is that sentiment of a sentence will be classified as 'neutral', if there is an equal count of positive and negative words.

Comparsion with traditional PSS approach: The traditional PSS approach determines the prior sentiment score of words from existing lexicon. Instead of counting positive and negative sentiment words as in the first approach, it sums up the sentiment score of positive and negative sentiment words in each sentence, and the final sentiment classification is done on the basis of the total sentiment score. If the total score is positive (negative), it is considered as positive (negative). For example, consider the Hindi sentence given in (4.41) and its English translation is given in (4.42).

Hindi Sentence: इस फ़िल्म की कहानी कमज़ोर है पर गीत सुंदर हैं। (4.41)

Transliteration: *Is philm kee kahaanee kamajor hai, par geet sundar hain.*

Equivalent English Translation: *'The story of this film is weak, but the songs are beautiful.'* (4.42)

In sentence (4.41), the negative sentiment word (*i.e.,* कमज़ोर *kamajor 'weak'*) has a polarity score of -0.5 and positive sentiment word (*i.e.,*सुंदर *sundar 'beautiful'*) has a polarity score of 0.4. Therefore, this sentence is classified as 'negative' according to this approach as score of negative sentiment word is greater than positive sentiment word. The

major drawback of PSS approach is that when positive and negative scores are equal and sentiment score of the sentence gets neutralized while computing the overall score.

Comparsion with NRC lexicon based approach: In this approach, NRC lexicon developed by Mohammad and Turney (2013), a lexical resource to identify sentiment polarity is used. It is also known as EmoLex. It supports 40 languages, including some Indian languages like Hindi, Tamil, Gujarati, Marathi, Urdu, *etc*. NRC lexicon includes annotations for 14,182 unigram words for the English language and 8,116 for Hindi language. It is implemented using '*get_nrc_sentiment'* function of '*syuzhet*' package in "*R*" language and an NRC vector ($\vec{E}$) for each word that matches with lexicon. Each NRC vector contains a Boolean value for each of the eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) and two sentiments (positive and negative). If a token is matched with a token in NRC lexicon, corresponding emotion vector is returned. If more than one token are matched, then the sum of their corresponding emotion and sentiment values is returned. It increases the value of the ten (eight emotion + two sentiment) parameters of the NRC vector by 1 depending upon the number of positive and negative words found in lexicon. The drawback of this approach is that it does not handle negations and intensifiers.

The major drawback of traditional lexicon-based approaches such as SWC, PSS and NRC emotion lexicon based is that these approaches do not handle the negations and intensifiers in a sentence and also gives overall positive and negative sentiment polarity instead of giving aspect-based sentiment analysis unlike proposed approach. Such type of drawbacks is overcome by aspect-based approach as it assigns sentiment polarity to each aspect instead of giving overall polarity to the sentence and also the proposed approach is able to handle the linguistic issues, like handling of transliteration, negations and intensifiers**.**

Table 4.10 presents the comparison of the accuracy of the proposed approach with accuracies of its traditional lexicon based approaches for Hindi.

Table 4.10: Comparison of the proposed approach with traditioanl lexicon based approaches

Approach	**Accuracy**
SWC using HSWN	72.7%
PSS using HSWN	75.2%
NRC emotion lexicon based	71.9%
Proposed approach	83.2%

4.6.2 Comparison with Existing Aspect-based SA Works

The proposed approach is also compared with other existing works on feature/aspect based sentiment analysis as shown in Table 4.11.

Table 4.11 : Comparison of proposed approach with existing works for Hindi language

Author	**Domain**	**Corpus Size**	**Accuracy**
Bakliwal et al., (2012)	Products Reviews	700 reviews	79.03%
Mittal et al., (2013)	Movie Reviews	662 reviews	80.21%
Sharma et al., (2014)	Movie Reviews	Not Specified	65%
Sharma et al., (2015)	Tweets	100 tweets	77.75%
Akhtar et al., (2016)	Products/Services reviews	5417 reviews	65.96%
Sharma and Moh (2016)	Election Tweets	42,235 tweets	34%
Garg and Buttar (2017)	General	201 reviews	82.3%
Hussaini et al., (2018)	Book Reviews	700 sentences	82.4%
Our Approach	**Movie Reviews**	**2,247 sentences**	**83.2%**

From the comparative analysis, it is analyzed that proposed aspect-based lexicon approach has been tested on corpus of movie reviews with corpus size 2,247 sentences and is able to achieve an accurcay of 83.2% which shows that the proposed approach outperforms when compared to the existing aspect-based approaches for SA of Hindi.

Figure 4.16 presents the comaprsison of accuracies of existing works on aspect-based sentiment analysis for Hindi language on different domains and size of corpra using lexicon-based approach

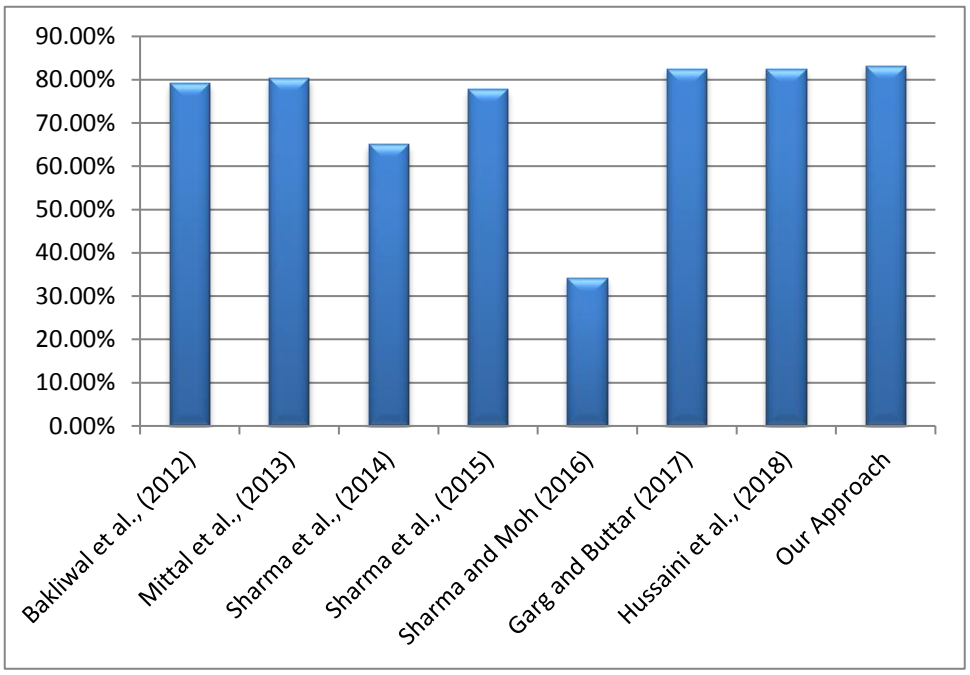


Figure 4.16 Comparison of accuracy of proposed approach with existing works

4.7 Error analysis

In this section, errorneous cases in sentiment classification are discussed. Misclassification of the sentences is due to some limiting factors which affect the accuracy of the system. The proposed system relies on the accuracy of the Hindi Dependency Parser. If HDP does not identify the POS accurately, then the feature list and the sentiment list will not be prepared correctly by the proposed system as discussed in Section 4.4.4. The system also depends upon the coverage of sentiment words present in HSWN. Sometimes people use Hinglish text in their context which is difficult to handle. For example, consider the sentence given in (4.43) and its corresponding English translation is given in (4.44).

Hindi Sentence: फ़िल्म की कहानी थ्रिलर है। (4.43)

Transliteration: *film kee kahaanee thrilar hai.*

Equivalent English Translation: '*The story of the film is thriller.*' (4.44)

In sentence (4.43), the word थ्रिलर thrilar 'thriller' is Hinglish word and it is an adjective in this sentence. But when this sentence is parsed by the HDP, it does not identify it as an adjective which further affects the accuracy of the system. For this, a translation dictionary or Google API can be used to handle these type of issues. Also, one can write a word with different spelling variations in case of Hindi language which arised challenge while performing sentiment analysis. For example, the word महँगा *mahanga* '*expensive*' can be written with a different spelling variation such as महंगा *mahanga* '*expensive*'. To handle these issues, a list of sentiment words has been manually added to the existing lexicon. However, the coverage of HSWN lexicon is still an issue and it can further be enhanced to improve the accuracy of the system. Including the issues of linguistic resources like sentiment lexicons and dependency parsers, the proposed system also lacks in some cases like handling of multiword expressions as given in sentence (4.45) with its corresponding English translation given in (4.46) and the sentences consisting of more than one sentiment word for one aspect.

Hindi Sentence: फ़िल्म की कहानी यथार्थ से परे है। (4.45)

Transliteration: *philm kee kahaanee yathaarth se pare hai.*

Equivalent English Translation: '*The film is beyond reality.*' (4.46)

The sentence (4.45) represents the negative sentiment in general, but the system classifies it as 'neutral' due to lack of handling of multiword expressions, i.e., यथार्थ से परे yathaarth se pare 'beyond reality' by the proposed system.

As the proposed system has been tested for domain-independent corpus therefore, the accuracy can further be improved by considering only domain dependent features, *i.e.,* by improving aspect list according to the domain. However, the proposed system has achieved promising results which can help the users about the positive and negative opinions of different aspects about an entity.

Chapter Summary

In this chapter, aspect-based sentiment analysis system for Hindi language is proposed. The system has been experimented on reviews dataset about products and movies in Hindi language. The proposed system uses two lexical resources HDP and HSWN. It follows an efficient aspect extraction process to extract all the relevant aspects which include three steps, i.e., extraction of frequent nouns, identification of relevant nouns and removal of irrelevant nouns. The sentiment nodes are extracted using HSWN. The system uses HDP to determine the association between the aspect nodes and sentiment nodes. Also, the system generates a dependency graph and assigns the sentiment to the particular aspect having the least distance between sentiment word and aspect word. It is observed that the system has achieved an accuracy of 83.2%. The precision, recall and F-measure of the system are 0.85, 0.83 and 0.84 respectively. The results of the proposed system are compared with its traditional lexicon based approaches, and also with existing works on the aspect-based SA. The error analysis discusses about errornous cases arised while performing SA. From the performance of the system, it has been analyzed that the proposed system can benefit the people in getting the different opinions as well as overall opinion about an entity.

Chapter 5

Sentiment Analysis System for Education: A Case Study

5.1 Background

In the teaching and learning process, it is crucial to evaluate the teaching performance. This evaluation is one of the most complex processes in any University, since various factors and criteria should be met to be concentrated in order to provide a final assessment to the professional. It not only helps in improving the course contents and quality but is also often used during the annual appraisal process of teachers. Teacher evaluation can be performed by an observation guide or a rubric with different evaluation criteria.

Many Universities obtain such feedback via a student response system (SRS) during or at the end of a course to analyze the teacher's performance (Kechaou et al., 2011). Student feedback about teacher performance, the learning experience, and other course attributes can also be gathered through social media. In recent years, online learning portals like Coursera, FutureLearn, Udemy, Udacity *etc*. have attracted many students by providing free courses from a growing number of selected institutions (Munezero et al., 2013). Millions of students join these massive open online courses each year and share their opinions about the course content and quality of teaching on the course's discussion forum. Students also comment about their educational experiences in blogs, online forums such as College Confidential (www.collegeconfidential.com), and teacher review sites such as Rate My Professors (www.ratemyprofessors.com) (Altrabsheh et al., 2014). This feedback not only yields useful insights for University administrators and instructors but also plays a key role in influencing student decisions on which Universities to attend or courses to take. This feedback can also assist the instructors in revisiting the course contents from the ethical and technical challenges raised by students (Meldal et al., 2008). However, when teacher performance is evaluated by students, varied opinions are collected from the same established criteria. It is at this time when emerges the need to use sentiment analysis methods to the analysis these comments. Thus, SA plays an important role in education domain, where student feedback is important to evaluate the

use of learning technologies (Altrabsheh et al., 2013). Figure 5.1 represents the architecture of proposed sentiment analysis system for education domain.

As shown in figure, students give feedback in the form of comments or questionnaire about their learning activities and teaching performance as they experience different emotions or sentiments during their learning activities. This feedback is used to predict the sentiment polarity and emotions using sentiment analysis. The predicted sentiment polarity and emotions in the form of graphs or bar charts is shown to teachers to improve their teaching quality so that students can get the best lecture and improve their learning activities which will further help the administrative to take corrective actions.

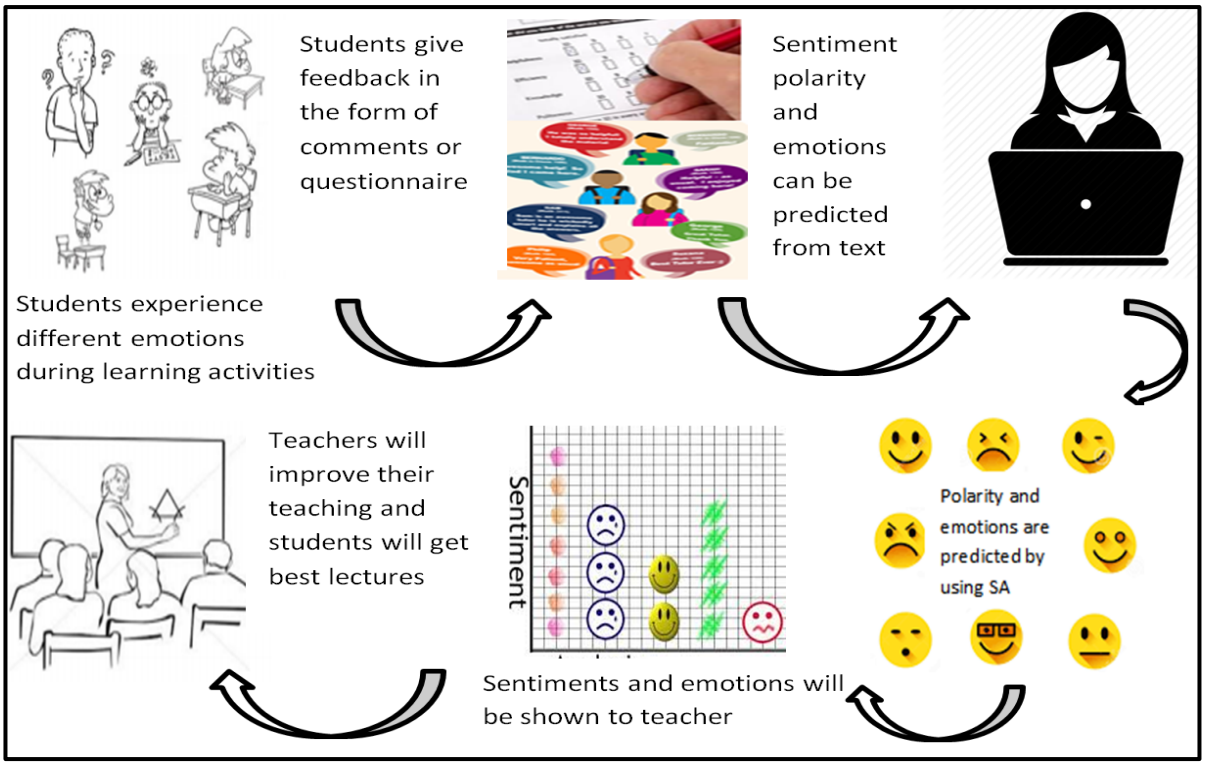


Figure 5.1 Architecture of sentiment analysis of students' feedback

5.2 Types of Assessment

Assessment is defined as all procedures that evaluate student's knowledge, understanding, abilities, or skills according to the quality assurance agency for higher education. Students receive feedback from their teachers during the semester and at the end of semester. The feedback received at the end of semester is called summative assessment that sums up all

student achievements and leads towards awarding the final grade, while the feedback which is given to student during the semester with the main aim to recognize students strengths and weaknesses to guide them accordingly (Garrison and Ehringhaus, 2007). Assessment is of two types direct and indirect.

Direct assessment type includes written and oral exams, essays, reports, portfolios, quizzes, presentation, projects, posters, theses, and many more. All form of assessment has general advantages and disadvantages. However, advantages to student A can be disadvantage to student B. For example, projects show depth of learning which suits a good coder who has great coding skills, but not to other students who don't have good coding skills. Students' perceptions of assessment remarkably affected their approaches to learning and studying, this means that different design and style of assessment would guide them in choosing the right method of studying to achieve better results. Assessment influences learning in three ways such as provides motivation to students; highlights the important part of the curriculum; and helps students to evaluate and judge the effectiveness of their learning. Despite the importance of Students' feedback on assessment as a constructive act of learners in higher education, there are relatively limited studies and reviews that consider students' perspectives (Evans, 2013).

Indirect assessment is based upon student observations of the learning experience and teaching quality. Institutions seek students' feedback using questions-based surveys in which choices are provided to choose from and also in the form the comments. The choices provided in question-based surveys are based on teacher's punctuality, management capability of class and syllabus, way of communication, fair checking and evaluation method of tests, quizzes and answer scripts *etc*. Students can also provide their feedback in the form of comments.

5.3 Sentiment Analysis of Student Assessment and its Challenges

SA of student feedback is a form of indirect assessment that analyzes text written by students. This feedback is collected from online platforms in the form of formal course surveys or informal comments. The sentiment analysis of student feedback helps in

determining students' interest in a class and to identify areas that could be improved through corrective actions. SA raises many technical challenges.

- The meaning of a word varies across different domains. For example, in an education context the word "early" connotes a negative sentiment in the sentence "The lecture is too early!" but in a consumer context it connotes a positive one in the sentence "The courier arrived early."
- It can be difficult to perform SA on text in different languages. In India, for example, people often express their opinions using a transliterated form of Hindi. For example, consider the following student comment given in (5.1) and its transliteration and corresponding English translation is given in (5.2) and (5.3) respectively.

  Student comment: Wo class mein achha padhate hain. (5.1)

  Transliteration: *वो क्लास में अच्छा पढ़ाते हैं।* (5.2)

  English translation: '*He teaches well in the class.*' (5.3)
- Sometimes people use Hinglish words while providing feedback. For example, consider the following student comment given in (5.4) and and its transliteration and corresponding English translation is given in (5.5) and (5.6) respectively.

  Student comment: Wo class mein boring padhate hain. (5.4)

  Transliteration: *वो क्लास में बोरिंग पढ़ाते हैं।* (5.5)

  English translation: '*He teaches boring in the class.*' (5.6)

  In sentence (5.4), "boring" is a Hinglish word. These types of words need to be translated first for further processing.
- Most SA studies have focused on user-review corpora—for example, product, movie, and hotel reviews—with researchers generally classifying the reviews into positive, negative, and sometimes neutral. SA has not been extensively applied to education, though work in this area has grown recently as described in the next section.
- Most of the approaches limit the classification of sentiments to the two or three categories without considering the wide range of emotions that can also affect student feedback.

- Also, no sentiment analysis system is available that can process multilingual data.
- Till now, researchers have not attempted to validate their systems by comparing the results of their analysis with those of traditional direct-assessment methods.

These types of challenges motivate the need to develop a context sensitive, multilingual SA system.

5.4 Related Work

In recent years, researchers have begun to apply sentiment analysis (SA) to the education field using various machine learning and natural language processing techniques. Some of the research works in this field are summarized as follows.

Kechaou et al., (2011) performed sentiment classification of e-learning blogs and forums using a supervised hybrid technique that combined hidden Markov models with Support Vector Machines (SVMs). They performed experiments using three feature-selection methods—mutual information, information gain, and chi statistics; and determined that the chi-statistics method outperformed the other two. Munezero et al., (2013) performed emotion analysis of student learning diaries and classified them into Robert Plutchik's eight emotion categories. They also computed frustration and anxiety from these eight emotions. Altrabsheh et al., (2014) performed SA of student feedback using Naive Bayes (NB), Complement NB (CNB), SVM, and Maximum Entropy (ME) classifiers using unigrams as features. They concluded that an SVM with a radial basis function kernel and the CNB technique achieved good results for real-time feedback analysis. They also observed better performance without including the neutral class. Patel et al., (2015) analyzed feedback from meetings of students' parents using the General Architecture for Text Engineering (GATE) tool and its ANNIE application to classify comments as positive or negative.

Balahadia et al., (2016) proposed an SA system to evaluate teacher performance in courses from student responses in both English and Filipino. They calculated sentiment scores from qualitative and quantitative response ratings using an NB algorithm and graphically represented the percentage of positive and negative sentiments to help

University administrators be aware of students' concerns. Dhanalakshmi et al., (2016) performed SA on feedback from a student evaluation survey of Middle East College in Oman. They used the Rapid Miner tool to classify the comments into positive and negative on the basis of features like teacher, exam, module content, and resources. The researchers compared the performance of their approach using NB, SVM, k-nearest neighbors, and neural-network classifiers. Mishra and Sahoo (2016) used CUDA C programming with GPU architecture to evaluate faculty performance. They categorized faculty members as excellent, very good, good, average, or poor on the basis of average marks given by students in feedback form. The researchers favorably compared their approach in terms of time execution to a similar performance evaluation using CPU architecture. Esparza et al., (2016) proposed a model for SA of student tweets about teacher performance in Spanish. They used an SVM algorithm to classify the tweets into positive, negative, and neutral; they also proposed a syntactic pattern model to compare results using SVMs and syntactic patterns. Aung and Myo (2017) used lexicon based approach to perform the sentiment analysis of students' feedback comments by constructing their sentiment words database in education domain. They classified the sentiment results into different classes such as strongly negative, or moderately negative or weakly negative, or strongly positive, or moderately positive, or weakly positive, or neutral. Nasim et al., (2017) used a hybrid approach of machine learning and lexicon based to perform SA of students' feedback from University SRS. They classified the feedback into three classes such as positive, negative and neutral. It was analyzed that their approach performs better than existing available APIs. Yu et al. (2018) performed SA on self-evaluated comments using SVM and CNN model for early identification of at-risk students. They developed Dynamic Diagnostic and Self-regulated (DDS) system that provides useful information for self-regulated learning with positive emotions, and the line chart of the valence value provides instructors with important information for the application of external guidance to students with negative emotions. Esparza et al., (2018) used SVM with three kernels such as linear, radial and polynomial, to predict a classification of comments in positive, negative or neutral for analyzing teacher's performance.

From the literature, it has been observed that researchers have used different lexicon, machine and deep learning based techniques to perform sentiment analysis of students' feedback. However, the results given by the existing systems are not analyzed properly by comparing them with the results given by direct assessment methods to check the correlation between direct and indirect assessment methods of feedback. The proposed sentiment analysis system developed for education domain is presented in next section.

5.5 Proposed System

The proposed SA system helps to improve teaching and learning by performing temporal sentiment and emotion analysis of student feedback in terms of teacher performance and course satisfaction. Figure 5.2 shows the system architecture, which has five main components: data collection, data preprocessing, sentiment and emotion identification, satisfaction and dissatisfaction computation, and data visualization. The proposed system has been implemented using the open source R language to perform data preprocessing and sentiment classification.

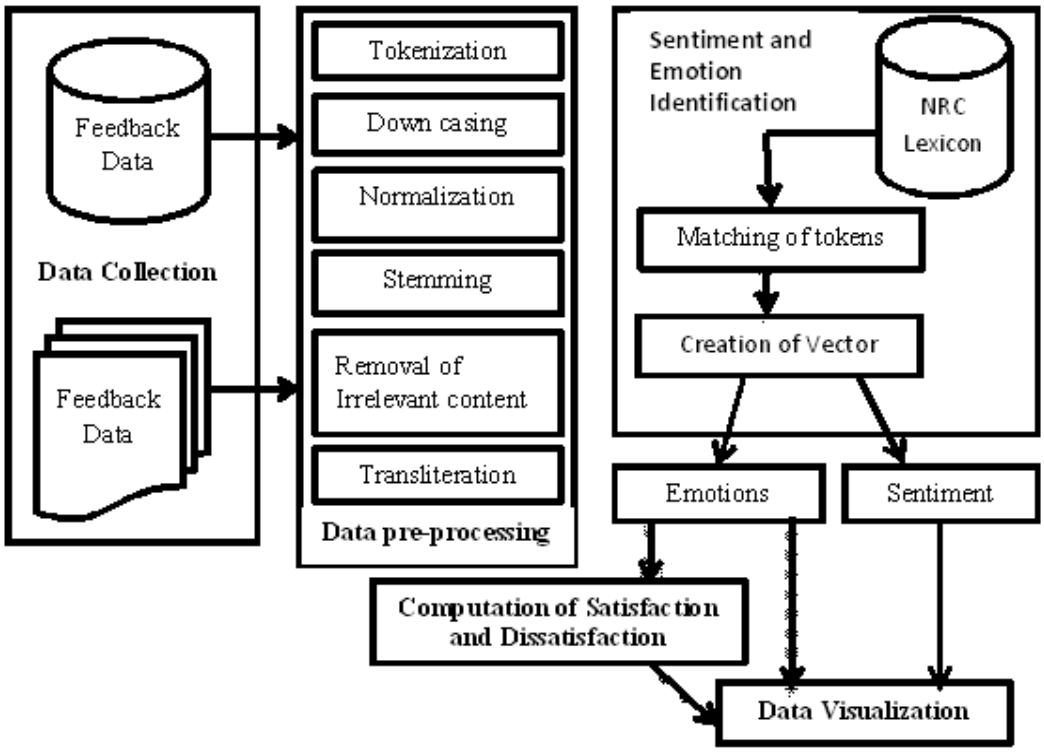


Figure 5.2 Proposed sentiment analysis system for education domain

The first component data collection collects student feedback from both formal sources such as course surveys and informal sources such as blogs and forums. After data collection, the system performs the pre-processing of data using different techniques such as transliteration, tokenization, down-casing, normalization, stemming and removal of irrelevant content *etc.* After preprocessing of input data, the system uses natural language processing in conjunction with the NRC Emotion Lexicon to classify sentiments and emotions. Sentiments are classified into two categories, positive and negative, and emotions are classified into one of Robert Plutchik's eight categories such as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust from which the system computes satisfaction or dissatisfaction. The proposed SA system also includes a data visualization component to facilitate analysis.

The detailed description about each component of proposed sentiment analysis system is given as follows.

5.5.1 Data Collection

The initial data corpus consists of student feedback about a Coursera course as well as data obtained from a University SRS. The Coursera dataset includes approximately 4,000 student comments made during the course, which ran from August 2015 to August 2016, and 1,700 student comments made after completion of the course. The SRS dataset includes about 500 student comments and ratings for lecture and lab sessions after midterm and final semester examinations for a course taught by one teacher over the past 10 years. It also includes student surveys and comments for 25 courses taught by different teachers at the University over the past 2 years, which we used in conjunction with direct assessments of student performance to evaluate the system's reliability. The summary of the dataset is given in Table 5.1.

Table 5.1: Summary of students' feedback dataset

Platform	**Comments**	**Remarks**
Coursera	4000	During course
	1700	After completion of course
University SRS	500	For lecture and lab sessions

The sample of students' feedback from Coursera along with ratings is shown in Table 5.2.

Table 5.2: Sample of students' feedback from Coursera

Sr. No.	Comment	Rating
1.	I love the professor! He is so funny!	★★★★★
2.	Too slow paced	★★
3.	Great teaching, easy language, some extra work and thinking	★★★★★
4.	Very good and easy learning techniques	★★★★
5.	I think this is great course, the professor, videos and all course content is very good!	★★★★★
6.	Horrible	★
7.	It was extremely hard and the activities was not explained properly	★★★
8.	Time wasting & rubbish	★
9.	Nice course helpful for beginners	★★★★
10.	Too easy and full of straw	★★
11.	Well-paced, well explained. The only bad side I found, was the lack of a solution to the exercises.	★★★★★
12.	Too easy for cs major	★★
13.	It's perfect.	★★★★★
14.	Great course Great Instructor! I really enjoyed the course.	★★★
15.	Nice teacher- frustrating assignments with little guidance…	★★★
16.	it is a great course for beginners	★★★★★
17.	Horrible experience.	★
18.	The course is very helpful and easy to understand.	★★★★
19.	very interesting course	★★★★★
20.	Very detailed course. Nice contents.	★★★★★
21.	it worth every penny.	★★★★★
22.	very simple	★★
23.	This course is awesome, I really enjoyed this course.	★★★★★
24.	Feels like lectures are being given by an AI instead of an actual teacher.	★★★★
25.	Overall the course is satisfactory although some areas still need work.	★★★

The sample of students' feedback from University SRS is shown in Table 5.3.

Table 5.3: Sample of students' feedback from University SRS

Sr. No.	Comment
1.	Wonderful teacher
2.	Lectures sometimes became monotonous and resulted in boring lectures partially due to use of same pitch by sir.
3.	U are so sincere and dedicated…it is amazing
4.	Good presentation word allotted for extra knowledge
5.	The teacher is good
6.	It was wonderful experience with you. You are wonderful teacher. thank u!!!
7.	Tutorials were very good
8.	Sir is great!!!!!!
9.	Awsum teacher…….
10.	A true professional
11.	The course content is best…..he teaches very well
12.	Excellent instructor
13.	No suggestion. He is the ideal teacher.
14.	The course is too much bulky as if two semesters' course is bundled into one. It should be somewhat lessened.
15.	You are my inspiration and motivation sir. I feel lucky to have you as my teacher…
16.	Good presentation word alloted for extra knowledge
17.	Very good explanation of fundamentals before moving on advance concepts
18.	everything is good
19.	he is simply excellent teacher
20.	he's a gem...
21.	Ideal Teacher
22.	awesome teaching....
23.	Boring lectures
24.	he is a marvelous teacher
25.	should lessen the bourdon

5.5.2 Data Pre-processing

During this phase, the SA system prepares collected data for further processing. This involves six steps.

a) **Transliteration:** To address the issue of use of mixed language in student comments, the text is transliterated using the Google Transliterate API.
b) **Lowercasing**: Characters are converted to lower case to ease the process of matching words in student comments to words in the NRC Emotion Lexicon. Consider the words “Something” and “something”. For humans, these words have the same meaning. The only difference between them is that the first word is capitalized, because it may be the word in the sentence. A SA system considers these words as different words because of change in their case. Therefore, after transliteration, first step of pre-processing is to make all words lowercase words. This step is performed using the ***tm_map*** function in R’s *tm* package.
c) **Tokenization:** Students’ comments are split into words, or tokens, using the tokenize function in R.
d) **Normalization**: Abbreviated content is normalized by using a dictionary to map the content to frequently used slang words. For example, “gud” and “awsm” are mapped to “good” and “awesome”, respectively.
e) **Stemming:** To further facilitate word matching, words in student comments are converted to their root word using the ***tm_map*** function in R’s *SnowballC* package. For example, “moving,” “moved,” and “moves” are all converted to “move.”
f) **Removal of irrelevant content:** Punctuation and stop words, which are irrelevant for SA, are removed to improve system response time and effectiveness.

For example, consider the one of the student feedback comment given as follows in (5.7).

Student Comment: Sir u r great..!! (5.7)

After Tokenization: Sir, u, r, great, ., !!

After normalization and down casing: sir, you, are, great, ., !

After stemming: sir, you, are, great, .,!!

After removal of irrelevant content: sir, great

After pre-processing of student comment given in (5.7), sentiment and emotion will be identified for the remaining words “sir” and “great”.

5.5.3 Sentiment and Emotion Identification

During this phase, the SA system analyzes the preprocessed data to identify instances of sentiment and emotion. It uses the NRC Emotion Lexicon, also known as EmoLex, to associate words with positive or negative sentiment and the eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). The lexicon supports 40 languages including several Indian languages like Hindi, Tamil, Gujarati, Marathi, and Urdu. It includes annotations for 14,182 unigram words for English and 8,116 for Hindi.

Each word in the lexicon has an emotion vector ($\vec{E}$) containing a Boolean value (*b*) for each sentiment (*s*) and emotion (*e*) as given in equation (5.8).

$$\vec{E} = \vec{E}_e + \vec{E}_s$$
$$where\ \vec{E}_e \in \{b_0, ..., b_7\}\ and\ \vec{E}_s \in \{b_8, b_9\}, \forall b_i \in \{0,1\} \quad (5.8)$$

In equation (5.8), $\vec{E}_e$ represents the boolean values of eight emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise and trust. $\vec{E}_s$ represents the Boolean values of sentiments negative and positive. If a word in a student comment matches a word in the lexicon, the corresponding emotion vector is returned.

Student Comment: “Sir, you are great!” (5.9)

For example, for the student comment given in (5.9), the SA system would return the following emotion vector as given in Table 5.4 as the word “great” in the comment matches with word in the lexicon. This equates to the positive sentiment, as *trust* and *positive* parameters have a *b* value equal to 1.

Table 5.4: Emotion vector for sentence (5.9)

anger	anticipation	disgust	Fear	joy	sadness	surprise	Trust	negative	positive
0	0	0	0	0	0	0	1	0	1

If more than one word in the comment matches with the words in the lexicon, the sum of the corresponding emotion vectors is returned. For example, consider the student comment given in (5.10).

Student Comment: "He is good and wonderful teacher." (5.10)

In comment (5.10), two words "good" and "wonderful" match with the words in the lexicon and the summation of the corresponding emotion vector of "good" and "wonderful" words is returned.

For example, the emotion vector returned for word "good is given in Table 5.5.

Table 5.5: Emotion vector for word "good"

anger	anticipation	Disgust	fear	Joy	sadness	surprise	Trust	negative	positive
0	1	0	0	1	0	1	1	0	1

Similarly, the emotion vector returned for word "wonderful" is given in Table 5.6.

Table 5.6: Emotion vector for word "wonderful"

anger	anticipation	Disgust	fear	Joy	sadness	surprise	Trust	negative	positive
0	0	0	0	1	0	1	1	0	1

Now, the emotion vector returned for the complete sentence given in (5.10) is the summation of the emotion vectors returned for words "good" and "wonderful" given in Table 5.7.

Table 5.7: Emotion vector for sentence (5.10)

anger	anticipation	Disgust	fear	joy	sadness	surprise	Trust	negative	positive
0	1	0	0	2	0	2	2	0	2

In this way, an emotion vector is created for each comment representing the different emotions and sentiments contained within it.

To enable temporal analysis of sentiments and emotions, the SA system generates a mean emotion vector ($\vec{E}_j$) for each month and year as given in following equation (5.11).

$$\vec{E}_j = \frac{1}{n}\sum_{i=0}^{n-1}\sum_{j=0}^{p-1}\vec{E}_{ji}, \forall \vec{E}_{ji} \in \square \ where \square \geq 0; \tag{5.11}$$

Here, *n* represents the number of comments in each month/year and *p* represents the emotion and sentiment parameters such that $p \in$ *{anger, anticipation, disgust, fear, joy, sadness, surprise, trust, negative, positive}.* This vector is created to avoid the anomalies that might result from increase in the value of particular emotion in that month/year.

For example, the sum of emotion vectors of all the comments in a year is given in the Table 5.8. In this table, the summation of emotion vector of all the comments in each year is represented. To compute the mean emotion vector, for each year, the summation of all the emotions is performed in each particular year which is represented in sum of emotions column. From the Table (5.8), it is clearly visible that the weightage of trust emotion is more as comparison to other emotion.

Table 5.8: Emotion vectors of students' comments for five years (2011-2015)

year	anger	anticipation	disgust	fear	joy	sadness	Surprise	trust	sum of emotions
2011	2	9	1	2	10	1	4	19	48
2012	2	6	2	3	7	3	7	24	14
2013	2	5	3	2	9	2	4	26	53
2014	1	13	0	0	19	2	7	46	88
2015	2	12	0	3	12	5	6	46	86

Therefore, to reduce the weightage of a particular emotion, normalization is performed by calculating the mean value for each emotion in the range between 0-1 as shown in Table 5.9.

Table 5.9: Mean values of emotion vectors for five years (2011-2015)

Year	anger	anticipation	disgust	fear	joy	Sadness	surprise	trust
2011	0.042	0.188	0.021	0.042	0.208	0.021	0.083	0.396
2012	0.037	0.111	0.037	0.056	0.129	0.056	0.129	0.444
2013	0.038	0.094	0.057	0.038	0.169	0.038	0.075	0.491
2014	0.011	0.148	0	0	0.216	0.023	0.079	0.523
2015	0.023	0.139	0	0.035	0.139	0.058	0.069	0.535

5.5.4 Satisfaction and Dissatisfaction Computation

Satisfaction and dissatisfaction are crucial parameters in education. The following six emotion parameter namely, joy, trust, anticipation, anger, disgust, and sadness play an important role in computing satisfaction and dissatisfaction. The parameters anticipation and trust clearly connote satisfaction, but in some circumstances joy could have a negative connation—for example, a student could feel joy at skipping a boring class. Therefore, in computing student satisfaction, the sum of anticipation and trust is multiplied by a constant ($\alpha = 0.6$) to give these parameters more weight. The same mechanism is employed in computing student dissatisfaction to give more weight to anger and disgust than to sadness. The calculations are as follows as given in equations (5.12) and (5.13) respectively.

$$satisfaction = [\alpha(TA) + (1-\alpha)(J)]/n \tag{5.12}$$

$$dissatisfaction = [\alpha(AD) + (1-\alpha)(S)]/n \tag{5.13}$$

where $TA = trust + anticipation$, $J = joy$, $AD = anger + disgust$, $S = sadness$ and n = *max(TA or AD, J or S).*

In equation (5.12) and (5.13), the variable *n* selects the maximum value of *TA* or *J* in case of satisfaction and maximum value of *AD* or *S* in case of computing the parameter dissatisfaction to normalize the value in the range 0 to 1 for both satisfaction and dissatisfaction parameters.

For example, consider the comment given in (5.14).

Student Comment: "He is good at teaching." (5.14)

The SA system returns the following emotion vector as given in Table 5.10 from the NRC lexicon for the word "good".

Table 5.10: Emotion vector for word “good”

anger	anticipation	disgust	fear	joy	sadness	surprise	trust	Negative	Positive
0	1	0	0	1	0	1	1	0	1

Here, $TA = 2$, $J = 1$, $n = max(TA, J) = max(2, 1) = 2$ and $\alpha = 0.6$

Satisfaction is thus calculated as using equation (5.12) for comment (5.14).

$$Satisfaction = [0.6(2) + (1-0.6)\ (1)]/2$$

$$= [0.6(2) + 0.4\ (1)]/2$$

$$= 1.6/2$$

$$= 0.8$$

Thus, satisfaction computed for comment (5.14) is 0.8 which shows the satisfaction of student about the performance of teacher.

Let us consider another comment given in (5.15) to compute the dissatisfaction parameter.

Student Comment: “He is bad at teaching and every student has doubts about the class.” (5.15)

The system returns the following emotion vector given in Table 5.11 which is summation of emotion vectors for the words “bad” and “doubts” from sentence (5.15).

Table 5.11: Emotion vector for sentence (5.15)

anger	anticipation	disgust	fear	joy	sadness	surprise	Trust	negative	Positive
1	0	1	2	0	2	0	1	2	0

In this case, $AD = 2$, $S = 2$, $n = max(AD, S) = \max(2, 2) = 2$ and $\alpha = 0.6$

Dissatisfaction is therefore calculated as follows using the equation (5.13) for the sentence (5.15).

$$Dissatisfaction = [0.6(2) + (1-0.6)(2)]/2$$

$$= [0.6(2) + 0.4(2)]/2$$

$$= 2.0/2$$

$$= 1$$

Thus, dissatisfaction value for sentence (5.15) is 1 which is maximum value and shows the dissatisfaction of students about the performance of teacher.

Table 5.12 presents the some examples of Students' feedback taken from University SRS and online course dataset with different emotion and sentiment values

Table 5.12: Examples of students' feedback with emotion and sentiment values

Sr. No.	Students' Feedback Comment	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Negative	Positive	Satisfaction	Dissatisfaction
1.	Wonderful teacher	0	0	0	0	0	0	0	1	0	1	0.6	0.4

Sr. No.	Students' Feedback Comment	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Negative	Positive	Satisfaction	Dissatisfaction
2.	Time wasting & rubbish	0	1	2	1	0	1	0	0	2	0	0.2	0.8
3.	Very good and easy learning techniques	0	1	0	0	1	0	1	1	0	2	0.8	0.2
4.	Great course Great Instructor! I really enjoyed the course.	0	1	0	0	0	0	0	1	0	1	0.6	0.4
5.	Well-paced, well explained. The only bad side I found, was the lack of a solution to the exercises.	1	0	1	1	1	1	0	1	2	2	0.5	0.5
6.	Excellent instructor	0	1	0	0	1	0	0	2	0	2	0.7	0.3
7.	Tutorials were very good.	0	1	0	0	1	0	0	2	0	2	0.8	0.2
9.	Horrible	1	0	1	1	0	0	0	0	1	0	0.4	0.6
9.	You are my inspiration and motivation sir. I feel lucky to have you as my teacher.	0	1	0	0	2	0	1	2	0	4	0.8	0.2

Sr. No.	Students' Feedback Comment	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Negative	Positive	Satisfaction	Dissatisfaction
10.	Nice course helpful for beginners	0	0	0	0	1	0	0	1	0	1	1	0

5.5.5 Data Visualization

To facilitate analysis of student feedback about course satisfaction and teacher performance, the proposed SA system has a data-visualization component. This component helps in analyzing students' sentiments and emotions by creating sentiment and emotion word clouds of positive and negative words. Also, temporal sentiment and emotion analysis of students' feedback of University SRS and online course over the time range of 10 years has been presented in this section with the help of line graphs.

a) Sentiment and Emotion Word Clouds

Students use a variety of words to convey their sentiments or emotions while giving feedback.

Figure 5.3 Sentiment cloud of positive words

Visualizing frequently used positive words (“great,” “excellent,” “interesting,” and so on) as shown in Figure 5.3 and negative words (“dull,” “confusing,” “terrible,” and so on) as shown in Figure 5.4 in the form of word clouds can help identify student learning behavior—for example, whether or not they are taking an interest in lectures and lab sessions.

Figure 5.4 Sentiment cloud of negative words

b) Temporal Sentiment and Emotion Analysis

As indicated earlier, the proposed SA system groups together positive and negative comments and ratings in student feedback by month and year. This makes it possible to track teacher performance and course satisfaction over time.

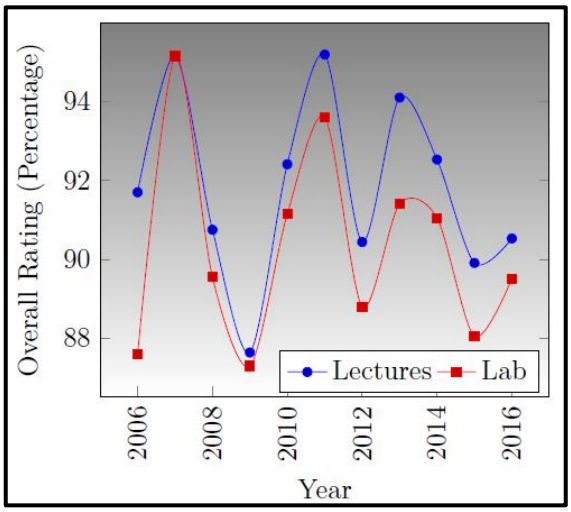


Figure 5.5 Temporal sentiment analysis of students’ ratings in lectures and labs

Figure 5.5 plots overall student ratings (ranging from 0 to 100 percent) of one teacher's performance in lectures and lab sessions of a University course from 2006 to 2016; the graph shows that students rated the teacher's performance in lectures slightly higher than that in lab sessions and that the average overall rating was more than 90 percent during the last six years.

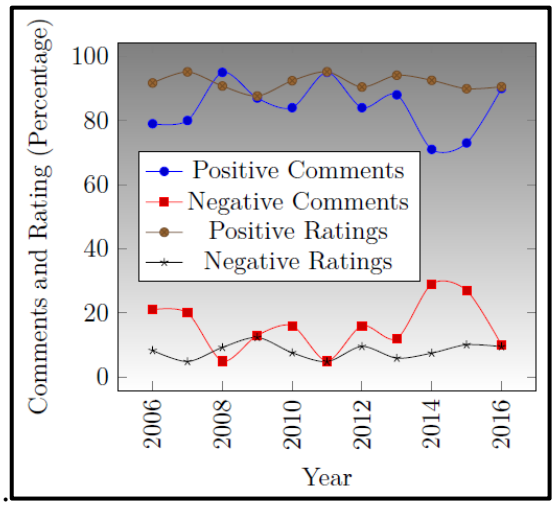


Figure 5.6 Temporal SA of percentage of positive and negative student comments and ratings

Figure 5.6 plots the percentage of positive and negative student comments about and ratings of the teacher over the same period; the graph reveals that, on average, 85 percent of comments were positive and 15 percent were negative. Sentiment polarity can also be tracked across different teachers and courses over time to analyze overall teaching quality at a given institution. The proposed SA system also groups together emotions identified in comments about courses and teachers by month and year, providing more granular insight.

Figure 5.7 plots the percentage of emotions extracted from student feedback on a one year Coursera course by month; the graph shows that students expressed the positive emotions of trust, joy, anticipation, and surprise more than the negative emotions of sadness, fear, disgust, and anger.

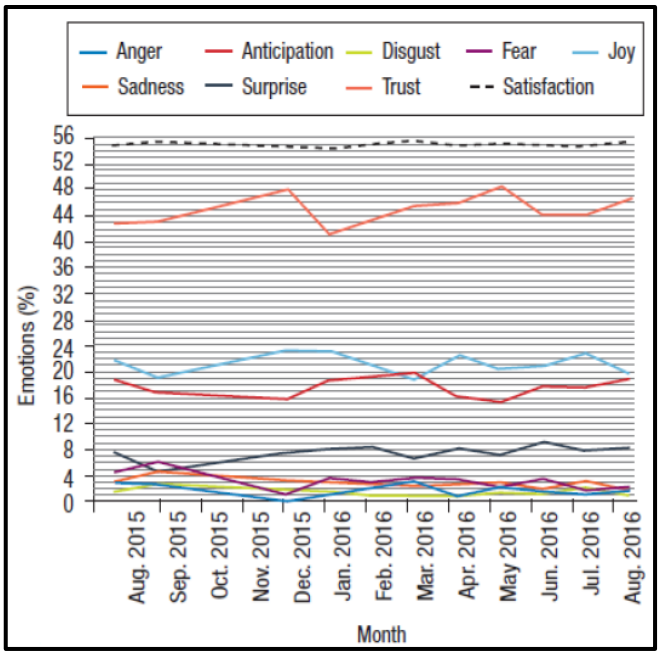


Figure 5.7 Temporal emotion analysis of feedback on one year Coursera course

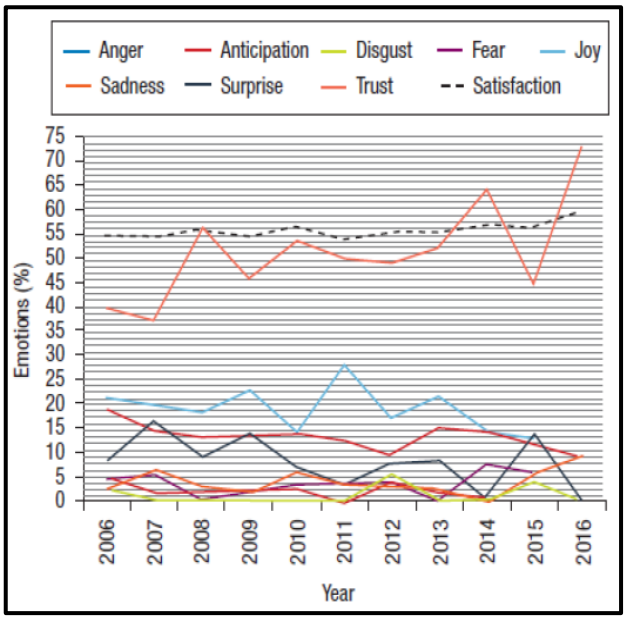


Figure 5.8 Temporal emotion analysis of feedback of comments about a teacher

Figure 5.8 plots the percentage of emotions extracted from student comments about the teacher; it shows that students' trust in the instructor gradually increased over the decade and that each year the percentage of positive emotions exceeded that of negative emotions.

5.6 System Evaluation

The results of the proposed system have been analyzed using standard evaluation methods of Course Outcomes (COs). COs are the characteristics that the students are expected to prove after finishing the course. The evaluation of COs is very important to test whether the student or learner has achieved what is expected out of them. The evaluation results are used for continuous quality improvement. There are two methods for assessment and evaluation of COs, i.e., direct and indirect. In case of direct assessment method, the evaluation of the achievement of course outcomes are carried out using the data from continuous tests, mid/end semester exams, assignments, laboratory practical and project reports to compute the student class performance. While in case of indirect assessment method, the evaluation of the student course survey, SRS feedback in the form of ratings and comments about a teacher and course is carried out to provide information about student's perception of their learning. The direct assessment methods are strong evidences of student learning as teacher continuously evaluates the students' performance in a particular course which decides the actual learning of students in terms of grade or CGPA attained by students in that course. Indirect assessment measures the students' implicit qualities such as values, attitudes and sentiments about the teacher and the course which help in measuring students' perception about the course. Both of these assessment methods provide converging evidences of student learning.

In education, there is a general consensus that direct and indirect assessments of teaching quality and learning behavior should agree. Students who perform well in a course, for example, would be expected to give the teacher high ratings and favorable comments; conversely, those who perform poorly are likely to be dissatisfied.

To validate the proposed SA system, student surveys and comments obtained from a University SRS system were analyzed for 25 different courses over a two-year period and compared the percentage of positive sentiments students had about each course with the average course grade on a 0–100 scale.

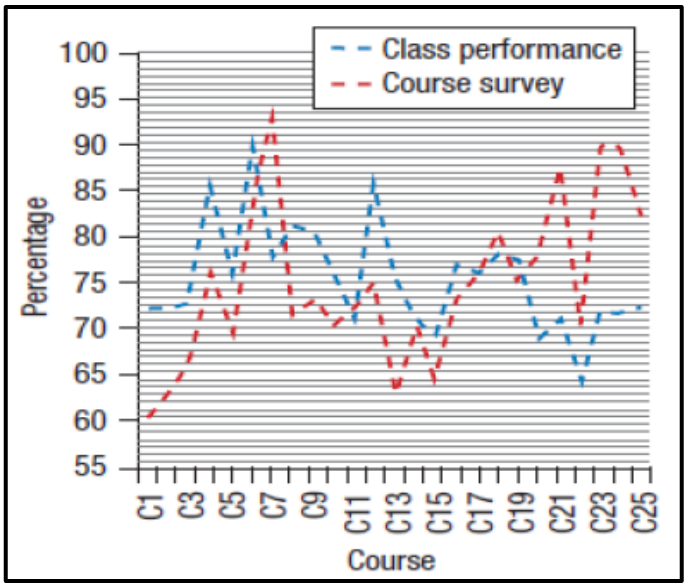


Figure 5.9 Comparison of student performance with surveys

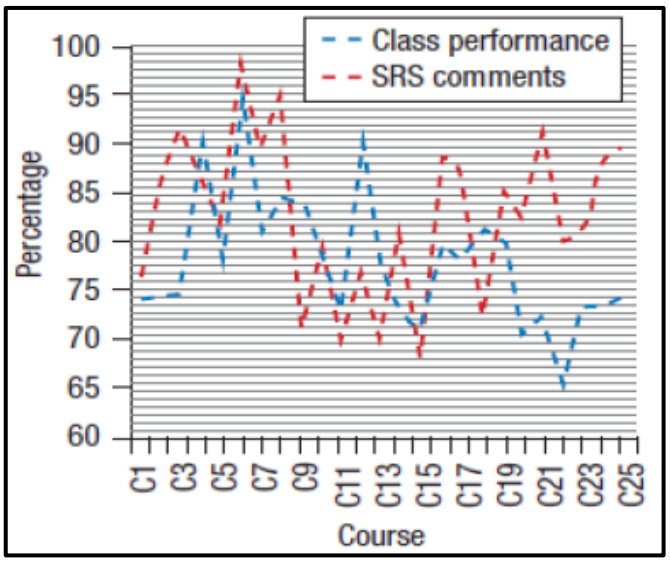


Figure 5.10 Comparison of student performance with comments

As Figure 5.9 and Figure 5.10 show that the results generally agreed, with less than 20 percent absolute difference between the methods. In those courses where student performance exceeded satisfaction, there could be a number of explanations such that the exams were relatively easy, the course had a particularly bright or hard-working group of students, or students did not like the teacher for personal reasons or felt they did not gain much value from the class. In those courses where student satisfaction exceeded

performance, perhaps the exams were exceptionally challenging or students failed to adequately prepare. In either case, the discrepancy in results obtained from both approaches invites continued analysis.

5.7 Limitations of the System

Despite its promise, the system has some limitations. It is only as good as the data it analyzes, so care must be taken in collecting feedback from students. SRSs must be well designed to ensure that they are engaging, and instructors must make a concerted effort to ensure that as many students as possible provide complete and accurate feedback. The accuracy of the system depends on the lexicon coverage of NRC emotion lexicon. Also, the system is not able to handle negations, intensifiers and sarcasm detection in the text.

Chapter Summary

In this chapter, a sentiment analysis system for education domain has been presented. The proposed sentiment analysis system analyzes students' feedback collected from Coursera and SRS of the University using "R" language with natural processing techniques. The sentiments of students have been analyzed in the form of different emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, trust as well as positive and negative sentiments. This chapter also presents the derivation of two new emotions satisfaction and dissatisfaction from the existing emotions along with example. The proposed system has been tested using direct and indirect assessment methods of course evaluation and it has been analyzed that both the methods provide converging evidence of student learning and teaching quality. Thus, the proposed sentiment analysis system can help an organization in improving student learning and teaching quality.

Chapter 6

Web-based Sentiment Analysis System for Hindi

6.1 Introduction

In this chapter, a web-based sentiment analysis system developed for the Hindi language has been presented. This system performs sentiment analysis at three levels, i.e., aspect, sentence, and document level and classifies the text into three classes, i.e., positive, negative and neutral. This web-based system also performs sentiment analysis of Twitter posts about any user-defined hashtag. The next sections of this chapter present the brief description about the tools and technologies used for the development of this web-based sentiment analysis system along with the snapshots of working of this system at different levels with example sentences.

6.2 Tools and Technologies Used

This web-based Hindi system has been implemented in Python language and Flask has been used for its Web Interface. A prototype of this system is available on the public repository at Github and its web application is made online at URL http://www.hindisenti.com/ to all users by deploying it to Heroku which supports both Python and Java languages. The Scikit-learn and Keras library of Python with backend TensorFlow has been used for training of machine learning and deep learning models respectively. The basic description of the libraries used for the development of the web-based sentiment analysis system for Hindi language is given in Table 6.1.

Table 6.1: Description of libraries of Python

Sr. No.	**Library**	**Version**	**Description**
1.	Gunicorn	==19.7.1	Gunicorn is a Python Web Server Gateway Interface (WSGI) HTTP server implementation that is commonly used to run Python web applications. It was developed by Benoit Chesneau in 2010.

Sr. No.	Library	Version	Description
2.	TensorFlow	==1.1.0	TensorFlow is a free and open-source software library for dataflow. It is a symbolic math library and is used for machine learning applications such as neural networks. It was developed by the Google Brain team for internal Google use and released under the Apache License 2.0 on November 9, 2015.
3.	Flask	>=0.12.3	Flask is a web application framework written in Python and it was created by Armin Ronacher of Pocoo, an international group of Python enthusiasts.
4.	h5py	==2.7.0	The h5py package of Python provides both a high- and low-level interface to the HDF5 library. The low-level interface is intended to be a complete wrapping of the HDF5 API, while the high-level component supports access to HDF5 files, datasets, and groups using established Python and NumPy concepts.
5.	Keras	==2.0.5	Keras is an open-source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System).
6.	Scikit-learn	==0.18.2	Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Sr. No.	Library	Version	Description
7.	NumPy	==1.12.1	NumPy is a Python library to add support for large, multi-dimensional arrays and matrices, along with a high-level mathematical functions to operate on these arrays. It was created by Travis Oliphant in 2005 by incorporating features of the competing Numarray into Numeric, with extensive modifications.
8.	NetworkX	==1.11	NetworkX is a Python package to study, create and manipulate the functions, synamics and structure of complex networks.
9.	Tweepy	==3.5.0	Tweepy is a Python library. It enables Python to communicate with the Twitter platform and uses its API.

6.3 Features of the System

Some of the important features provided by the web-based Hindi SA system are given as follows.

a) The developed web-based system performs SA for resource-poor Hindi language. It is available online at http://www.hindisenti.com/.
b) The system is supported by Quillpad plugin which converts the English input into Hindi script so that users can easily type the Hindi sentence.
c) The system is also able to abbreviate the short forms used in Hindi with the help of mapping dictionary.
d) The system performs SA at three sentiment levels such as aspect, sentence, and document.
e) The system classifies the sentences into three sentiment classes, i.e., positive, negative and neutral.
f) The system generates the dependency graph of the sentence while performing SA at aspect-level and also highlights the aspects, sentiment-bearing words, intensifiers, and negations contained in a sentence through GUI.

g) The system is also trained on different machine and deep learning algorithms to perform SA. The sentiment polarity is represented through bar-chart.
h) The interface of the system also shows the sentiment polarity of a sentence or document predicted by different ML and deep algorithms.
i) The system performs SA of real-time tweets about any user-defined trending Hashtag.

Thus, the developed web-based Hindi sentiment analysis system can be used for other NLP applications and has a great impact on the benefits of society.

The next sections of this chapter present the brief description about the interface and working of the web-based sentiment analysis system for Hindi langauge.

6.4 Home Page

Figure 6.1 shows the front page of web-based sentiment analysis system for Hindi language.

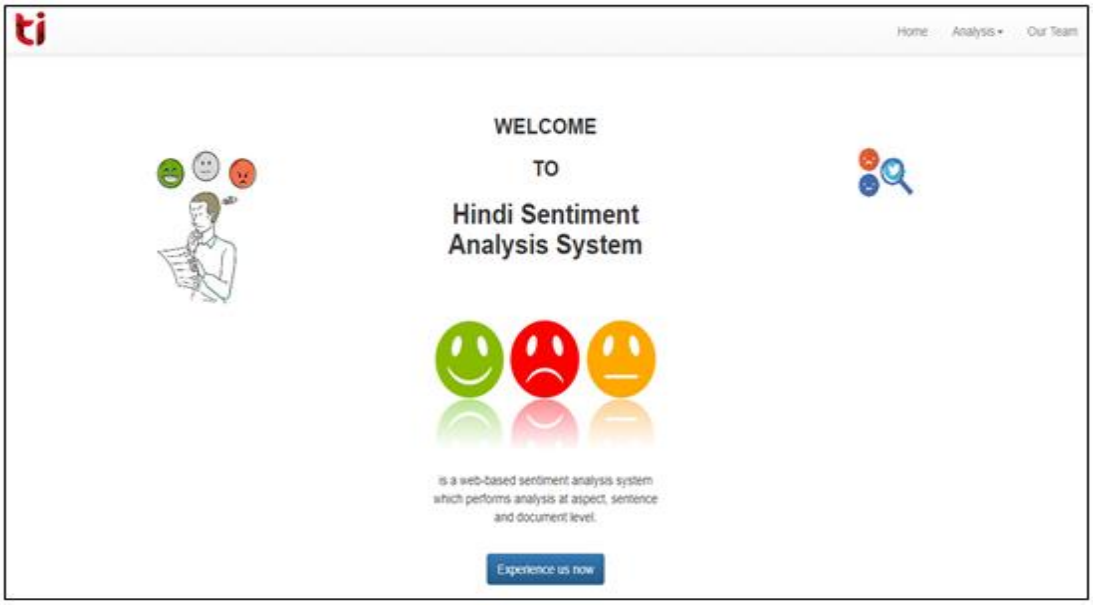


Figure 6.1 Home page of web-based sentiment analysis system

The interface of the system consists of three tabs, i.e., “Home”, “Analysis” and “Our Team”. There is a button “Experience us now” which redirects the user to an interface which performs the aspect-based sentiment analysis. The “Home” tab redirects the user to home page of the system. The “Analysis” tab consists of drop-down menu which consists

of the links to other four interfaces, i.e., “Sentence Analysis”, “Document Analysis”, “Tweets Analysis” and “Aspect Analysis” as shown in Figure 6.2.

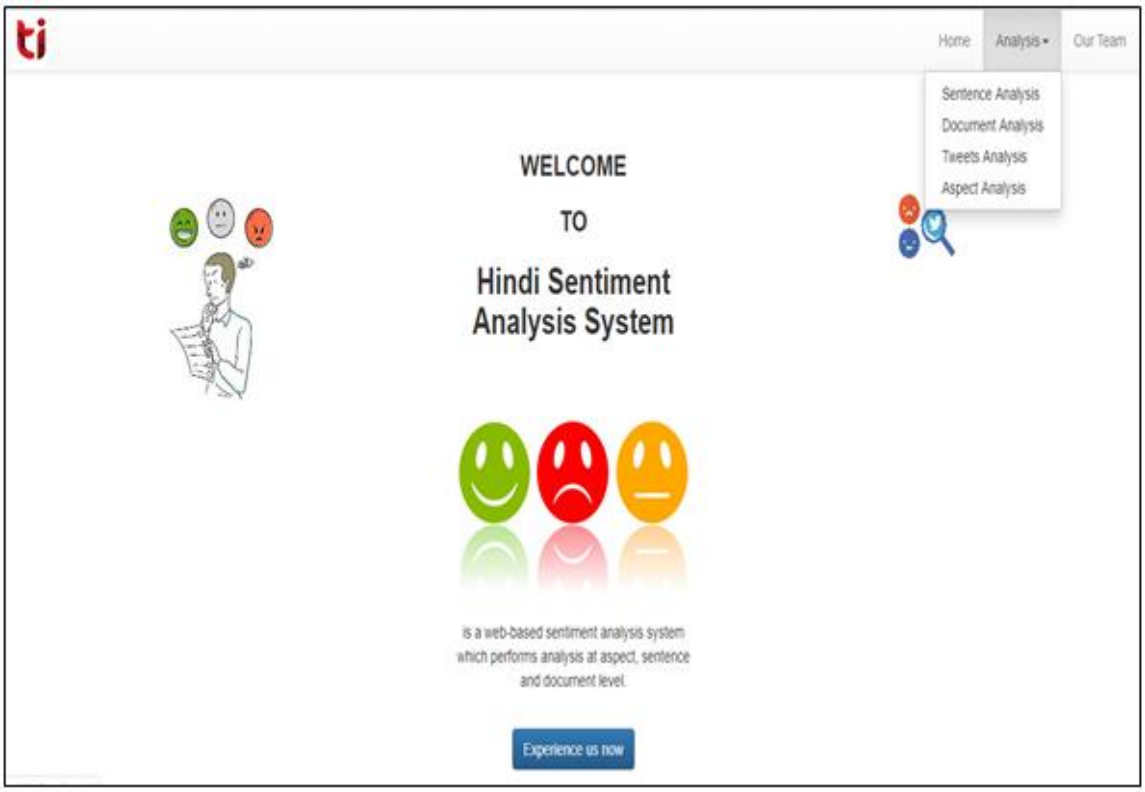


Figure 6.2 Analysis drop-down menu

The tab “Our Team” presents the details about the team who have worked for the development of this web-based sentiment analysis system.

6.5 Sentence-Based Sentiment Analysis

The interface of sentence-based sentiment analysis consists of text-box for inputting a Hindi sentence. It consists of a button “Predict”, on clicking this button; the system presents the predicted output in the form of bar chart and also displays the sentiment polarity of the sentence. As shown in Figure 6.3, each interface of the system also consists of other three tabs on the right-hand side of the page from which user can easily go to any other interfaces for performing sentiment analysis at other levels. The text-box of the developed web-based sentiment analysis system is integrated with Quillpad plugin so that it can transliterate the English text into Hindi. The user can either copy-paste the

sentence into the text box from any source for which sentiment analysis is to be performed or can type in English for its transliteration to Hindi text.

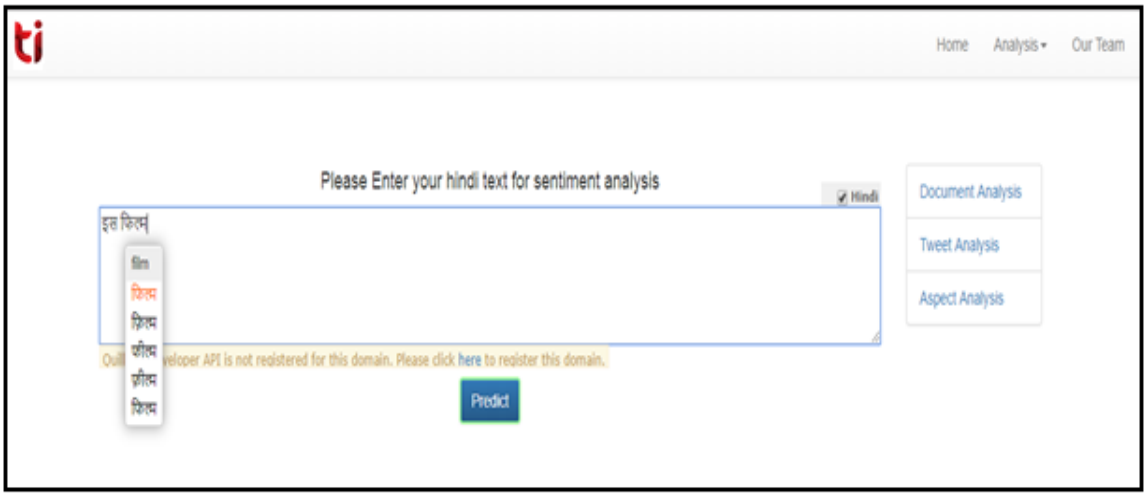


Figure 6.3 Transliteration of English text into Hindi text

For example, as the user types the English word "film" in the text box then it will automatically get transliterated into Hindi word फ़िल्म *philm "film"* as shown in Figure 6.3.

The developed web-based sentiment analysis system also supports mapping of abbreviations to its full form. For example, consider the Hindi sentence given in (6.1) and its transliteration and English translation are given in (6.2) and (6.3) respectively.

Hindi Sentence: यूपी में हिंदू मुस्लिम दंगे हो गये।. (6.1)

Transliteration: *yoopee mein hindoo muslim dange ho gae.* (6.2)

English Translation: "*Hindu Muslim riots broke out in UP.*" (6.3)

In sentence (6.1), the word यूपी *yoopee* "UP" will be mapped to word उत्तर प्रदेश *Uttar Pradesh* "Uttar Pradesh" (abbreviated form of UP) by the system before performing sentiment analysis. The sentence (6.1) after mapping of abbreviations is shown in the text area "Expanded Abbreviations" as shown in Figure 6.4. The system predicts the sentiment polarity depending upon the majority of vote of different machine learning and

deep learning algorithms. For the sentence (6.1), the system predicts the negative sentiment polarity.

Figure 6.4 Mapping of abbreviations

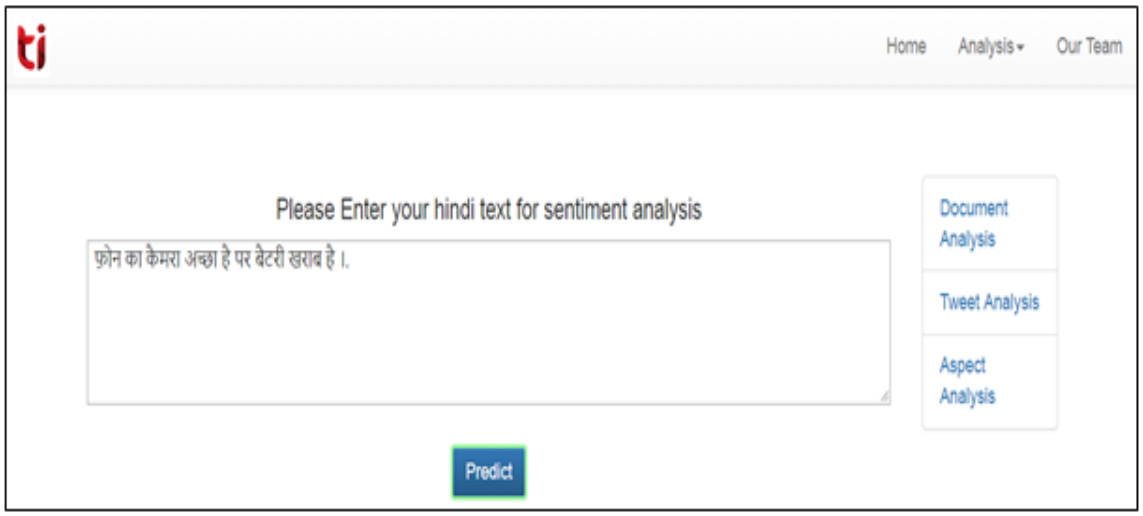


Figure 6.5 Input interface for sentence-based sentiment analysis

Let us take an another example to analyze the sentence-based sentiment analysis as shown in Figure 6.5. Consider the inputted Hindi sentence given in (6.4) and its transliteration and English translation are given in (6.5) and (6.6) respectively.

Hindi Sentence: फ़ोन का कैमरा अच्छा है पर बैटरी खराब है।. (6.4)

Transliteration: *fon ka kaimara achchha hai par baitaree kharaab hai.* (6.5)

English Translation: "The camera of phone is good but battery is bad." (6.6)

The predicted output of input Hindi sentence (6.4) is shown in Figure 6.6. It shows that the overall sentiment of the sentence is negative and it is also depicted in the form of bar chart.

Figure 6.6 Predicted output for sentence-based sentiment analysis

6.6 Document-Based Sentiment Analysis

In document-based analysis, a document is uploaded for which the sentiment analysis is to be performed. Figure 6.7 presents the input interface for document-based analysis. It consists of "Choose file" button which helps in browsing the file from the system as shown in Figure 6.8.

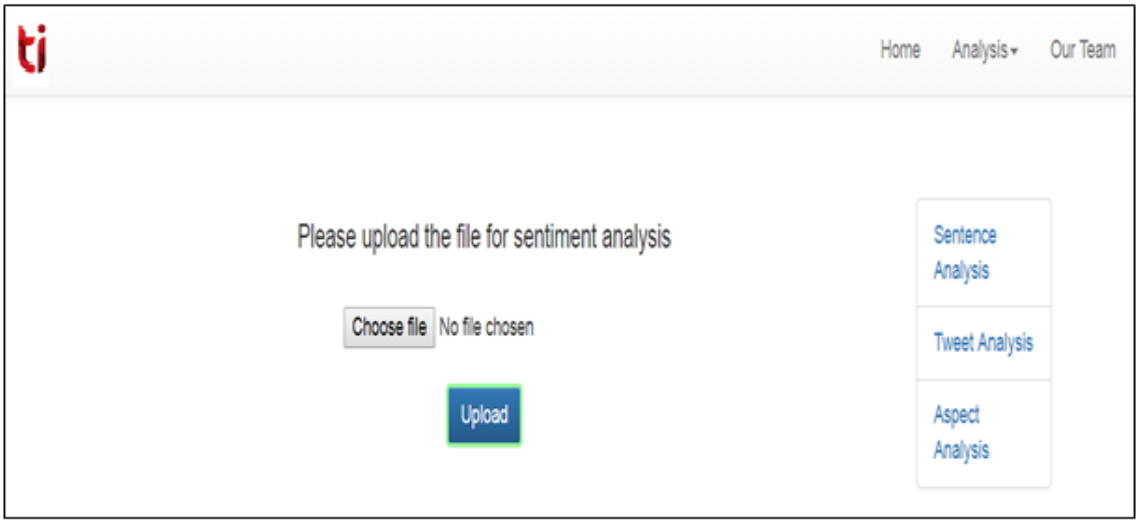


Figure 6.7 Input interface for document-based sentiment analysis

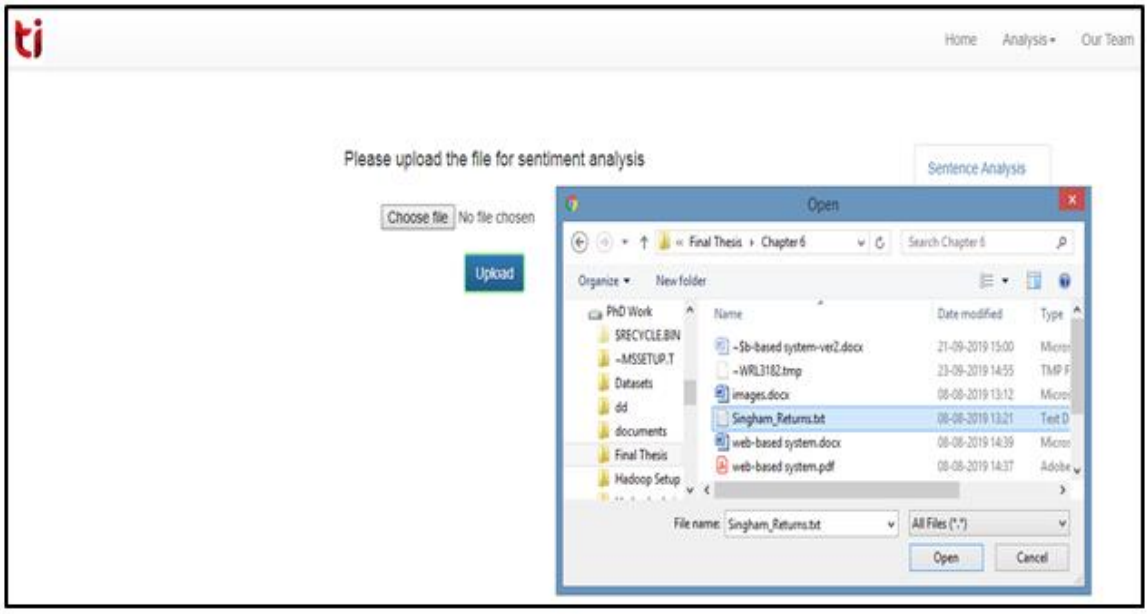


Figure 6.8 Browsing of review for document-based sentiment analysis

As shown in Figure 6.8, “Singham_Returns.txt” (a movie review about the movie “Singham Returns”) is selected from the system to perform its sentiment analysis. On clicking “Upload” button, the system presents the predicted output in the form of overall sentiment polarity of the document as shown in Figure 6.9. As shown in Figure 6.9, the system shows the text of the uploaded document in text area “Expanded Abbreviations” after resolving all the abbreviations used in the document.

Figure 6.9 Predicted output for document-based sentiment analysis

The system predicts the output of the document using different machine learning and deep learning algorithms as shown in the left-hand side of Figure 6.9. The system also shows the predicted output with number of sentences with positive, negative and neutral polarity on bar chart. For the given document, the system predicts the 18 sentences as neutral, 8 sentences as negative and 2 sentences as positive out of total sentences. As the document consists of majority of neutral sentences therefore, the system predicts the overall neutral sentiment depending upon the majority of vote of all the machine learning and deep learning algorithms. However, in a review, a document consits of more number of neutral sentences as comparison to positive and negative sentences and also the number of negative sentences predicted by the system are more than positive sentences therfore, the given document “Singham_returns.txt” consists of negative polarity.

6.7 Aspect-Based Sentiment Analysis

Figure 6.10 presents the input interface for aspect-based sentiment analysis. For example, consider the inputted Hindi sentence given in (6.7) and its transliteration and English translation is given in (6.8) and (6.9) respectively.

Hindi Sentence: फ़ोन का कैमरा अच्छा नहीं है और बैटरी भी खराब है।. (6.7)

Transliteration: *fon ka kaimara achchha nahin hai aur baitaree bhee kharaab hai.* (6.8)

English Translation: "The camera of the phone is not good and battery is also bad." (6.9)

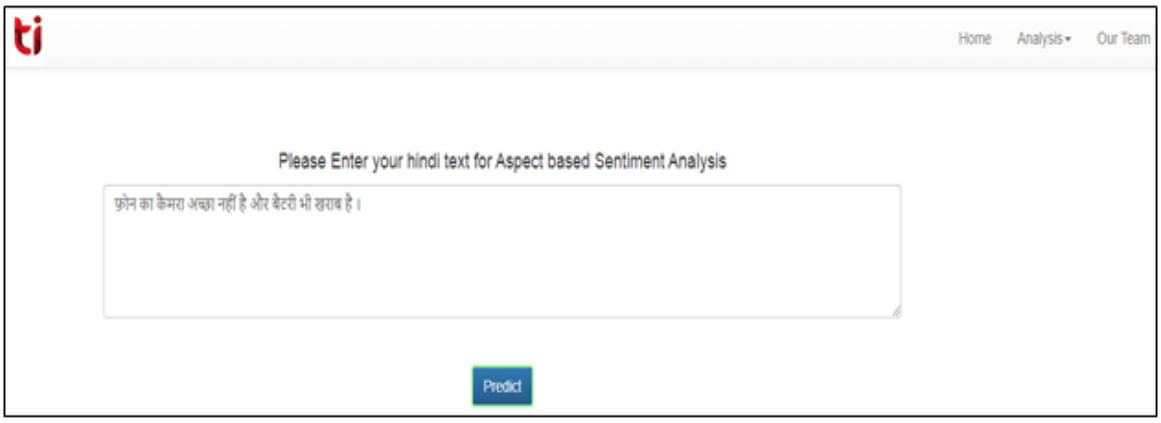


Figure 6.10 Input interface for aspect-based sentiment analysis

On clicking "Predict" button, the system presents the predicted output for aspect-based sentiment analysis as shown in Figure 6.11.

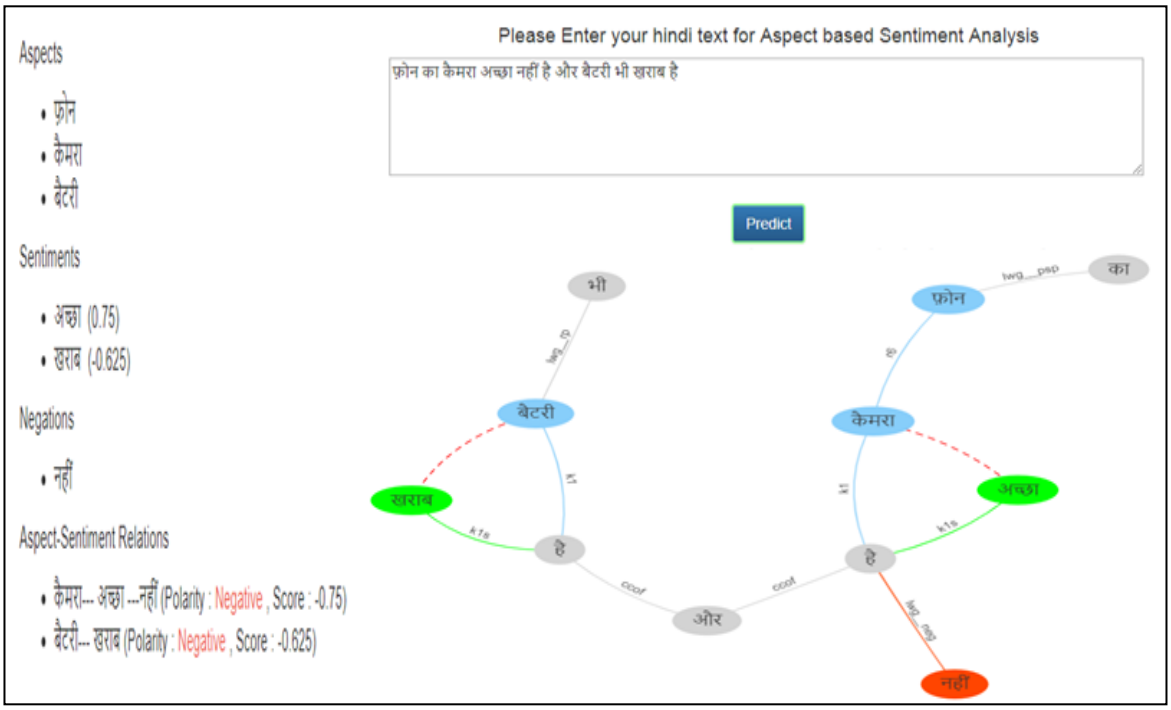


Figure 6.11 Predicted output for aspect-based sentiment analysis

The system is able to handle negations as well as intensifiers. The interface of the system highlights the aspect-terms, sentiment words, negations. and intensifiers identified from the sentence with different colors in the dependency graph. In dependency graph, the aspects are shown in “blue” color, sentiments are shown in “green” color and negations are shown with “red” color. The edges of the graph are represented with dependency relations between the nodes. The final sentiments which are assigned to aspects are shown with dotted edges in dependency graph and also the sentiment score is displayed on the interface.

For example, for the sentence (6.7), the system identifies the फ़ोन *fon* “phone”, कैमरा *kaimara* “camera” and बैटरी *baitaree* “battery” as aspects; अच्छा *achchha* “good” with polarity score 0.75 and खराब *kharaab* “bad” with polarity score -0.625 as sentiments and नहीं *nahin* “not” as negation. Then, the system assigns positive sentiment word अच्छा *achchha* “good” to aspect कैमरा *kaimara* “camera” depending upon the minimum distance between sentiment word and aspect word and also assigns the negation नहीं *nahin* “not” to aspect अच्छा *achchha* “good” which in return changes the polarity to negative and its score to -0.75. The negative sentiment word खराब *kharaab* “bad” is assigned to aspect word बैटरी *baitaree* “battery” by the system. Thus, the system assignes the negative polarity to both the aspects कैमरा *kaimara* “camera” and बैटरी *baitaree* “battery”.

6.8 Tweets Analysis

The developed web-based SA system is also able to perform the sentiment analysis of tweets. The system first extracts the tweets in Hindi about any user-defined trending Hashtag. Currently, the system extracts the recent hundred tweets. The system performs SA of the extracted tweets on the basis of the trained model over different machine learning and deep learning algorithms. All the trained models predict the sentiment class

of extracted tweets individually and final sentiment polarity is assigned based on the majority of the vote. The system represents the sentiment polarity of tweets through the bar-chart in all the three sentiment classes.

Figure 6.12 shows the input interface for tweets analysis. It consists of a text-box in which user can enter any user-defined Hashtag or current trending Hashtag on Twitter.

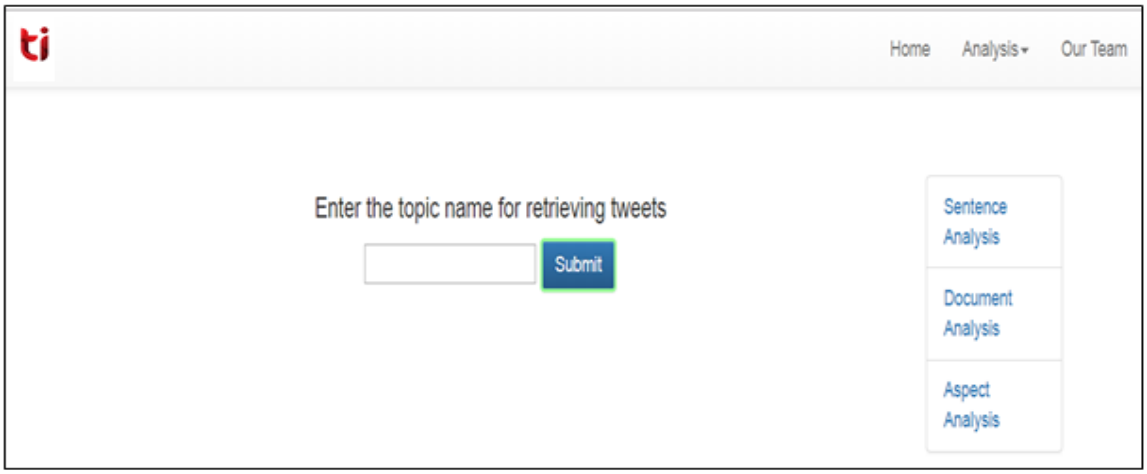


Figure 6.12 Input interface for tweets analysis

On clicking "Submit" button, the system presents the predicted output of extracted tweets as shown in Figure 6.13.

Figure 6.13 Predicted output of extracted tweets

For example, the predicted output for the Hashtag "#biharchilddeaths" is shown in Figure 6.13. The system has predicted the negative polarity for the given Hashtag depending upon the majority of vote of different machine learning and deep learning algorithms. It has been observed from the Figure 6.13 that most of the extracted tweets about the given Hashtag are identified with negative polarity and out of which, some of the tweets are of neutral polarity and none of the tweets are identified with positive polarity. Thus, the system assigns the negative polarity as overall sentiment about the given Hashtag.

The developed Hindi sentiment analysis system can benefit in many application areas of NLP. In future, this web-based Hindi sentiment analysis system can be integrated with other applications by developing its web API. Also, the performance of the system can be improved by extending the Hindi corpus. The lexicon coverage of sentiment words can be enhanced by using the wordnet based approach to improve the accuracy of the system at aspect level. The extension of mapping dictionary can also help in improving the performance of the system.

Chapter Summary

This chapter presents web-based sentiment analysis system for Hindi language. It discusses about the libraries and frameworks of Python language that have used for the development of this web-based system. This chapter also highlights the important features of web-based system. The working of the system is discussed in detail at each level, i.e., sentence, document and aspect along with examples. The capability of the system to perform the sentiment analysis of Twitter posts is also explained in detail in this chapter.

Chapter 7

Conclusions and Future Scope

7.1 Conclusions

In this thesis, a detailed description about sentiment analysis is provided along with different sentiment levels, i.e., aspect, sentence and document; and sentiment classes positive, negative and neutral. As a part of this thesis, the need of sentiment analysis for Hindi language is discussed in detail and various research gaps have been identified which motivated us to perform the sentiment analysis for Hindi language.

In this thesis, a comprehensive and detailed review has been performed on the works which carried the research for different Indian languages in the field of sentiment analysis in the last 10 years. This review was performed systematically by developing review methodology following an inclusion and exclusion criteriato consider the relevant studies. From this review, it has been observed that research work has been performed for 15 Indian languages till now. Various annotated datasets and pre-processing linguistic resources were identified for different Indian languages with their online availability. A detailed summary of research works in terms of approach used, domain, courpus size, tools/language used and evaluation measures used to perform sentiment analysis are also provided. The percentage of status of research works in different sentiment analysis techniques, classes, levels and domains is represented in the form of pie charts which can help the researchers to further carry out the research in their native language.

In this thesis, sentiment analysis is performed at all the three levels such as aspect, sentence and document level. For experimentation, the corpus of Hindi movie reviews and tweets has been extracted from online review websites and Twitter, respectively. The extracted corpus has been manually annotated by three native speakers of Hindi. The validation of the corpus has been performed using kappa statistic measure. For sentence and document level, different machine learning algorithms, i.e., NB, MNB, BNB, SVM, DT, k-NN, RF, LR, AdaBoost, Gradient Boosting, LSTM and CNN have been trained on the annotated corpus of reviews and tweets for Hindi language. NB performs better out of

traditional ML algorithms and it has been identified that CNN outperforms while comparing with traditional ML algorithms with an accuracy of 95%.

For aspect-based sentiment analysis, the system uses two lexical resources HDP and HSWN. It follows an efficient aspect extraction process to extract all the relevant aspects which include three steps, *i.e.,* extraction of frequent nouns, identification of relevant nouns and removal of irrelevant nouns. The sentiment nodes are extracted using HSWN. The system uses HDP to determine the association between the aspect nodes and sentiment nodes. Also, the system generates a dependency graph and assigns the sentiment to the particular aspect having the least distance between sentiment word and aspect word. It is observed that the system has achieved an accuracy of 83.2%. The precision, recall and F-measure of the system are 0.85, 0.83 and 0.84 respectively. The results of the proposed system are compared with its traditional lexicon based approaches as well as with existing works on the aspect-based SA.

In this thesis, a case study of sentiment analysis for education domain is also presented which performs the sentiment analysis of students feedback collected from online blogs, discussion boards and SRSs of the University. The comments of students are analysed in the form of different emotions such as anger, anticipation, disgust, joy, fear, sadness, surprise and trust. Two new emotions, i.e., satisfaction and dissatisfaction are computed from these existing emotions. The results given by system are validated by comparing the indirect assessment with the direct assessment. The results given by the system are very promising and can help the administration of the University to improve the quality of teaching and learning of students.

A web-based sentiment analysis system has been developed for Hindi language which performs the sentiment analysis at aspect, sentence and document level. This web-based system is developed in Python language and is available at http://www.hindisenti.com/. The system to able to perform the sentiment analysis of Twitter posts about any user-defined Hashtag. The system is integrated with Quillpad API so that user can type in English and text will get transliterated into Hindi language; and a mapping dictionary is used to handle the abbreviations. The results given by system are represented using bar charts and graphs.

7.2 Future Work

The developed web-based Hindi sentiment analysis system can benefit in a number of areas of NLP. However, the system has some limitaions which can be improved in future. The experimentation with other deep learning models can be performed along with extension of the corpus for other domains such as products, restaurant reviews, social media analysis and political analysis etc. to build the system in general so that can be used in other applications.

The accuracy of the aspect-based sentiment analysis system can be improved by handling the issues like handling of multiword expressions, word sense disambiguation and sarcasm detection etc. Also, WordNet based approach can be used in future for the extension of the sentiment lexicons to achieve better accuracy. The coverage of mapping dictionary can help in improving the perfoemance of the system.

In future, the SA system API can be integrated with SRSs and online learning portals to enable real-time analysis of student feedback so that the administrators of the University/Institute can take corrective actions before time. The aspect-level sentiment analysis of student feedback can further help in improving the accuracy of the system. Also, the other Indian languages can be added to extend the system's multilingual capabilities.

However, the results given by the system are very encouraging and the proposed system has a potential to provide impetus to different emerging applications like SA of product reviews, social media analysis for effective policing, transparency in governance, public participation in decision-making, women empowerment, controlling riots and crime *etc.* for the benefits of society.

References

[1] (2005) Indian language families. http://www.indianetzone.com/39/indian_language_families.htm [Accessed 20 June 2017]

[2] (2012) Shallow parsers, Language Technologies Research Centre (LTRC), IIIT Hyderabad. http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php [Accessed 25 June 2017].

[3] (2014) Indo-Aryan languages. http://www.indianetzone.com/11/indo_aryan_languages.htm [Accessed 22 June 2017].

[4] Agarwal, A., (2019). Interesting Statistics On Google Plus!. https://trak.in/tags/business/2011/07/22/google-plus-statistics/. [Accessed on: August 15, 2019].

[5] Akhtar, M.S., Ekbal, A., Bhattacharyya, P., (2016a). Aspect based sentiment analysis: category detection and sentiment classification for Hindi. *In: 17th International conference on intelligent text processing and computational linguistics*, pp. 1–12.

[6] Akhtar, M.S., Ekbal, A., Bhattacharyya, P., (2016b). Aspect based sentiment analysis in Hindi: resource creation and evaluation. *In: Proceedings of the 10th international conference on language resources and evaluation*, pp. 1–7.

[7] Akhtar, M.S., Kumar, A., Ekbal, A., Bhattacharyya P., (2016c). A hybrid deep learning architecture for sentiment analysis. *In: Proceedings of the 26th international conference on computational linguistics*, pp. 482–493.

[8] Al-Amin, M., Islam, M.S. and Uzzal, S.D., (2017). Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words. *In International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 186-190.

[9] Alansary, S., (2012). A formalized reference grammar for UNL-based machine translation between English and Arabic. *In Proceedings of 24th International Conference on Computational Linguistics*, pp. 33-42.

[10] Alansary, S., Nagi, M., Adly, N. and Egypt, A., (2006). Towards a language-independent Universal digital library. *In Second International Conference on*

Universal Digital Libraries (ICUDL). Alexandria, Egypt, pp. 1-10.

[11] Ali, M. and Wagan, A.I., (2017). Sentiment summerization and analysis of Sindhi text. *International Journal of Advanced Computer Science and Applications*, 8(10), pp. 296-300.

[12] Altrabsheh, N., Cocea, M. and Fallahkhair, S., (2014). Learning sentiment from students' feedback for real-time interventions in classrooms. *In International Conference on Adaptive and Intelligent Systems*, pp. 40-49.

[13] Altrabsheh, N., Gaber, M.M. and Cocea, M., (2013). SA-E: sentiment analysis for education. *In International Conference on Intelligent Decision Technologies*, 255, pp. 353-362.

[14] Amin, A., Hossain, I., Akther, A. and Alam, K.M., (2019). Bengali VADER: A Sentiment Analysis Approach Using Modified VADER. *In International Conference on Electrical, Computer and Communication Engineering (ECCE),* pp. 1-6.

[15] Anagha, M., Kumar, R. R., Sreetha, K., Rajeev, R., Raj, P.R., (2014). Lexical resource based hybrid approach for cross domain sentiment analysis in Malayalam. International Journal of Engineering Sciences, 15:18–21.

[16] Anagha, M., Kumar, R.R., Sreetha, K. and Raj, P.R., (2015). Fuzzy logic based hybrid approach for sentiment analysis of malayalam movie reviews. *In International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1-4.

[17] Araque, O., Corcuera, I., Román, C., Iglesias, C. A., and Sánchez-Rada, J. F., (2015). Aspect based Sentiment Analysis of Spanish Tweets. *In Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, CEUR Series, 1397: pp. 29-34.

[18] Angelica, A. D., Kurzweil Accelerating Intelligence. Social networks, surveillance, and terrorism [Blog]. Available: http://www.kurzweilai.net/social-networks-surveillance-and-terrorism [Accessed: 8 August 2019]

[19] Arora, P., (2013). Sentiment Analysis for Hindi Language, *Doctoral diss.,* International Institute of Information Technology Hyderabad.

[20] Arora, P. and Kaur, B., (2015). Sentiment analysis of political reviews in

Punjabi language. *International Journal of Computer Applications*, 126(14), pp. 1-4.

[21] Asghar, M.Z., Sattar, A., Khan, A., Ali, A., Masud Kundi, F. and Ahmad, S., (2019). Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. *Expert Systems*, pp. 1-19.

[22] Ashna, M.P. and Sunny, A.K., (2017). Lexicon based sentiment analysis system for malayalam language. *In International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 777-783.

[23] Aung, K.Z. and Myo, N.N., (2017). Sentiment analysis of students' comment using lexicon based approach. *In 16th International Conference on Computer and Information Science (ICIS)*, pp. 149-154.

[24] Baccianella, S., Esuli, A. and Sebastiani, F., (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *In: Proceedings of language resources evaluation*, 10, pp. 2200-2204.

[25] Bae, M.H., Kim, Y.I., Kim, S.C., Lee, D.H. and Otgonbayar, B., (2009). Performance analysis of mobile wimax mmr system with vertical handover. *The Journal of Korean Institute of Communications and Information Sciences*, 34(11A), pp. 844-851.

[26] Bakliwal, A., Arora, P., Varma, V., (2012). Hindi subjective lexicon: a lexical resource for Hindi polarity classification. *In: Proceedings of the eight international conference on language resources and evaluation*, pp. 1189–1196.

[27] Balahadia, F.F., Fernando, M.C.G. and Juanatas, I.C., (2016). Teacher's performance evaluation tool using opinion mining with sentiment analysis. *In IEEE 10 Symposium (TENSYMP)*, pp. 95-98.

[28] Balamurali, A., Joshi, A., Bhattacharyya, P., (2012). Cross-lingual sentiment analysis for Indian languages using linkedWordnets. *In: Proceedings of 24th international conference on computational linguistics: posters*, pp. 73–82.

[29] Bansal, N., Ahmed, U.Z., and Mukherjee, A., (2013). Sentiment analysis in Hindi. Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India, pp.1-10.

[30] Bhattacharyya, P., (2017). Indowordnet. *In: The WordNet in Indian languages*. Springer, pp. 1–18.

[31] Bora, N. N., (2011). Feature Based Sentiment Analysis on Twitter, *Doctoral diss.*, Indian Institute of Technology Guwahati.

[32] Chand, S., (2016). Indian languages: classification of Indian languages. http://www.yourarticlelibrary.com/language/indian-languages-classification-of-indian- languages/19813/. [Accessed 22 June 2017].

[33] Chaudhari, C.V., Khaire, A.V., Murtadak, R.R., Sirsulla, K.S., (2017). Sentiment analysis in Marathi using Marathi WordNet. *Imperial Journal of Interdisciplinary Research (IJIR)*, 3(4), pp. 1253–1256.

[34] Clement, J., (2019a) Leading countries based on number of Facebook users as of July 2019 (in millions). https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/. [Accessed on: August 15, 2019].

[35] Clement, J., (2019b) Leading countries based on number of Twitter users as of July 2019 (in millions). https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. [Accessed on: August 15, 2019].

[36] Das, A. and Bandyopadhyay, S., (2010a). Phrase-level polarity identification for Bangla. *International Journal of Computational Linguistics and Applications (IJCLA)*, 1(1-2), pp.169-182.

[37] Das, A., Bandyopadhyay, S. (2010b). Sentiwordnet for Bangla. *Knowledge Shared Event Task* 2:1–9.

[38] Das, A., Bandyopadhyay, S., (2010c). Sentiwordnet for Indian languages. *In: Asian federation for natural language processing*, pp. 56–63.

[39] Deepamala, N., Kumar, R. (2015). Polarity detection of Kannada documents. *In: International advance computing conference*. pp. 764–767.

[40] Dehkharghani, R., Yanikoglu, B., Saygin, Y. and Oflazer, K., (2017). Sentiment analysis in Turkish at different granularity levels. *Natural Language Engineering*, 23(4), pp. 535-559.

[41] Denef, S., Bayerl, P.S. and Kaptein, N.A., (2013). Social media and the police: tweeting practices of british police forces during the August 2011 riots. *In*

proceedings of the SIGCHI conference on human factors in computing systems, pp. 3471-3480.

[42] Deshmukh, S., Patial, N., Rotiwar, S., and Nunes, J., (2017). Sentiment analysis of marathi language. *International Journal of Research Publications in Engineering and Technology (IJRPET)*, 3(6), pp. 93-97.

[43] Dhanalakshmi, V., Bino, D. and Saravanan, A.M., (2016). Opinion mining from student feedback data using supervised learning algorithms. *In 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, pp. 1-5.

[44] Di Caro, L., and Grella, M., (2013). Sentiment Analysis via Dependency Parsing. *Computer Standards & Interfaces*, 35: pp. 442-453.

[45] Erdmann, M., Ikeda, K., Ishizaki, H., Hattori, G., and Takishima, Y., (2014). Feature based Sentiment Analysis of Tweets in Multiple Languages. *In Proceedings of Web Information Systems Engineering*, pp. 109-124.

[46] Esparza, G.G., Díaz, A.P., Canul-Reich, J., De-Luna, C.A. and Ponce, J., (2016). Proposal of a Sentiment Analysis Model in Tweets for improvement of the teaching-learning process in the classroom using a corpus of subjectivity. *International Journal of Combinatorial Optimization Problems and Informatics*, 7(2), pp. 22-34.

[47] Esparza, G.G., de-Luna, A., Zezzatti, A.O., Hernandez, A., Ponce, J., Álvarez, M., Cossio, E. and de Jesus Nava, J., (2018). A sentiment analysis model to analyze students reviews of teacher performance using support vector machines. *In International Symposium on Distributed Computing and Artificial Intelligence*, pp. 157-164.

[48] Esuli, A. and Sebastiani, F., (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *In: International conference on language resources and evaluation*, 6: pp. 417-422.

[49] Esuli, A., Sebastiani, F. (2007). Sentiwordnet: a high-coverage lexical resource for opinion mining. *In: International conference on language resources and evaluation*, pp. 1–26.

[50] Evans, C., (2013). Making sense of assessment feedback in higher education. *Review of educational research*, 83(1), pp.70-120.

[51] Fondekar, A., Pawar, J.D., Karmali, R., (2016). Konkani sentiwordnet: resource for sentiment analysis using supervised learning approach. *In: Workshop on Indian language data: resources and evaluation (WILDRE3)*, Portoroz, Slovenia, pp. 55–59.

[52] Garg, K. and Buttar, P.K., (2017). Aspect based sentiment analysis of hindi text review. *International Journal of Advanced Research in Computer Science*, 8(7), pp. 831-836.

[53] Garrison, C. and Ehringhaus, M., (2007). Formative and summative assessments in the classroom, pp. 1-3.

[54] Ghosal, T., Das, S.K. and Bhattacharjee, S., (2015). Sentiment analysis on (Bengali horoscope) corpus. *In Annual IEEE India Conference (INDICON)*, pp. 1-6.

[55] Gohil, L., and Patel, D., (2019). A Sentiment Analysis of Gujarati Text using Gujarati Senti word Net. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(9), pp. 2290-2292.

[56] Gupta, A. and Kumaraguru, P., (2012). Credibility ranking of tweets during high impact events. *In Proceedings of the 1st workshop on privacy and security in online social media*, pp. 1-8.

[57] Gupta, A., Kumaraguru, P., Castillo, C. and Meier, P., (2014). Tweetcred: Real-time credibility assessment of content on twitter. *In International Conference on Social Informatics*, pp. 228-243.

[58] Gupta, C.P., Bal, B.K., (2015). Detecting sentiment in Nepali texts: a bootstrap approach for sentiment analysis of texts in the Nepali language. *In International conference on cognitive computing and information processing*. pp. 1–4.

[59] Hasan, K.A. and Rahman, M., (2014). Sentiment detection from bangla text using contextual valency analysis. *In 17th International Conference on Computer and Information Technology (ICCIT)*, pp. 292-295.

[60] Hassan, A., Amin, M.R., Al Azad, A.K. and Mohammed, N., (2016). Sentiment analysis on bangla and romanized bangla text using deep recurrent models. *In International Workshop on Computational Intelligence (IWCI)*, pp. 51-56.

[61] Hegde, Y. and Padma, S.K., (2015). Sentiment analysis for Kannada using

mobile product reviews: a case study. *In International Advance Computing Conference (IACC),* pp. 822-827.

[62] Hegde, Y. and Padma, S.K., (2017). Sentiment analysis using random forest ensemble for mobile product reviews in Kannada. *In 7th International Advance Computing Conference (IACC)*, pp. 777-782.

[63] Hoque, M.T., Islam, A., Ahmed, E., Mamun, K.A. and Huda, M.N., (2019). Analyzing Performance of Different Machine Learning Approaches with Doc2vec for Classifying Sentiment of Bengali Natural Language. *In International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1-5.

[64] Hu, M. and Liu, B., (2004). Mining and summarizing customer reviews. *In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168-177.

[65] Hussaini, F., Padmaja, S. and Fatima, S.S., (2018). Score-based sentiment analysis of book reviews in hindi language. *International Journal on Natural Language Computing (IJNLC)*, 7(5), pp. 115-127.

[66] Jawale, M.A., Kyatanavar, D.N. and Pawar, A.B., (2013). Design of automated sentiment or opinion discovery system to enhance its performance. *In Proceedings of International Conference on Advances in Information Technology and Mobile Communication 2013 (AIM 2013)* and *in ACEEE 2013 Digital Library*, pp. 48-53.

[67] Jayan, P., Nair, D.S. and Elizabeth Jisha, S., (2015). A subjective feature extraction for sentiment analysis in Malayalam language. *International Journal of Engineering Sciences,* 14: pp.1-4.

[68] Jena, M.K. and Chandra, B.R., (2014). Opinion mining for online Oriya text. *European Journal of Academic Essays (EJAE),* pp. 44-48.

[69] Jha, V., Manjunath, N., Shenoy, P.D., Venugopal, K.R. and Patnaik, L.M., (2015). Homs: Hindi opinion mining system. *In 2nd International Conference on Recent Trends in Information Systems (ReTIS),* pp. 366-371.

[70] Jiménez-Zafra, S.M., Martín-Valdivia, M.T., Martínez-Cámara, E. and Ureña-López, L.A., (2016). Combining resources to improve unsupervised sentiment

analysis at aspect-level. *Journal of Information Science*, 42(2), pp. 213-229.

[71] Joshi, A., Balamurali, A., Bhattacharyya, P., (2010). A fall-back strategy for sentiment analysis in Hindi: a case study. *In: Proceedings of the 8th international conference on natural language processing*, pp. 1–6.

[72] Joshi, V.C. and Vekariya, V.M., (2017.) An approach to sentiment analysis on Gujarati tweets. *Advances in Computational Sciences and Technology*, 10(5), pp.1487-1493.

[73] Jupyter [Online]. http://jupyter.org/. Accessed 2 June 2017

[74] Kaur, A. and Gupta, V., (2014). N-gram based approach for opinion mining of Punjabi text. *In International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pp. 81-88.

[75] Kaur, A., Gupta, V. (2014b). Proposed algorithm of sentiment analysis for Punjabi text. *Journal of Emerging Technologies in Web Intelligence,* 6(2):180–183.

[76] Kaur, A. and Gupta, V., (2017). A Novel Approach for Sentiment Analysis of Punjabi Text using SVM. *International Arab Journal of Information Technology (IAJIT),* 14(5), pp. 707-712.

[77] Kaur, G. and Kaur, K., (2017). Sentiment Detection from Punjabi Text using Support Vector Machine. *International Journal of Scientific Research in Computer Science and Engineering*, 5(6), pp. 39-46.

[78] Kaur, J., and Saini, J.R., (2014). A study and analysis of opinion mining research in Indo-Aryan, Dravidian and Tibeto-Burman language families. *International journal of data mining and emerging technologies*, 4(2), pp. 53-60.

[79] Kar, A., and Mandal, D. P., (2011). Finding Opinion Strength using Fuzzy Logic on Web Reviews. *International Journal of Engineering and Industries*. 2(1): pp. 37-43.

[80] Kechaou, Z., Ammar, M.B. and Alimi, A.M., (2011). Improving e-learning with sentiment analysis of users' opinions. In global engineering education conference (EDUCON), pp. 1032-1038.

[81] Khan, M.Y., Emaduddin, S.M., and Junejo, K.N., (2017). Harnessing English

Sentiment Lexicons for Polarity Detection in Urdu Tweets: A Baseline Approach. *In: Proceedings of 11th International Conference on Semantic Computing (ICSC)*, pp. 242-249.

[82] Kitchenham, B., and Charters, S., (2007). Guidelines for performing systematic literature reviews in software engineering. EBSE technical report.

[83] Kumar, A., Kohail, S., Ekbal, A., and Biemann, C., (2015a). IIT-TUDA: System for sentiment analysis in indian languages using lexical acquisition. *In International Conference on Mining Intelligence and Knowledge Exploration*, pp. 684-693.

[84] Kumar, K.A., Rajasimha, N., Reddy, M., Rajanarayana, A. and Nadgir, K., (2015b). Analysis of users' Sentiments from Kannada Web Documents. *Procedia Computer Science*, 54: pp. 247-256.

[85] Kumar, S.S., Premjith, B., Kumar, M.A., and Soman, K.P., (2015c). AMRITA_CEN-NLP@ SAIL2015: sentiment analysis in Indian Language using regularized least square approach with randomized feature learning. *In International Conference on Mining Intelligence and Knowledge Exploration*, pp. 671-683.

[86] Kumar, S.S., Kumar, M.A. and Soman, K.P., (2017). Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. *In International Conference on Mining Intelligence and Knowledge Exploration*, pp. 320-334.

[87] Kumar, S.S., Kumar, M.A. and Soman, K.P., (2019). Identifying Sentiment of Malayalam Tweets Using Deep Learning. *In Digital Business*, pp. 391-408.

[88] Liu, B., (2012). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2nd edn. pp. 1-38.

[89] Liu, B., (2010). Sentiment Analysis: A Multi-Faceted Problem. *IEEE Intelligent Systems*, pp. 1-5.

[90] Liu, T., Cho, K., Broadwell, G. A., Shaikh, S., Strzalkowski, T., Lien, J., Taylor, S. M., Feldman, L., Yamrom, B., Webb, N., and Boz, U., (2014). Automatic Expansion of the MRC Psycholinguistic Database Imageability Ratings. *In Proceeedings of the 9th International Conference on Language*

Resources and Evaluation, pp. 2800-2805.

[91] Manson, D., Meldal, S., Sledge, C., Maurer, S.M., Mitchell, J.C., Spengler, E., Sztipanovits, J. and Torner, J., (2006). Panel Session-Learning Modules for Security, Privacy and Information Assurance In Undergraduate Engineering Education. *In Proceedings of 36th Annual Conference Frontiers in Education*, pp. 1-2.

[92] Meldal, S., Gates, K., Smith, R. and Su, X., (2008). Security, Safety and Privacy–Pervasive Themes for Engineering Education, *In International Conference on Engineering Education (ICEE)*, pp. 1-8.

[93] Miranda, D.T. and Mascarenhas, M., (2016). Kop: an opinion mining system in Konkani. *In International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 702-705.

[94] Mishra, B.K. and Sahoo, A.K., (2016). Evaluation of Faculty Performance in Education System using Classification technique in Opinion mining based on GPU. *In Computational Intelligence in Data Mining*, 2, pp. 109-119.

[95] Mittal, N., Agarwal, B., Chouhan, G., Pareek, P., and Bania, N., (2013). Discourse based Sentiment Analysis for Hindi Reviews. *In Proceedings of Pattern Recognition and Machine Intelligence*, pp. 720–725.

[96] Mohammad, S. M., and Turney, P. D., (2013). Crowdsourcing a Word–emotion Association Lexicon. *Computational Intelligence*, 29(3): pp. 436-465.

[97] Mosha, C., and Tianfang, Y., (2010). Combining Dependency Parsing with Shallow Semantic Analysis for Chinese Opinion-Element Relation Identification. *In 4th International Universal Communication* Symposium, pp. 299-305.

[98] Mukherjee, S. and Bhattacharyya, P., (2012). Feature Specific Sentiment Analysis for Product Reviews. *In Proceedings of Computational Linguistics and Intelligent Text Processing*, pp. 475-487.

[99] Mukhtar, N. and Khan, M.A., (2018). Urdu sentiment analysis using supervised machine learning approach. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(02), pp. 1-15.

[100] Mukhtar, N. and Khan, M.A., (2019). Effective lexicon-based approach for

Urdu sentiment analysis. *Artificial Intelligence Review*, pp.1-28.

[101] Mukhtar, N., Khan, M.A. and Chiragh, N., (2017). Effective use of evaluation measures for the validation of best classifier in Urdu sentiment analysis. *Cognitive Computation*, 9(4), pp. 446-456.

[102] Mukku, S.S., Choudhary, N. and Mamidi, R., (2016). Enhanced Sentiment Classification of Telugu Text using ML Techniques. *SAAIP@ 25th international joint conference on artificial intelligence*, pp.29-34.

[103] Mukku, S.S., and Mamidi, R., (2017). Actsa: Annotated corpus for Telugu sentiment analysis. *In Proceedings of the first workshop on building linguistically generalizable NLP systems*, pp. 54-58.

[104] Munezero, M., Montero, C.S., Mozgovoy, M. and Sutinen, E., (2013). Exploiting sentiment analysis to track emotions in students' learning diaries. *In Proceedings of the 13th Koli Calling International Conference on Computing Education Research*, pp. 145-152.

[105] Naidu, R., Bharti, S.K., Babu, K.S. and Mohapatra, R.K., (2017). Sentiment analysis using Telugu sentiwordnet. *In International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 666-670.

[106] Nair, D.S., Jayan, J.P. and Sherly, E., (2014). SentiMa-sentiment extraction for Malayalam. *In International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1719-1723.

[107] Nair, D.S., Jayan, J.P., Rajeev, R.R. and Sherly, E., (2015). Sentiment Analysis of Malayalam film review using machine learning techniques. *In International conference on advances in computing, communications and informatics (ICACCI)*, pp. 2381-2384.

[108] Nakagawa, H. and Mori, T., (2002). A Simple but Powerful Automatic Term Extraction Method. *In 2nd International Workshop on Computational Terminology*, 14: pp. 1-7.

[109] Nanda, C., Dua, M. and Nanda, G., (2018). Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning. *In International Conference on Communication and Signal Processing (ICCSP)*, pp. 1069-1072.

[110] Nasim, Z., Rajput, Q. and Haider, S., (2017). Sentiment analysis of student

feedback using machine learning and lexicon based approaches. *In International Conference on Research and Innovation in Information Systems (ICRIIS)*, pp. 1-6.

[111] Nivedhitha, E., Sanjay, S., Anand Kumar, M. and Soman, K., (2016). Unsupervised word embedding based polarity detection for Tamil tweets. International Journal of Computer Technology and Applications, 9(10), pp. 4631-4638.

[112] Nongmeikapam, K., Khangembam, D., Hemkumar, W., Khuraijam, S. and Bandyopadhyay, S., (2014). Verb based manipuri sentiment analysis. *International Journal on Natural Language Computing (IJNLC)*, 3: pp.12-13.

[113] Padmamala, R. and Prema, V., (2017). Sentiment analysis of online Tamil contents using recursive neural network models approach for Tamil language. *In International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM)*, pp. 28-31.

[114] Pandey, P. and Govilkar, S., (2015). A framework for sentiment analysis in Hindi using HSWN. *International Journal of Computer Applications*, 119(19): pp. 23-26.

[115] Pang, B., and Lee, L., (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), pp. 1-135.

[116] Pang, B., Lee, L. and Vaithyanathan, S., (2002). Thumbs up?: sentiment classification using machine learning techniques. *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 10: pp. 79-86.

[117] Parihar, M., (2012). Social media: The final frontier in customer experience management. Epoch Strategies for Marketing, Family Business and Entrepreneurship, Forthcoming, *Proceedings of the 15th Nirma International Conference on Management*, pp. 1-18.

[118] Patel, T., Undavia, J. and Patela, A., (2015). Sentiment analysis of parents feedback for educational institutes. *International Journal of Innovative and Emerging Research in Engineering*, 2(3), pp. 75-78.

[119] Patra, B.G., Das, D., Das, A., Prasath, R., (2015). Shared task on sentiment analysis in Indian languages (sail) tweets an overview. *In: International conference on mining intelligence and knowledge exploration.* Springer, pp. 650–655.

[120] Phani, S., Lahiri, S., and Biswas, A., (2016). Sentiment analysis of tweets in three Indian languages. *In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016),* pp. 93-102.

[121] Poria, S., Ofek, N., Gelbukh, A., Hussain, A., and Rokach, L., (2014). Dependency Tree-based Rules for Concept-level Aspect-based Sentiment Analysis. *In Proceedings of Semantic Web Evaluation Challenge*, pp. 41-47.

[122] Prasad, S.S., Kumar, J., Prabhakar, D.K., and Pal, S., (2015). Sentiment classification: an approach for Indian language tweets using decision tree. *In International Conference on Mining Intelligence and Knowledge Exploration*, pp. 656-663.

[123] Raghuwanshi, G. and Tyagi, V., (2019). Impact of feature extraction techniques on a CBIR system. *In International Conference on Advances in Computing and Data Sciences*, pp. 338-348.

[124] Rahmath, H., (2014). Opinion mining and sentiment analysis-challenges and applications. *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* 3(5), pp.1-3.

[125] Rai, V., Vijay, S. and Misra, D., (2017). Linguistic approach based Transfer Learning for Sentiment Classification in Hindi. *In Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017),* pp. 373-382.

[126] Rajput, H., (2014). Social media and politics in India: A study on Twitter usage among Indian political leaders, *Asian Journal of Multidisciplinary Studies*, 2(1), pp. 63-69.

[127] Rana, T. A., and Cheah, Y., (2016). Hybrid Rule-Based Approach for Aspect Extraction and Categorization from Customer Reviews. *In 9th International Conference on IT in Asia*, pp. 1-5.

[128] Rani, S. and Kumar, P., (2019). Deep Learning Based Sentiment Analysis

Using Convolution Neural Network. *Arabian Journal for Science and Engineering*, 44(4), pp. 3305-3314.

[129] Rehman, Z.U., and Bajwa, I.S., (2016). Lexicon-based sentiment analysis for Urdu language. *In 6th international conference on innovative computing technology (INTECH)*, pp. 497-501.

[130] Robaldo, L., and Di Caro, L., (2013). OpinionMining-ML. *Computer Standards & Interfaces*, 35(5), pp. 454-469.

[131] Rohini, V., Thomas, M. and Latha, C.A., (2016). Domain based sentiment analysis in regional language-Kannada using machine learning algorithm. *In International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 503-507.

[132] Sahu, S., Behera, P., Mohapatra, D., and Rakesh, C., (2016a). Information retrieval in web for an Indian language: an Odia language sentimental analysis context. *Journal of Computer Technology & Applications*, 9(22), pp. 249–256.

[133] Sahu, S.K., Behera, P., Mohapatra, D.P. and Balabantaray, R.C., (2016b). Sentiment analysis for Odia language using supervised classifier: an information retrieval in Indian language initiative. *CSI transactions on ICT*, 4(2-4), pp.111-115.

[134] Salama, H. and Alansary, S., (2016). Building a POS-Annotated corpus for Egyptian children. *The Egyptian Journal of Language Engineering*, 3(1), pp. 12-23.

[135] Salas-Zárate, M.D.P., Valencia-García, R., Ruiz-Martínez, A. and Colomo-Palacios, R., (2017). Feature-based opinion mining in financial news: an ontology-driven approach. *Journal of Information Science*, 43(4), pp. 458-479.

[136] Sarkar, K., (2018). Using Character N-gram Features and Multinomial Naïve Bayes for Sentiment Polarity Detection in Bengali Tweets. *In 5th International Conference on Emerging Applications of Information Technology (EAIT)*, pp. 1-4.

[137] Sarkar, K., (2019). Sentiment Polarity Detection in Bengali Tweets Using Deep Convolutional Neural Networks. *Journal of Intelligent Systems*, 28(3), pp.377-386.

[138] Sarkar, K. and Bhowmick, M., (2017). Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines. *In Calcutta Conference (CALCON),* pp. 31-36.

[139] Sarkar, K. and Chakraborty, S., (2015). A sentiment analysis system for Indian language tweets. *In International Conference on Mining Intelligence and Knowledge Exploration,* pp. 694-702.

[140] Se, S., Vinayakumar, R., Kumar, M.A. and Soman, K.P., (2015). AMRITA-CEN@ SAIL2015: sentiment analysis in Indian languages. *In International Conference on Mining Intelligence and Knowledge Exploration,* pp. 703-710.

[141] Se, S., Vinayakumar, R., Kumar, M.A., and Soman, K.P., (2016). Predicting the sentimental reviews in tamil movie using machine learning algorithms. *Indian Journal of Science and Technology,* 9(45), pp. 1-5.

[142] Segura, M.I.S., Dominguez, F.M., Dugarte-Peña, G., and Goñi, A. J., (2016). Software engineers must speak the systemic intangible process assets language. *SWEBOK Evolution: Virtual Town Hall Meeting, IEEE Computer Society*, pp. 1-5.

[143] Seshadri, S., Madasamy, A.K., Padannayil, S.K., and Anand Kumar, M., (2016). Analyzing sentiment in Indian languages micro text using recurrent neural network. *Institute of Integrative Omics and Applied Biotechnology,* 7: pp. 313-318.

[144] Shaikh, S., Cho, K., Strzalkowski, T., Liu, T., and Lien, J., (2016). ANEW+: Automatic Expansion and Validation of Affective Norms of Words Lexicons in Multiple Languages. *In Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC),* pp. 1127-1132.

[145] Shalini, K., Ravikurnar, A., Vineetha, R.C., Aravinda, R.D., Anand, K.M. and Soman, K.P., (2018). Sentiment Analysis of Indian Languages using Convolutional Neural Networks. *In International Conference on Computer Communication and Informatics (ICCCI),* pp. 1-4.

[146] Sharma, R., Nigam, S., and Jain, R., (2014). Polarity detection movie reviews in Hindi language. *International Journal on Computational Sciences & Applications (IJCSA),* 4(4): pp. 49-57.

[147] Sharma, R., and Bhattacharyya, P., (2014). A sentiment analyzer for hindi using hindi senti lexicon. *In Proceedings of the 11th International Conference on Natural Language Processing*, pp. 150-155.

[148] Sharma, Y., Mangat, V., and Kaur, M., (2015). A practical approach to sentiment analysis of Hindi tweets. *In 1st International Conference on Next Generation Computing Technologies (NGCT)*, pp. 677-680.

[149] Sharma, P. and Moh, T.S., (2016). Prediction of indian election using sentiment analysis on hindi twitter. *In International Conference on Big Data (Big Data)*, pp. 1966-1971.

[150] Sharmista, A. and Ramaswami, M., (2016). Tree based opinion mining in Tamil for product recommendations using R. *International Journal of Computational Intelligence and Informatics*, 6(2), pp.108-116.

[151] Sharmista, A. and Ramaswami, M., (2018). SVM and Fuzzy SVM Based Opinion Mining In Tamil Using R. *American Journal of Engineering Research (AJER)*, 7(4), pp. 45-53.

[152] Singh, J., Singh, G., Singh, R. and Singh, P., (2018). Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification. *Journal of King Saud University-Computer and Information Sciences*, pp. 1-10.

[153] Singh, V. K., Piryani, R., Uddin, A., and Waila, P., (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. *In International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, pp. 712-717.

[154] Stevens, L., Social Media in Policing: Nine Steps for Success [Online]. Available: http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=print_display&article_id=2018&issue_id=22010 [Accessed: 11 July 2019]

[155] Sumit, S.H., Hossan, M.Z., Al Muntasir, T. and Sourov, T., (2018). Exploring Word Embedding for Bangla Sentiment Analysis. *In International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1-5.

[156] Svensson, K., (2017). Sentiment analysis with convolutional neural networks: classifying sentiment in Swedish reviews, Bachelor Thesis Project, Linnaeus University, Sweden.

[157] Syed, A.Z., Aslam, M., and Martinez-Enriquez, A.M., (2010). Lexicon based sentiment analysis of Urdu text using SentiUnits. *In Mexican International Conference on Artificial Intelligence*, pp. 32-43.

[158] Syed, A.Z., Aslam, M. and Martinez-Enriquez, A.M., (2011). Sentiment analysis of urdu language: handling phrase-level negation. *In Mexican International Conference on Artificial Intelligence*, pp. 382-393.

[159] Syed, A.Z., Aslam, M. and Martinez-Enriquez, A.M., (2014). Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text. *Artificial intelligence review*, 41(4), pp.535-561.

[160] Tensorflow [Online]. https://www.tensorflow.org/. Accessed 3 June 2017

[161] TFLearn: Deep learning library featuring a higher-level API for TensorFlow [Online]. http://tflearn.org/. Accessed 5 June 2017

[162] Thapa, L.B.R. and Bal, B.K., (2016). Classifying sentiments in Nepali subjective texts. *In 7th International conference on information, intelligence, systems & applications (IISA)*, pp. 1-6.

[163] Thet, T. T., Na, J. C., and Khoo, C. S., (2010). Aspect-based Sentiment Analysis of Movie Reviews on Discussion Boards. *Journal of Information Science*, 36: pp. 823-848.

[164] Thulasi, P.K. and Usha, K., (2016). Aspect polarity recognition of movie and product reviews in Malayalam. *In International Conference on Next Generation Intelligent Systems (ICNGIS)*, pp. 1-5.

[165] Tripto, N.I. and Ali, M.E., (2018). Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments. *In International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1-6.

[166] Turney, P. D., (2002). Thumbs up or Thumbs down?: Semantic Orientation applied to Unsupervised Classification of Reviews. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424.

[167] Venugopalan, M. and Gupta, D., (2015). Sentiment classification for Hindi tweets in a constrained environment augmented using tweet specific features. *In International Conference on Mining Intelligence and Knowledge Exploration*, pp. 664-670.

[168] Vilares, D., Alonso, M. A. and Gómez-Rodríguez, C., (2015). A syntactic approach for opinion mining on Spanish reviews. *Natural Language Engineering*, 21 (01): pp. 139-163.

[169] Wiebe, J.M., Bruce, R.F. and O'Hara, T.P., (1999). 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*, pp. 246-253.

[170] Wu, Y., Zhang, Q., Huang, X., and Wu, L., (2009). Phrase Dependency Parsing for Opinion Mining. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3: pp. 1533-1541.

[171] Yadav, M. and Bhojane, V., (2019). Semi-Supervised Mix-Hindi Sentiment Analysis using Neural Network. *In 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 309-314.

[172] Yu, L.C., Lee, C.W., Pan, H.I., Chou, C.Y., Chao, P.Y., Chen, Z.H., Tseng, S.F., Chan, C.L. and Lai, K.R., (2018). Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning*, 34(4), pp. 358-365.

[173] Zhang, Q., Wu, Y., Li, T., Ogihara, M., Johnson, J., and Huang, X., (2009). Mining Product Reviews Based on Shallow Dependency Parsing. *In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 726-727.

Publications

Journals

[1] Sujata Rani and Parteek Kumar, “A Journey of Indian Languages over Sentiment Analysis: A Systematic Review", Artificial Intelligence Review, Springer, 2018 [SCIE indexed, IF=5.095]

[2] Sujata Rani and Parteek Kumar, “Deep Learning based Sentiment Analysis using Convolution Neural Network”, Arabian Journal of Science and Engineering, 2018 [IF = 1.518]

[3] Sujata Rani and Parteek Kumar, “Sentiment Analysis of Social Media using Machine Learning Techniques: Social Enablement”, Digital Scholarships in Humanities, Oxford Press, 2018 [IF = 0.418]

[4] Sujata Rani and Parteek Kumar, “Sentiment Analysis System to improve Teaching and Learning”, Computer, IEEE Computer Society, 50 (5), pp. 36-43, 2017 [SCIE indexed, IF= 3.564]

Conferences

[1] Sujata Rani and Parteek Kumar, “Desirable features for an effective Sentiment Analysis System”, International Conference on Information, Communication and Computing Technology, India International Centre (IIC), New Delhi, 2017.

[2] Sujata Rani and Parteek Kumar, “Rule based Sentiment Analysis System for Analyzing tweets”, International Conference on Infocom Technologies and Unmanned Systems, Amity University, Dubai, 2017.