

# **METHODS OF SENTIMENT ANALYSIS ON TWITTER DATA FOR ENGLISH AND HINDI LANGUAGE**

*Dissertation submitted in fulfilment of the requirements for the Degree of*

*Integrated B.Tech - M.Tech*

**MASTER OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

By

**Aarsh Agrawal**

**11703262**

Supervisor

**Vinay Bhardwaj**



**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

December 2021

@ Copyright LOVELY PROFESSIONAL UNIVERSITY, Punjab (INDIA)

December 2021

ALL RIGHTS RESERVED

# PAC FORM



## TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering (SCSE)

**Program :** P192-ND::Integrated B.Tech. - M.Tech. (Computer Science & Engineering)

**COURSE CODE :** CSE548

**REGULAR/BACKLOG :** Regular

**GROUP NUMBER :** CSERGD0083

**Supervisor Name :** Dr. Vinay Bhardwaj

**UID :** 23825

**Designation :** Assistant Professor

**Qualification :** \_\_\_\_\_

**Research Experience :** \_\_\_\_\_

SR.NO.	NAME OF STUDENT	Prov. Regd. No.	BATCH	SECTION	CONTACT NUMBER
1	Aarsh Agrawal	11703262	2017	K17MT	8934830158

**SPECIALIZATION AREA :** Networking and Security

**Supervisor Signature:** \_\_\_\_\_

**PROPOSED TOPIC :** Sentiment Analysis for Twitter data

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	6.80
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.20
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.60
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	8.00
5	Social Applicability: Project work intends to solve a practical problem.	6.20
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.80

PAC Committee Members		
PAC Member (HOD/Chairperson) Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member (Allied) Name: Cherry Khosla	UID: 13436	Recommended (Y/N): Yes
PAC Member 3 Name: Ravishanker	UID: 12412	Recommended (Y/N): Yes

**Final Topic Approved by PAC:** Sentiment Analysis for Twitter data

**Overall Remarks:** Approved

**PAC CHAIRPERSON Name:** 13897::Dr. Deepak Prashar

**Approval Date:** 17 Dec 2021

12/17/2021 7:44:09 PM

## ABSTRACT

---

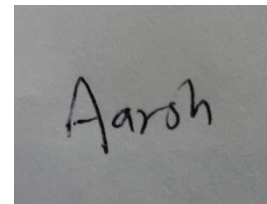
Social media is widely regarded as one of the most important unstructured data. Analyzing and extracting meaning from such data is a time-consuming process. Because of the enormous data available on social media platforms, sentiment extraction has gotten a lot of attention. Microblogging is a relatively new phenomenon, with Twitter being the most widely utilized. It's one of the most comprehensive free and open data sources available. Today's society sees a lot of differing viewpoints on Twitter. Researchers can use opinion mining to obtain the present emotion and mood of the public. Sentiment Analysis is defined as the technique of extracting and finding the polarity of a given material to get insight into the hidden information, emotion, feeling contained within a text. The ultimate objective of sentiment analysis is to extract meaningful material from various sources of information. The first analysis of tweets was done using the Natural Language Processing (NLP) method. For further analysis of the opinionated data, two approaches are available: the Lexicon Based Approach (LBA) and the Machine Learning Approach (MLA) based on supervised learning. The LBA approach employs a resource dictionary, namely the Hindi SentiWordNet, and a Hybrid Based Approach (HBA) that joins the Lexicon based and Machine learning for categorizing tweets as positive or negative.

## DECLARATION STATEMENT

---

I hereby declare that the research work reported in the dissertation/dissertation proposal entitled "METHODS OF SENTIMENT ANALYSIS FOR HINDI AND ENGLISH LANGUAGE FOR TWITTER DATA" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Vinay Bhardwaj. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

A photograph of a handwritten signature in black ink on a light-colored surface. The signature appears to be 'Aarsh'.

*Signature of Candidate*

**Aarsh Agrawal**

**11703262**

## SUPERVISOR'S CERTIFICATE

---

This is to certify that the work reported in the M.Tech Dissertation/dissertation proposal entitled “**METHODS OF SENTIMENT ANALYSIS FOR HINDI AND ENGLISH LANGAGE FOR TWITTER DATA**”, submitted by **Aarsh Agrawal** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Vinay Bhardwaj)

**Date:**

**Counter Signed by:**

**1) Concerned HOD:**

HoD's Signature: \_\_\_\_\_

HoD Name: \_\_\_\_\_

Date: \_\_\_\_\_

**2) Neutral Examiners:**

**External Examiner**

Signature: \_\_\_\_\_

Name: \_\_\_\_\_

Affiliation: \_\_\_\_\_

Date: \_\_\_\_\_

**Internal Examiner**

Signature: \_\_\_\_\_

Name: \_\_\_\_\_

Date: \_\_\_\_\_

## **ACKNOWLEDGEMENTS**

---

I would like to convey my most heartfelt and sincere gratitude to my guide Mr. Vinay Bhardwaj of Lovely Professional University, for his valuable guidance and advice. His willingness to motivate me contributed tremendously to achieve the goal successfully. I would also like to thank God and my parents who has always inspired and encouraged me to achieve my goal successfully.

**Aarsh Agrawal**

**11703262**

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE NO.</b>
Inner first page – Same as cover	i
PAC form	ii
Abstract	iii
Declaration by the Scholar	iv
Supervisor’s Certificate	v
Acknowledgement	vi
Table of Contents	vii
List of Acronyms / Abbreviations (If any)	viii
List of Symbols (If any)	ix
List of Figures	x
List of Tables	xi
<b>CHAPTER 1: INTRODUCTION</b>	<b>12</b>
<b>1.1 WHAT IS SENTIMENT ANALYSIS</b>	<b>13</b>
<b>1.2 LEVELS OF SA</b>	<b>13</b>
<b>1.2.1 DOCUMENT LEVEL</b>	<b>13</b>
<b>1.2.2 SENTENCE LEVEL</b>	<b>13</b>
<b>1.2.3 ASPECT LEVEL</b>	<b>14</b>
<b>1.3 APPLICATIONS OF SENTIMENT ANALYSIS</b>	<b>15</b>
<b>1.4 CHALLENGES IN SENTIMENT ANALYSIS</b>	<b>15</b>
<b>1.4.1 FOR ENGLISH LANGUAGE</b>	<b>15</b>
<b>1.4.2 FOR HINDI LANGUAGE</b>	<b>17</b>
<b>CHAPTER 2: OBJECTIVE &amp; SCOPE OF THE STUDY</b>	<b>18</b>
<b>CHAPTER 3: REVIEW OF LITERATURE</b>	<b>19</b>

3.1 LITERATURE REVIEW	19
3.2 TABLE OF RIVIEW	20
3.3 LITERATURE REIVIEW OF HINDI LANGUAGE	21
3.4 TWITTER	22
<b>CHAPTER 4: RESEARCH TERMINOLOGIES</b>	<b>25</b>
4.1 NLTK	25
4.2 STEMMING	25
4.3 WORD LEMMENTIZATION	26
4.4 TEXTBLOB LIBRARY	26
4.5 TWEETPY	26
4.6 VADER	26
4.7 POS TAGGING	26
4.8 WORDNET	27
4.9 SENTI WORDNET	27
4.10 BAG OF WORDS	27
<b>CHAPTER 5: RESEARCH METHODLOGIES</b>	<b>28</b>
5.1 PYTHON	28
5.2 DATA PRE-PROCESSING	28
<b>CHAPTER 6: SENTIMENT ANALYSIS CLASSIFICATION METHODS</b>	
6.1 MACHINE LEARNING APPROACH	33
6.1.1 NAÏVE BAYES	33
6.1.2 BAYESIAN NETWORK	33
6.1.3 MAXIMUM ENTROPY	33
6.1.4 SVM	34
6.1.5 DECESION TREE	34
6.2 LEXICON BASED APPROACH	34
6.2.1 DICTIONARY APPROACH	34
6.2.2 CORPUS BASED	34
6.2.2.1 STATISCAL APPROACH	34
6.2.2.2 SEMANTIC APPROACH	34
6.3 MEHTODS FOR HINDI LANGUAGE	



6.3.1	BILINGUAL DICTIONARY	35
6.3.2	MACHINE TRANSLATION	35
6.3.3	WORDNET LINKING	35
6.3.4	HINDI SENTI-WORDNET	35
<b>CHAPTER 7: IMPLEMENTATION</b>		<b>36</b>
7.1	DATA COLLECTION	36
7.2	PRE-PROCESSING	37
7.3	SENTIMENT ANALYSIS	38
7.4	IMPLEMENTATION FOR ENGLISH TWEET	38
7.4.1	TEXTBLOB	38
7.4.2	SENT-WORDNET	39
7.5	IMPLEMENTATION FOR HINDI TWEET	40
7.5.1	HSWN	40
7.5.2	MACHINE TRANSALTION	41
<b>CHAPTER 8: RESULT</b>		<b>43</b>
<b>CHAPTER 9: CONCLUSION</b>		<b>46</b>
9.1	CONCLUSION	46
9.2	FUTURE WORK	47
<b>CHAPTER 10: REFERENCES</b>		<b>48</b>

## LIST OF FIGURES

<b>FIGURE NO.</b>	<b>FIGURE DESCRIPTION</b>	<b>PAGE NO.</b>
<b>Figure 1</b>	Types of people reactions	13
<b>Figure 2</b>	Process of sentiment analysis.	13
<b>Figure 3</b>	Challenges in sentiment analysis	16
<b>Figure 4</b>	Users of twitter (country wise).	22
<b>Figure 5</b>	Twitter functionalities.	23
<b>Figure 6</b>	Fetching tweets using API	28
<b>Figure 7</b>	Data pre-processing steps.	29
<b>Figure 8</b>	Cleaning Steps.	29
<b>Figure 9</b>	Sentiment analysis paradigm	32
<b>Figure 10</b>	Method Classification	32
<b>Figure 11</b>	Steps for tweet analysis	36
<b>Figure 12</b>	Yogi Adityanath using Textblob	38
<b>Figure 13</b>	Akhilesh Yadav using Textblob	39
<b>Figure 14</b>	Adityanath using SWN	40
<b>Figure 15</b>	Akhilesh Yadav using SWN	40
<b>Figure 16</b>	Adityanath using HSWN	40
<b>Figure 17</b>	Adityanath using Machine Translation	41
<b>Figure 18</b>	Akhilesh Yadav using Machine Translation	42
<b>Figure 19</b>	Positive words for Adityanath	43
<b>Figure 20</b>	Negative words for Adityanath	43
<b>Figure 21</b>	Positive words for Akhilesh Yadav	44
<b>Figure 22</b>	Negative words for Akhilesh Yadav	44
<b>Figure 23</b>	Scatter plot for Yogi Adityanath	47
<b>Figure 24</b>	Scatter plot for Akhilesh Yadav	48

## LIST OF TABLES

TABLE NO.	TABLE DESCRIPTION	PAGE NO.
<b>Table 1</b>	Other notable research	21
<b>Table 2</b>	Users of twitter (year wise.	22
<b>Table 3</b>	Tokenization	25
<b>Table 4</b>	Stemming and Lemmatization	26
<b>Table 5</b>	POS Tagging	27
<b>Table 6</b>	Data cleaning algorithm	30
<b>Table 7</b>	Example Dataset	31
<b>Table 8</b>	Sample of dataset	37
<b>Table 9</b>	Data Cleaning code	37
<b>Table 10</b>	Tokenization	38
<b>Table 11</b>	POS tagging example	39
<b>Table 12</b>	POS tagging in dataset	39
<b>Table 13</b>	picture of final dataset with senti score	41
<b>Table 14</b>	Translation code Hindi to English	41
<b>Table 15</b>	Translated dataset	42
<b>Table 16</b>	Result for English tweets	45
<b>Table 7</b>	Result for Hindi tweets	45

# Chapter 1

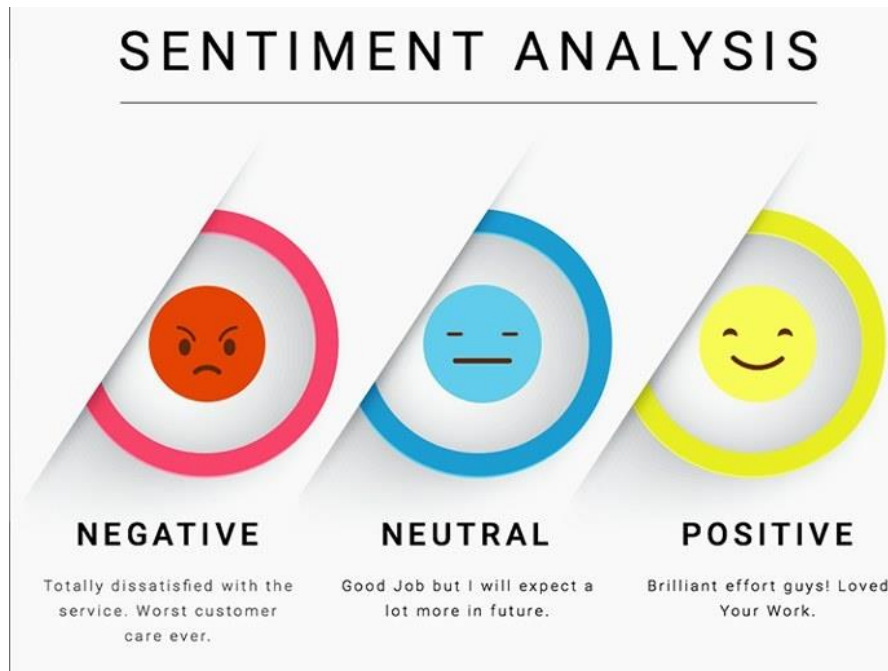
## Introduction

---

This is a new era of technology which connects people to each other no matter how far they are. This credit goes to social media. Social media is the platform of sharing and receiving information, data, as well as communication system of people. They share their psychology, thinking, ideas, behaviours and sentiments. It is very powerful weapon of increasing literature and business. People use social media to gain education and power for a better life and health. Every single day in the world millions and billions amount of data is generated but only a small fraction of these data are structured and well organized. Consider 80 percent as unstructured and the rest are structured. There are many ways by which data is generated but the most common and widely used is social media. People are nowadays so much indulged in these platforms. People use social networking sites to communicate and express their joy, rage, grief, and any other emotion. They not only use it as a communication tool, but they also use it to document their daily lives. In another way, social media has evolved into a repository of feelings. As a result, dealing with it is a time consuming task. As there are a large amount of opinion data are available, new theme such as sentiment analysis and opinion mining is getting popular [1].

Twitter is the most popular among all social media for opinion sharing. Twitter messages are generally short status updates from Twitter app users limited to 140 characters in length. Every day, Twitter generates millions of brief messages and user status updates on hundreds of different subjects. Extracting data from these messages has shown to be quite effective in categorizing and rating the popularity of subjects.[2]-[3] The technique of extracting meaning from vast amounts of data is known as text mining or text analytics. Sentiment analysis and opinion mining are two terms that can be used interchangeably. Sentiment analysis recognizes and analyses the emotion portrayed in a text. Opinion mining extracts and analyses people's opinions on a topic, SA seeks strong viewpoints, recognizes the feelings they express and then determines their polarity.

Majority of the work on sentiment analysis is done in English. Our project is a venture into Hindi sentiment analysis. With approximately 200 million speakers, most of the people in northern India speaks Hindi in their daily life for communication and expressing their expressions. Hindi is more order free and has many words that can be used in many ways as compared to English. As a result, scarcity of resources creates problems at every phase of the process, including data collecting and production. The number of users and web content for Hindi language has been increasing tremendously. This part of the web hasn't been explored much in the direction of sentiment and opinion analysis.



*Figure 1: Types of people reactions*

## 1.1 WHAT IS SENTIMENT ANALYSIS

Sentiment analysis (also known as opinion mining) is the technique of identifying and extracting subjective information from source materials using natural language processing, text analysis, and computational linguistics. Sentiment analysis is commonly used in reviews and social media for a number of purposes, including marketing and customer service.



*Figure 2: Process of sentiment analysis*

Sentiment Analysis is used to evaluate a writer's sentiments and views in a single subject area or a multi-topic domain. It determines the overall sentiment polarity of online actual reviews for a certain topic using sentiment categorization levels such as positive and negative. Existing sentiment analysis techniques may be divided into four groups: word level, sentence level, document level, and aspect/entity level.

Several websites have recently sprung up to allow academics to express and communicate their thoughts, ideas, and opinions on scientific studies. Sentiment analysis seeks to determine a writer's attitude toward certain themes or the general polarity of a text's sentiment, such as positive or negative and neutral. The polarity of emotion and the sentiment score are two factors that influence sentiment analysis. The positive or negative polarity of a sentiment is a binary value. Sentiment score, on the other hand, is based on one of three mathematical models. Bag-of-words (BOW), part-of-speech (POS), and semantic links are the three models. The BOW model, based on the representation of words, is the most common among scholars. In a Bag-of-Words paradigm, the phrase refers to words. It disregards grammatical rules and the arrangement of words in a sentence. The POS tagging model tags verbs, adjectives, and adverbs grammatically..

## **1.2 LEVELS OF STUDY IN SA**

Sentiment analysis can be categorized according to the granularity of text on three level:

### **1.2.1 DOCUMENT LEVEL**

This level's analysis is used to assess whether a document's overall sentiment is good or bad. For instance, given product reviews, the algorithm might assess the overall sentiment polarity [4]. Document level analysis presupposes that a piece of text conveys a single target's mood. While this is frequently true for product reviews, movie reviews, restaurant reviews, and other similar scenarios, it is unlikely to be true in cases when a document critiques numerous targets.

### **1.2.2 SENTENCE LEVEL**

The analysis of this level is to determine whether the opinions expressed in a sentence is positive, negative or neutral. Sentence level analysis can be conducted in two ways. One way is to simply regard such analysis as a 3-way classification task, where the labels are positive, negative and neutral. The second way is to first detect subjectivity in the sentence to split opinionated texts from those un-opinionated texts, then classify those subjective texts with one of two labels(positive or negative). The challenge of sentence level analysis is that each individual sentence is semantically and syntactically connected with other parts of the text. Therefore, this task requires both local and global contextual information. Yang [5] analyzes product review on sentence level and addresses this challenge successfully.

### **1.2.3 ASPECT LEVEL**

Unlike document or sentence level analysis, aspect level analysis explores what the holder likes or hates about the target. The tasks of such fine-grained analysis is three folded [6]: (1) extracting features of target, (2) determining feature-wise polarity, (3) summarizing the

overall evaluation. Aspect level sentiment analysis is one the most challenging tasks compared to other level of analysis.

Twitter sentiment analysis falls into document level. However, since Twitter allows a maximum of 140 characters, each tweet status tends to be very short. Usually a tweet contains only one simple sentence or just several words. Therefore, twitter sentiment analysis also calls for a wide variety of strategies utilized on other levels of analysis.

### 1.3 APPLICATIONS OF SENTIMENT ANALYSIS

#### ***A. In product reviews***

In this digital era, people like to buy products online to save time and resources. So a customer review and feedback for products help to improve or do anything required to make the product more useful.

#### ***B. In Business***

The area of sentiment analysis is ideally suited for a wide range of commercial and intelligence applications. One of the most important aspects of the corporate world is business intelligence. The job of extracting opinions from unstructured human materials is a fantastic commercial task.

#### ***C. In government intelligence***

Governments might poll the public before or after implementing a policy to assess its efficacy and acceptability. To get the mood and reaction of the public on a specific topic or policy government uses sentiment analysis.[7].

#### ***D. In other domains***

SA may be used in a variety of situations, including product evaluations, stock market news[8,9] stories[10], and political arguments[11]. For example, we may learn about people's views on specific election candidates or political parties in political discussions. Political positions may also be used to forecast election outcomes.

### 1.4 CHALLENGES IN SENTIMENT ANALYSIS

#### 1.4.1 FOR ENGLISH LANGUAGE

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

***Unstructured Data-*** The data on the internet is largely unstructured; there are many distinct types of data discussing the same entities, people, places, things, and events. Data from many sources, including as books, journals, web articles, health records, company logs, internal files of an organisation, and even data from multimedia platforms, such as texts, photos, audios, and videos, is available on the web. The complexity of the analysis is increased by the fact that the data comes in a variety of forms.

***Noise (slangs, abbreviations)-*** The web content available is very noisy. In today's era of 140 characters texting, for their ease people use various abbreviations, slangs, emoticons in normal text which makes the analysis more complex and difficult.

**Example-** *mvie ws awsummm :D*

The web content reports a large number of spelling variations for the same word. Eg a word **awesome** can be found in various forms as- “*awsum, awssuummm, awesome*” the repetition of the characters can be in any combination.

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

Ex-1 The movie was long.

Ex-2 Lecture was long.

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

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

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

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

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

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

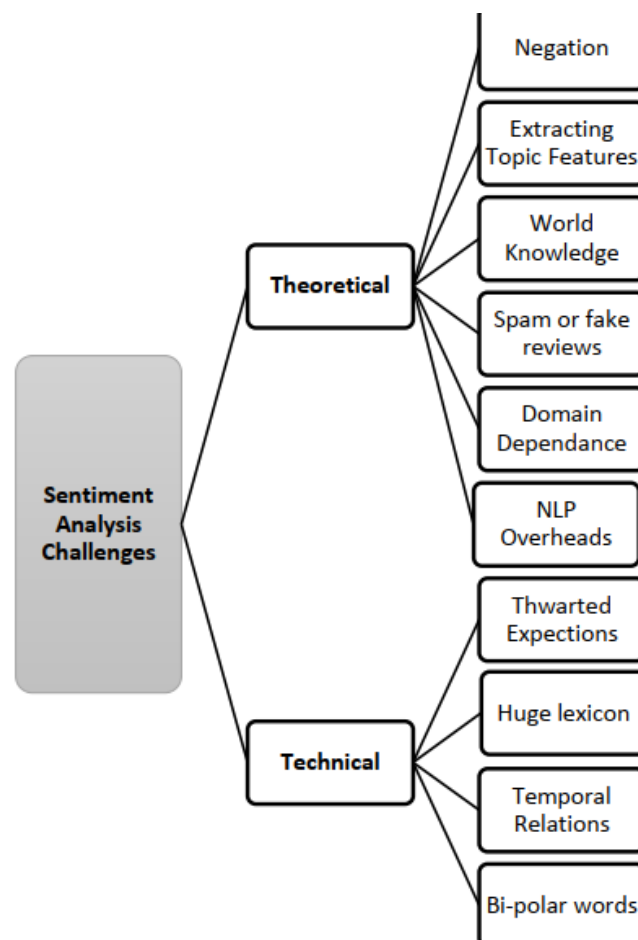


Figure 3: Challenges in sentiment analysis



## 1.4.2 FOR HINDI LANGUAGE

Each language has its own nature and style of writing which is accompanied by its own challenges and specifications. Challenges while dealing/working with Hindi language are as follows-

**Word Order-** The order of words in a phrase is critical in determining the text's subjective quality. Hindi is a free-order language, meaning that the subject, object, and verb can all appear in any order, whereas English is a fixed-order language, meaning that the subject comes first, followed by the verb, and then the object.

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

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

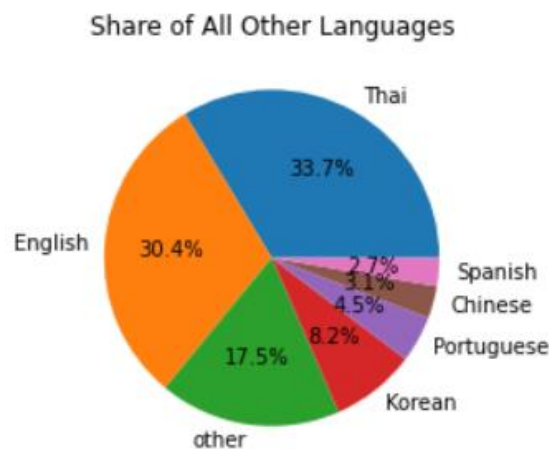
## Chapter 2

# Scope and Objective of the Study

---

### 2.1 SCOPE

1. Sentiment Analysis is used to evaluate a writer's sentiments and views in a single subject area or a multi-topic domain. It determines the overall sentiment polarity of online actual reviews for a certain topic using sentiment categorization levels such as positive and negative.
2. This is a new era of technology which connects people to each other no matter how far they are. This credit goes to social media. Social media is the platform of sharing and receiving information, data, as well as communication system of people. They share their psychology, thinking, ideas, behaviours and sentiments.
3. Most of the sentiment analysis that has been done till now is mostly for English language. The dataset that used in most of the research is Movie Review site, Blogging site and product reviews etc.
4. Twitter is the top social media platform where people share their opinions.



### 2.2 RESEARCH GAP

Lot work has been done on sentiment analysis topic, but most of them are for English language and dataset that has been used is movie review, product review, blogging sites and for Hindi language blogging site. We will do sentiment analysis for both Hindi tweets and English tweets. Indian elections 2021 tweets will be taken as dataset. As performance measured for Hindi language sentiment analysis is not so much we will try to increase accuracy by combining and trying different algorithms and approaches.

## **2.3 RESEARCH TOPIC INTRODUCTION**

Assembly election 2022 in Uttar Pradesh is slated from 10th February to 7<sup>th</sup> March and will be done in seven phases. The counting of the votes starts on 10<sup>th</sup> March and the result will be declared on that same day. Uttar Pradesh is the state which has the highest number of constituencies, 403 seats. This is the most prestigious state for any government to hold. In 2017 Bhartiya Janta Party(BJP) came into power and selected Yogi Adityanath as chief minister. The main opponent of the BJP in UP is Samajwadi Party (SP), Akhilesh Yadav is the face of the SP, he has been chief minister from 2012-to 2017.

Yogi Adityanath's face has become the most visible in the state. According to an NBT Online poll, 53.81 percent of people say Chief Minister Yogi Adityanath is the most identifiable face of the Uttar Pradesh election results. Meanwhile, 31.17 percent of respondents still consider PM Modi to be the most important vote-getter for the BJP. Meanwhile, 15.02 percent of people feel they have no notion who would be the BJP's vote-getter between Modi and Yogi. In this sense, public opinion appears to be shifting in favour of job creation and development.

According to the poll, the BJP would face a frontal fight from the SP. On the other side, the BSP and the Congress are not viewed as serious contenders in this election. According to a poll conducted by NBT Online, 67.07 percent of respondents believe the SP is the BJP's main opponent. Exit polls, on the other hand, have been proven wrong in Bengal and Bihar Assembly elections. Because so many things happen between election stages and the public's decision takes several twists and turns, the true conclusion can only be seen in the results. But one thing is certain: the real fight is between the SP and the BJP.

## Chapter 3

# Review of Literature

---

### 3.1 LITERATURE RIVIEW

Although sentiment analysis and opinion mining became one of the most important sources in decision making in business still several challenges need further attention. In the following, we discuss the related work with sentiment analysis and challenges:

Farzindar [12] separates sentiment analysis and emotion analysis to emphasize the subtle difference. Emotion analysis is more meticulously classified into finer granularity. In [13], emotion is categorized into six class: anger, disgust, fear, joy, sadness, surprise, which are most widely used in the literature. There is currently no consensus on how many classes of emotions should be used. Emotion analysis is also referred to as mood detection.

Kasper & Vela proposed a “Web Based Opinion Mining system” for hotel reviews [14]. The paper introduced an evaluation system for online user’s reviews and comments to support quality controls in hotel management system. It is capable of detecting and retrieving reviews on the web and deals with German reviews. It has multi-topic domain and is based on multi-polarity classification; the system could recognize the neutral e.g., “don’t know” to “classify sentiment polarity that as neutral” and the multi-topic cases identified in their corpus.

Mobile devices products reviews were analyzed by (Zhang, et-al) in [14]. This research can help in evaluate accuracy. It is useful in a judgment of the product quality and status in the market [14]. This research used three machine learning algorithms (Naïve Base Classifier, K-nearest neighbor, and random forest) to calculate the sentiments accuracy. The random forest improves the performance of the classifier.

Hemalatha *et al.* (2012)[15] shows a very nice approach for pre-processing Twitter data following simple steps, and demonstrates how to prepare the data for training in machine learning technique. This approach eliminates unnecessary noise such as slang, abbreviation, URL, special characters from the linguistic data and also reduces the size of data set by eliminating noise. Extending the work in other literature; Hemalatha and her colleagues derived a combined pre-processing and classification approach executed parallel to achieve high performance, reduced data size and produced more accurate results by classifying the features from the sentiment words by adding polarity of it, and applied machine learning techniques to the derived data set. Bandgar and Kumar (2015)[16] using their research methodology illustrated how to create a windows application for real-time Twitter data for pre-processing of text data using available natural language processing resources like WordNet, SMS dictionary, Stanford dictionary

There are many papers about election prediction such as (KokilJaidka, Saifuddin Ahmed, el, 2018)[17] the election prediction of three different countries India, Pakistan and Malaysia. The accuracy of results is awesome. They only shows volumetric performance, Supervised and unsupervised model, And Show the result on histogram graph chart and expression but not gave an open result that an average people can understand.

Some research paper compares two or more than two parties like USA (Alexandre Bovet, et, 2016) Trump versus Clinton, they have large scale of twitter data 0.73 million and gave good results and prediction but results are reversed, Clinton being more popular than Trump. They cannot show number of tweets for each candidate and also gave line graph which does not show number of tweets.

The research in sentiment analysis trend is not limited yet. In order to improve accuracy and performance of the proposed techniques, applications, or algorithms. It enables them to be more compatible with understanding meaning and features. But still there are some problems and challenges in text analysis of reviews/documents and evaluate sentiment scores.

### 3.2 REVIEW TABLE

Reference	Year	Description	Technology/Outcome
[1]	2014	Discussed all types of sentiment analysis technique	It is stated that SVM has higher accuracy
[2]	2014	All the methods, pros, and cons of sentiment analysis techniques	Survey Paper
[3]	2018	prediction of US presidential election 2016	Twython, Textblob, NLTK
[4]	2019	Prediction of general elections of 2019 is done using decision tree classifier and accuracy was 97%	Decision tree, Text blob library used
[5]	2020	Polarity of tweets was calculated for the general election of India of 2019	R programming, lexicon based approach
[6]	2015	SA of Hindi language and discussion on unsupervised lexicon method for classification	Build a corpus of Hindi tweets using the Lexicon approach and Twitter archiver.
[7]	2018	The data from Twitter relating to the 2017 Punjab assembly elections are used in this study.	SVM, DT, and KNN are used. Proposed an algorithm for seat forecasting.
[8]	2016	Several challenges are facing the sentiment analysis and evaluation process	The purpose of this study is to survey sentiment analysis problems that are important to their methodologies and techniques.

[9]	2019	Forecast the President and Vice President of Indonesia's elections from 2019 to 2024.	Support Vector Machine with Particle Swarm Optimization and Genetic Algorithms selection characteristics
[10]	2016	In this different classification, algorithm accuracy is compared.	An extensive assessment of six supervised classification methods is presented in this work.
[11]	2020	Data collection from social media over time, as well as similarity recognition based on similar decisions made by people in social networks	The methods for assessing emotions have been investigated, classified, and contrasted.

*Table 1: Other notable research*

### 3.3 LITERATURE REIVIEW OF HINDI LANGUAGE

In the area of opinion mining and sentiment analysis, very little and countable research has been done for the Hindi language. Das and Bandopadhyya created a lexicon for the Bengali language using an English-Bengali dictionary. They came up with 35,805 words. Dipankar Das and Bandopadhyya[19] performed sentiment analysis for Bengali. They classified words in six different sorts of emotions (aggression, hatred, fear, joyful, sadness, and astonishment), as well as three different types of classes (high, medium, and low).

Hindi-SentiWordNet was produced by Joshi, Balamurali, and Bhattacharyya[20] utilizing two prebuilt lexical tools, English SentiWordNet and English-Hindi WordNet. They used wordnet to match English terms with their Hindi meanings to create a Hindi lexicon.

Piyush Arora[21] suggested a graph-based technique for constructing a subjective lexicon for Hindi based on Wordnet. They began using a small set of seed words and expanded on them using wordnet, synonyms, and antonyms. Each word in the seed list is considered a node, with synonyms and antonyms linked to it. They were able to classify evaluations with 74 percent accuracy.

For predicting sentiment, Namita Mittal [22] proposed effective. Accuracy of Hindisentiwordnet increases by adding additional opinion words in HSWN. They developed guidelines for dealing with negation and speech that had an impact on sentiment prediction. Their suggested method obtained an accuracy of 80%.

Rao and Ravichandran [23] published a thorough investigation on the difficulty of recognising sentiment polarity in words. It is known as bi-polar sentiment classification of words, which means that a term can have either a positive or negative connotation. They used semi-supervised label propagation in a graph to detect word sentiment polarity. Each of these words represents a graph node whose polarity must be determined. They primarily worked on three languages: English, French, and Hindi, but they insist that their work be expanded to

any other language for which WordNet exists. There has been a lot of effort put into building the subjective lexicon for English languages. [24]

### 3.4 TWITTER

Twitter is one of the most popular social media platforms on the planet. People update hundreds of activities, thoughts, and on any issue on Twitter every second of the day, making it a treasure trove of sentiments from all over the world. It is known as one of the largest psychological databases, which is constantly updated and allows us to evaluate millions of data using machine learning. Twitter has a strong presence in social media networks. Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams established Twitter in March 2006. (Way back Machine, 2012). Twitter has 336 million active users, with more than 100 million of them posting every day, totaling more than 500 million posts with a limit of 280 characters.

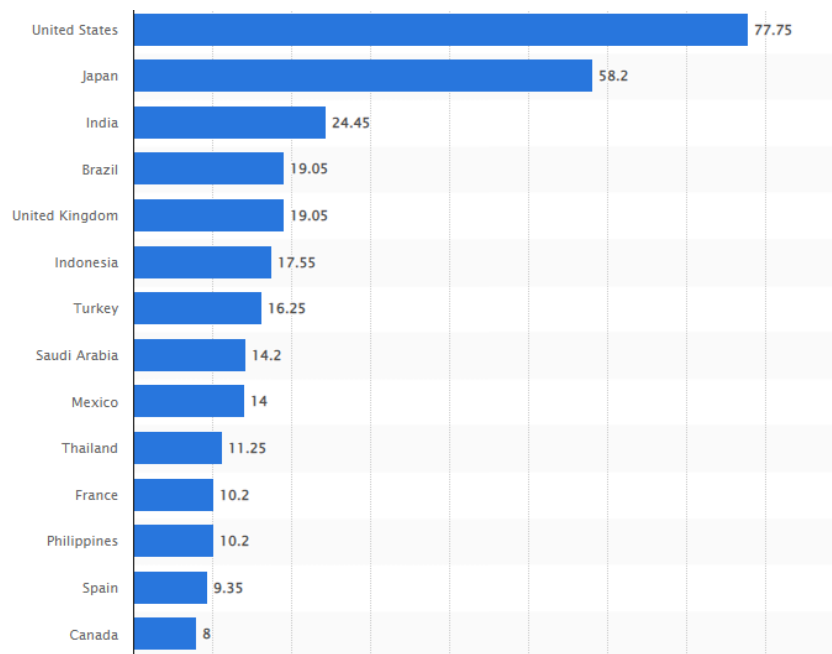


Figure 4: Users of twitter (country wise)

Year	Users
2017	110 million
2018	112 million
2019	139 million
2020	186 million

Table 2: Users of twitter (year wise)

Twitter has released the most powerful API for developers, which is ranked in the top ten APIs worldwide. There are two types of Twitter accounts: one for regular users and the other for developers (using API). Normal users can share and read information (tweets), but developer accounts can utilise the API to access Twitter data (Application programme interface). Data can be collected from developer accounts using keys issued by Twitter. Consumer keys, consumer secret keys, token keys, and token secret keys are the four types of keys. These keys are one-of-a-kind and are utilised in a variety of programming languages to capture twitter data. Twitter is a major commercial and advertising centre as well. Twitter's "SMS" feature employs authentication for account security. Twitter is an open source platform as well.

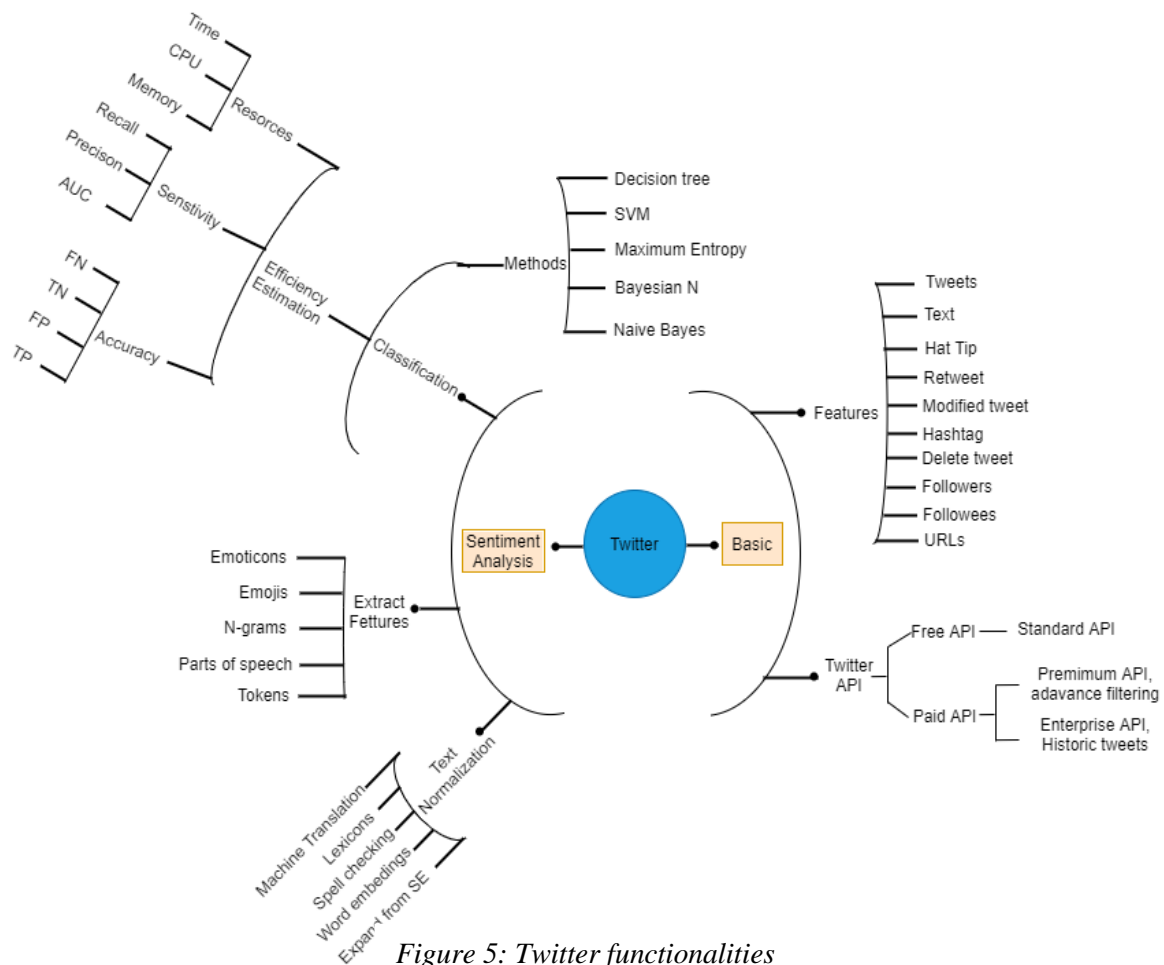


Figure 5: Twitter functionalities

Twitter terminology since we are going to have to deal with in the tweets:

- **Hashtag:** Any word or phrase preceded by the # sign is referred to as a hashtag. You'll see other Tweets with the same term or topic if you click on a hashtag.
- **@username:** On Twitter, you're identifiable by your username, which is always preceded by the @ sign. Selena Gomez, for example, has the handle @selenagomez.
- **Retweet:** A Retweet is a Tweet that you forward to your followers. Retweets are often used on Twitter to share news or other significant findings, and they always keep the original attribution.
- **Emoticons:** Composed using punctuation and letters, they are used to express emotions concisely, ";) :) ...".



## Chapter 4

# Research Terminologies

---

### NATURAL LANGUAGE PROCESSING

It is a branch of computer science, artificial intelligence, and computational linguistics concerned with computer-human (natural) language interactions. NLP is fundamentally linked to the field of human–computer interaction. Natural language understanding, or the ability for computers to extract meaning from human or natural language input, is one of the many issues in NLP [25]. Another is natural language generation. Natural Language Processing is a broad phrase that refers to a collection of techniques for automating the production, manipulation, and analysis of natural or human languages. Despite the fact that most NLP approaches are based on linguistics and AI, they are also influenced by disciplines such as Machine Learning, Computational Statistics, and Cognitive Science.

#### 4.1 NLTK

Toolkit for natural language It's also known as NLTK. It's a collection of libraries that assist Python English-written programmes, including symbolic and statistical natural language processing. It contains over 50 corpora and lexical resources, such as WordNet, combined with language processing libraries for text document analysis, such as word tokenization, categorization and stemming, tagging, parsing, and semantic rules. Graphical demos and sample data are included in NLTK.

- **Token:** Before any real processing can be done on the input text, it needs to be segmented into linguistic units such as words, punctuation, and numbers or alphanumeric. These units are recognized as tokens.
- **Sentence:** This refers to an ordered sequence of tokens.
- **Tokenization:** The operation of splitting a sentence into its constitutive tokens.

<b>Input:</b> Filtered Tweets
<b>Output:</b> Tokenize words
<b>For all words in Processed Tweets</b> <i>Tokenize the word passing to Tweet Tokenizer Method and append Tokenize Sentence</i> <b>Return Tokenize Sentence</b>

Table 3: Tokenization

- **Corpus:** This means a body of text, usually including a large number of sentences.

#### 4.2 STEMMING

The procedures of stemming and lemmatizing words provide the normalised version of a word in the text[26]. Word stemming is a method for determining the root (base) of a word in a text. It normalises the term by deleting the suffix, which gives the word its basic meaning.

### 4.3 WORD LEMMENTIZATION

Word lemmatization is another significant natural language processing job. It is a method of converting a word's structural form to its base or dictionary form by filtering the affixation or replacing the vowel in the word. The result of the word is referred to as a lemma[27]. The lemmatized word is the entrance to the WordNet, and lemmas are the words that have a root meaning of the term requested. As a result, lemmatizing the word with an algorithm creates a lemma, which is then passed to the WordNet dictionary, which extracts the word's meaning and its sense number, which is the goal of gaining a better sentiment score for the word[28].

<b>I/p:</b> Tokenized tweets <b>O/p:</b> Stemmized text
<b>For every tokenized word</b> <b>Initialize</b> empty string 'Stem_Sen_List' <i>If word.length() &gt; 2</i> <i>call fuction for stemming the word by creating object of PorterStemmer.</i> <i>call function for Lemmatizing the word by creating object of WordNetLemmatizer.</i> <i>Add into Stem_Sen_list</i> <b>Return Stem_Sen_List</b>

Table 4: Stemming and Lemmatization

### 4.4 TEXTBLOB LIBRARY

Textblob is a Python module for handling textual data. Textblob provides an API that allows you to access its functions and conduct NLP tasks with ease. The key rationale for using Textblob is because it functions similarly to a Python string, making it simple to use without having to worry about syntax. Textblob includes features such as part of speech, noun phrase, sentiment analysis, tokenization, word inflection and lemmatization, wordlist, spelling correction, translation or language recognition, and N-gram detection..

### 4.5 TWEETPY

This is the most crucial section of our thesis. We can't gather twitter postings (tweets) from the Twitter API without this module. This is an open-source library that connects to Twitter through an API. Tweepy works with Twitter's authentication credentials. Keys for consumer, consumer secret, token, and token secret are all unique to each user or API. We pull data from Twitter on various topics using these keys.

### 4.6 VADER

The VADER Emotion Lexicon [29] is a comprehensive collection of "gold-standard" sentiment terms that is especially useful for text data from microblogs and other social networks. VADER is validated by human specialists and provides both polarity and intensity. It includes information on emoticons, slang (such as "nah" and "meh"), and acronyms in addition to regular dictionary terms (\LOL", \LMAO" etc.)

### 4.7 PARTS OF SPEECH TAGGING

This technique analyses the sentence structure and annotates the part-of-speech (e.g. Noun, Adverb, Adjective, Subjects, Objects) to the words, resulting in the raw form of word

meaning disambiguation. It is the final stage in natural language processing for determining the emotion of a statement [30]. This stage allows you to get featured words, which indicate the sentence structure and meaning of the words in the domain they belong to in the phrase.

Example: POS Tagging	Text Data
Word Tokens	['Mamata', 'Banarjee', 'going', 'to', 'win', 'the', 'Bengal', 'elections', 'in', '2021']
POS(Part-of Speech)-Tagged sentence	[('Mamata', 'NNP'), ('Banarjee', 'NNP'), ('going', 'VB'), ('to', 'IN'), ('win', 'VB'), ('the', 'DT'), ('Bengal', 'JJ'), ('elections', 'NNS'), ('in', 'IN'), ('2021', 'CD')]

Table 5: POS tagging

## 4.8 WORDNET

With 58058 words and four parts of speech tags, WordNet is a huge lexical database in English. Cognitive synonyms (synsets) are a collection of nouns, verbs, adjectives, and adverbs that individually communicate a separate notion. It is one of the most widely used and well-behaved tools for processing natural language, and it includes emotional words as well as "semantic links" between words [31]. Synsets or Synonyms set or group of synonyms are the connectivity of semantic and lexical relationships for words and their meaning.

Dictionary (Lexicon) based technique, which automatically builds the dictionary of the words and their relationships in suitable size, is often utilised by WordNet type of lexical databases.

## 4.9 SENTI-WORDNET

SentiWordNet is a resource that consists of opinion information for the word derived from the WordNet database, where each term is assigned a numerical score that contains the sentiment value for the word as well as information about the word. Each item in this dictionary relates to a set of words belonging to the same Part-of-Speech (POS) and having the same meaning (meaning). Each category has three numerical sentiment scores that represent how good, negative, or neutral the words inside it are. These scores range from 0.0 to 1.0, and the total for each category is 1.0.

## 4.10 BAG OF WORDS

A document is represented as an unordered collection of words, disregarding grammar and even word order, in this model [32], which is the most often used in supervised text categorization. According to this approach, a dictionary is generated based on training data during the training phase and then utilised to distinguish between positive and negative articles in the testing step.

# Chapter 5

## Research Methodology

---

### 5.1 PYTHON

Python is one of the most fast growing programming language in terms of number of developers. Developers mostly use python because it is one of the platform for easiest and fast coding and compilation. Python has huge number of libraries (scientific computing and data sciences) and many big companies use python, such as Google, Yahoo, YouTube, Dropbox and NASA. Python also supports machine learning, GUI, software developing and web developing. Python is general purpose, interpreter, object oriented and high level language. Python is also multi-paradigms programming language like functional, imperative, object oriented and reflective language. Python consists of different syntax and semantics such as Indentation, Statement and control flow, Expression, Methods, Typing and mathematics.

### 5.2 DATA PRE-PROCESSING

Twitter data made available to conduct this research is in semi-structured data set. The data set contains 'text' field where the user generated tweets are used for research, which may consist of noise as well as partial and unreliable linguistic data. Hence, in order to analyze linguistic data from Twitter, it is necessary that this irregular data be cleared and removed, so the true meaning and sentiments can be accounted for from the data. This is where data preprocessing comes into play. To filter and remove the noise from the data, the algorithm implemented using Python Programming language and all the preprocessing tasks for filtering the noise from the data are discussed in the following

We use twitter API to fetch tweets

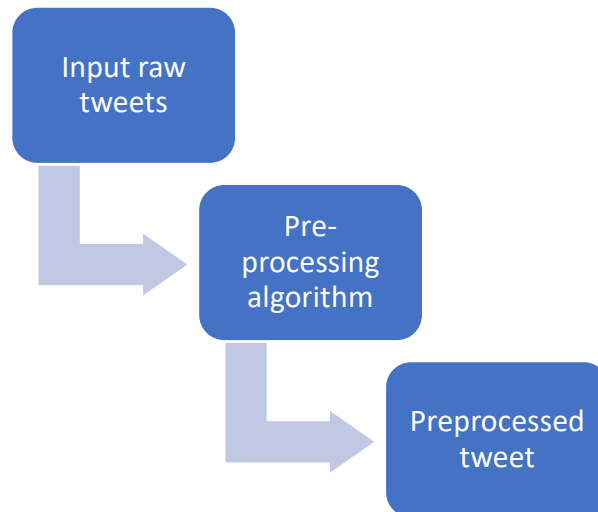
```
[ ] api_key="weuYSR9uH3TJLwXkeRyI5X1cV"
    api_secret="IrDAZBnDvw4GEswMGLALB71R0aXFK0Yiju3D5IfVbJqQJSbCSs"
    access_token="1402677717603028992-18ctdbUs0Gu4eXXpxw4DhyG1COuVn3"
    access_token_secret="5bDGJPwWlQNgmlzy7piYk5zinTq0BAVMtSoDxdwLwff5B"
    auth= tw.OAuthHandler(api_key,api_secret)
    auth.set_access_token(access_token,access_token_secret)

▶ api = tw.API(auth, wait_on_rate_limit=True)
  search_text="mamata banerjee"
  date_since="2021-02-01"
  tweets = tw.Cursor(api.search, q=search_text, lang="en", since=date_since).items(300)
  tweets
```

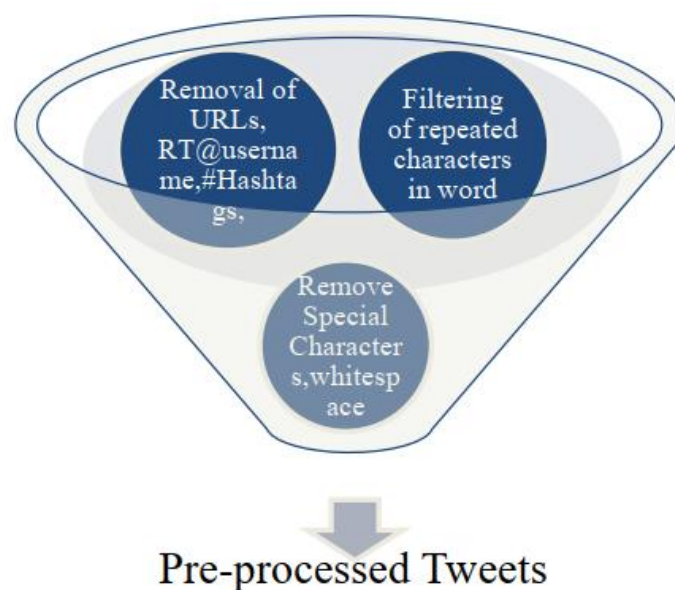
Figure 6: Fetching tweets using API

Tweets are in unstructured form of data and are full of noise and unwanted information. This textual data is full of unwanted text, special and repeated characters, and may contain unwanted space in it. Therefore, in order to perform sentiment analysis on this data set, the preliminary step is to pre-process this data and transform it so that the machine learning algorithm analysis can be performed to it.

Hence, in order to properly analyze this data from tweets, it is necessary that this irregular data is cleared and removed, so the true meaning and sentiments can be accounted from the sentence. Preprocessing of data normalizes the linguistic data, eliminates noise and simplifies the vocabulary used for analyzing sentiments from the tweets [33].



*Figure 7: Data pre-processing steps*



*Figure 8: Cleaning data*

There are various steps to be performed in order to reassess the data correctly and determine the true meaning behind it, through data processing. It is also necessary to follow proper sequence to pre-process data to achieve accuracy as well as consistency in data set.

**I/p:** Noisy and raw twitter data  
**O/p:** Filtered and noise free data for the next step of Natural Language Pre-processing Task.

**For every tweets in File**

*To store the output we initialize the new empty string name- filtered\_tweet*

- 1. Using regular expression approaches, replace any URLs or https:// links with the term 'URL' and save the result in filtered\_tweet.*
- 2. Replace all instances of "@username" with "AT\_USER" and save the result in filtered\_tweet.*
- 3. Remove #Hashtags and ReTweet from the tweet and save the result in filtered\_tweet.*
- 4. If there is repetetions of characters in a word then replace with the character itself, and save the result in filtered\_tweet.*
- 5. Delete all special characters (: \ / [ ] ; : { } - + ( ) < > ? ! @ # % \*,) from the statements.*
- 6. Delete the word 'URL' and AT\_USER and save the result in filtered\_tweet.*

**Return filtered\_weet.**

*Table 6: Data cleaning algorithm*

In the first step the algorithm clear out all the URLs present in the tweets. This step of preprocessing will eliminate all the 'URLs' from the Dataset and will result in reducing the noise as well as decreasing the size of dataset. However, the output generated will remain with the meaningful information in the tweet. Also, in the developed algorithm 'www. &' and 'https: //' is converted to the word 'URL' using regular expression function available in Python. This can be imported using regular expression (.re) module in Python, which gives programmer an embedded functionality inside Python language to operate textual or string data set.

The second step to perform preprocessing is to remove '@username' from the tweet. '@username' is the tag with '@' followed by the user id or the user name on Twitter platform. The information can be found with '@username' tag, and retweets have been abbreviated as '\RT'. Retweet is the process when any user re-posts the comment on others account, which describes the reaction of the user behavior to that particular post.

After eliminating URLs, retweet and username, and #Hashtag from the text data it becomes more meaningful and each word gives us some meaning. Again, this was only some basic steps to be performed to pre-process Twitter text data for analysis which not only removes noise

from the data, but also, reduces the size of the dataset as well as increases performance

[ ]	geo	text	user	Subject
0	None	"This is YOUR VICTORY!"Our leader, Mamata Ban...	Chandan33027631	mamata banerjee
1	None	"This is YOUR VICTORY!"Our leader, Mamata Ban...	SuvayanPal	mamata banerjee
2	None	Consolidation of Congress ( slyly and quietly...	singh_uvach	mamata banerjee
3	None	What's the role of Mamata Banerjee or TMC's F...	_SamratDas_	mamata banerjee
4	None	West Bengal Chief Minister Mamata Banerjee to...	MonaSyed017	mamata banerjee
5	None	Mamata Banerjee To Visit Delhi, Meet Oppositi...	Javed8877	mamata banerjee
6	None	Consolidation of Congress ( slyly and quietly)...	RohitashwT	mamata banerjee
7	None	"Congress doesn't look serious but Goans are ...	DidiforGoa	mamata banerjee
8	None	What's the role of Mamata Banerjee or TMC's F...	A_Worker4	mamata banerjee
9	None	West Bengal Chief Minister Mamata Banerjee to...	ManzarA81671181	mamata banerjee
10	None	"Congress doesn't look serious but Goans are ...	IMRANAL54690675	mamata banerjee
11	None	West Bengal CM Mamata Banerjee set to visit De...	sameepshastri	mamata banerjee
12	None	West Bengal CM Mamata Banerjee set to visit De...	ChhattisgarhThe	mamata banerjee
13	None	"Congress doesn't look serious but Goans are ...	m1a52sabina	mamata banerjee
14	None	"I want to defeat BJP in 2022 elections; Cong...	iamonlyabhik	mamata banerjee
15	None	#MamataBanerjee set to visit #Delhi from Nov	1967Anil	mamata banerjee

Table 7: Example dataset (fetched using API)

# Chapter 6

## Sentiment Analysis Classification Methods

Machine Learning, Lexicon Based, and Hybrid Approach methods are the three primary kinds of sentiment categorization methodologies. The Machine Learning Approach (ML) applies well-known ML techniques to language characteristics. The Lexicon-based method drives the opinion lexicon, which is nothing more than a collection of pre-compiled opinion words. It is primarily split into corpus based and dictionary based approaches, both of which employ semantic and statistical methodologies. The hybrid approach combines the two approaches mentioned above [34].

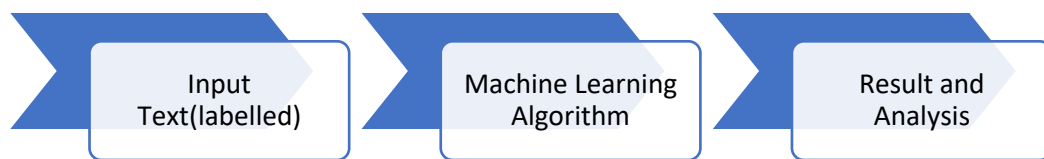


Figure 9: Sentiment analysis paradigm

Text categorization approaches based on machine learning may be split into supervised and unsupervised learning methods. In the supervised procedures, a large number of labeled training papers are utilized. Unsupervised techniques are used when locating these labeled training resources is difficult.

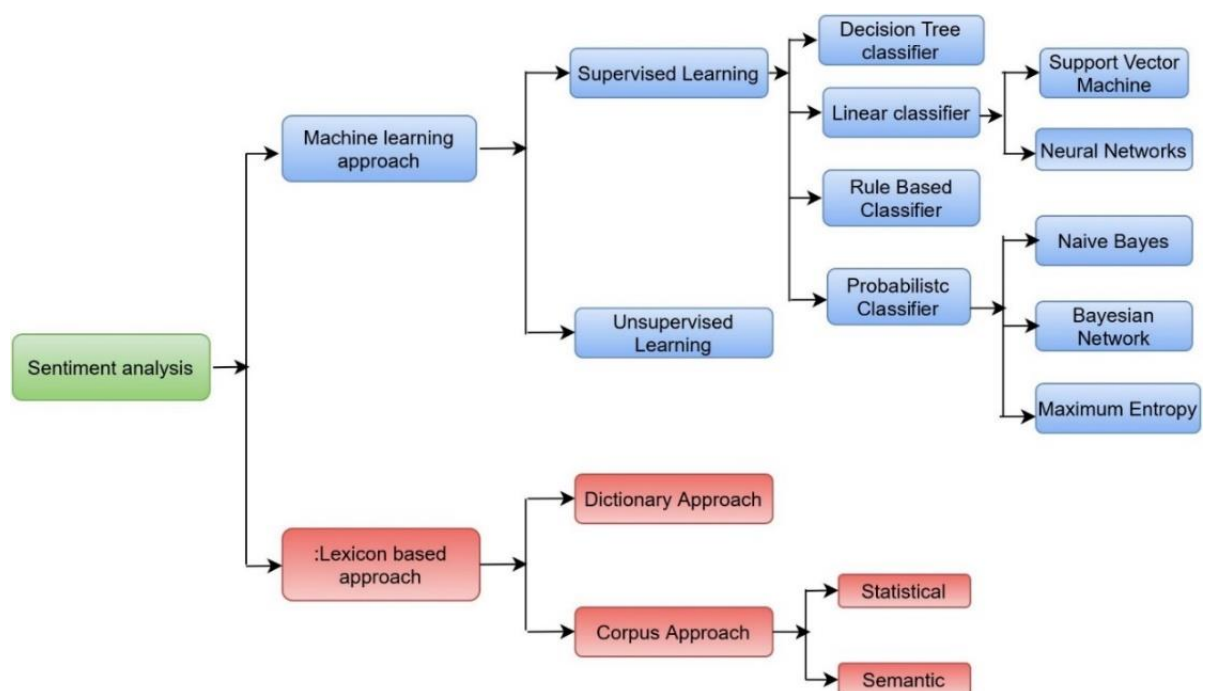


Figure 10: Method Classification



The lexicon-based method is predicated on locating an opinion lexicon that can be used to evaluate the text. This strategy employs two approaches. The dictionary-based method finds opinion seed words and then searches a lexicon for synonyms and antonyms. A seed list of opinion words is used to start the corpus-based technique. Following that, it searches a huge corpus for further opinion keywords to help in the discovery of opinion words with regard to context-specific orientations.

## 6.1 MACHINE LEARNING APPROACH

To solve the SA, the machine learning technique uses well-known ML algorithms. Each record has a class assigned to it. The classification model is linked to one of the class labels by characteristics in the underlying data. The model is then used to predict a class label for a given instance of an unknown class.

### 6.1.1 SUPERVISED LEARNING

In the supervised approach, we use labelled data for the training set. All the data have their predefined class. there are several different types of supervised classifiers.

A. **Naive Bayes classifier-** The simplest and most often used classifier is naive bayes. The probability of a class is calculated using the Naive Bayes method. For feature extraction, the model use Bag of Words. The Bayes Theorem is used to forecast the probability of the feature set corresponds to a specific label. The main advantage of this classifier it its simplicity as well as prediction of the correct class for a new instance [35]. It simply multiple all feature values which has been extracted from each instance in the class (for e.g. POSITIVE, NEGATIVE and NEUTRAL) are classified as a class. Every labeled sentiment tags (instance) contributes for the final classification result and given equally importance with respect to each other tokens in the data set.

The Naïve Bayes classifiers assumes that every feature or the attribute of an instance (in this case ‘Tweet Sentiment’ attribute) is considered independently from all other feature in the given class. And as a result, it will multiply all the members of feature vector in given class to compute Bayesian probability.

B. **Bayesian Network-** The fundamental thinking of the Naïve Bayes classifier is that the characteristics are not related. The opposite thinking is that all the characteristics are entirely interwinding. Bayesian nets (BN) are a network-based technology primarily used to evaluate and visualize uncertain models. The causal relationships are learned via Bayesian networks, which are then used to apply incremental learning. First, configure the input nodes with evidence, then query and evaluate the output nodes using normal Bayesian network inference. Extracting information using the Bayesian method is cost-effective so it is not used mostly.

C. **Maximum Entropy-** To convert pre-defined labeled feature sets to vectors, the Maximum Classifier uses encoding. The feature weights generated from this encoded vector may then be used to provide the most likely label for a feature set.

D. **Support Vector Machine-** SVMs are based on the notion of finding linear dividers in the solution space that can appropriately distinguish different classes. Due to the fragmented nature of language, few elements are unimportant, but they are likely to be connected and organized into linearly separable groups. SVM may be used to solve classification and

regression challenges. Classification predicts a label/group, whereas regression predicts a continuous value.

- E. **Decision Tree classifier**- A decision tree classifier splits the data depending on an attribute value condition, it hierarchically decomposes the training data set. The data space is recursively partitioned until the leaf nodes have a specified number of records to classify.

## 6.2 LEXICON BASED APPROACH

Lexicons are semantically meaningful words, phrases, idioms, and adjectives; this approach is based on lexicons to categorize data. It uses a scoring mechanism to determine emotion polarity by rating each phrase according to negative and positive terms. There are three basic approaches for extracting and collecting an opinion words list.

- A. **Dictionary based approach**- In this, we manually select some words whose orientation is known. Then using Wordnet we look for antonym and synonyms of original words. We continue to do this task recursively until there are no new words left. This method has one disadvantage: it is impossible to detect opinion words with specific domain and context orientations.
- B. **Corpus based approach**- We create a corpus containing words that are connected to any context using this method. This strategy is less successful than dictionary-based methods since huge corpora are difficult to create, but it can help in the discovery of domain and context-specific opinion words and their sources. Statistical or semantic methodologies are used in the corpus-based method.

**A) Statistical approach**- Statistical methods can be used to find frequency trends or main opinion words. If a term appears more frequently in positive literature, its polarity is positive. If it appears more frequently in negative material, it has a negative polarity. If the frequency of the terms is equal, it is a neutral word. When two words come together often in the same context, their polarities are likely to be the same. As a consequence, estimating the polarity of an unknown word may be done by calculating the relative frequency of co-occurrence with another word.

**B) Semantic Approach**- The Semantic method provides direct sentiment class and uses various concepts to compute word similarity. This concept assigns semantically comparable phrases similar to emotion ratings. For example, WordNet provides several types of semantic connections between words that may be utilized to compute sentiment polarity.

## 6.3 SENTIMENT ANALYSIS METHODS FOR HINDI

Hindi is an Indian language with a population of 53 million speakers. (Indian census, 2011; 43 crores in 2001). Hindi is becoming more popular. Today, Hindi and other Indian languages are supported by many websites and social media platforms. A few years ago, Twitter offered Hindi as a tweeting language. To build a model that does sentiment analysis for the Hindi language, we have three methods

- A. **Using Bilingual Dictionary**- This method involves utilizing bi-lingual dictionaries to map the polarity of words. We need a connecting or mapping between the words of the two languages and a subjective lexicon or a dictionary of the polarity of terms for one of the languages.

- B. **Using Machine translation**- there is a lack of corpus and wordnet resources for the Hindi language. So we use the machine translation technique. In this technique, we use translation technology to transform the subjective lexicon from source to target language. The Hindi documents are passed through a translation system to translate it into English, and then an English language trained classifier classifies the text. For translating subjective lexicon from one language to another, translation technologies such as Google Translation, Language-specific translation systems, and others are utilized. In this, we assume that the sentiment of the sentence will not be changed after translation.
- C. **Wordnet linking**- The WordNet linking map the synset of English words to any other language. It outputs the lexicon capturing the relations and similarity between the words of the wordnet.

## 6.4 Hindi Senti-WordNet

This is one of the most significant and widely used Hindi language resources. The Hindi Wordnet was created in 2002 [36] and contains terms relating to various elements of speech such as adjectives, adverbs, nouns, and verbs that are related to other words via syntactic and semantic relationships. It gives information on how words are linked to one another, words belonging to comparable systems, and synonyms, antonyms, and other information for a specific word.

*Example - For word अच्छा*

*Synonyms are बढ़िया, भला etc*

*Antonyms are बुरा, खराब, अनुचित etc*

We can always develop our own resources. IIT Bombay develop HindiSentiWordnet. They used SentiWordnet for creating Hindi resources. As sentiment scores are associated with English words can be projected to the corresponding synset in the Hindi language. It has different synsets, for example- Adjectives, adverbs, Noun verbs. It contains words with part of speech as Adjective, Adverb, Noun and Verb with each word having 3 scores positive, negative and objective. Sum of positive, negative and objective score sums to 1. The lexicon assigns single score to a word irrespective of the sense in which it is used

## Chapter 7

# Implementation

The proposed methodology consists of the 3 phases-

- 1) Data collection: It is performed 3 times in a week during a fixed time.
- 2) Data Pre-processing: Clearing tweets
- 3) Sentiment analysis

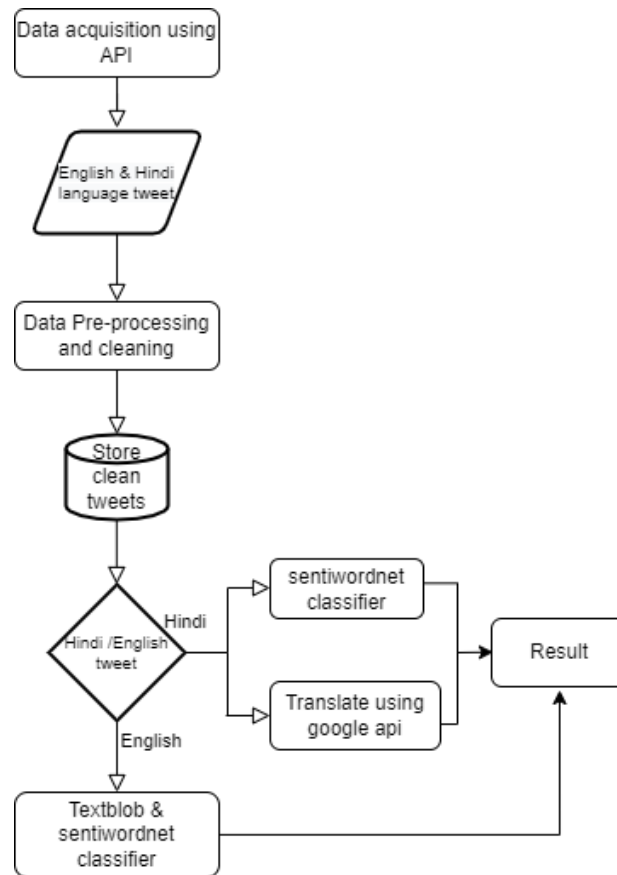


Figure 11: Steps for tweet analysis

### 7.1 Data Collection

Tweets are extracted from the Twitter database using Twitter developer API which works with the tweepy library. Tweepy provides connectivity to the Twitter database and saves tweets into .csv format file. Monday, Wednesday, and Saturday from 9 am tweets are fetched. Tweets are searched and collected which have hashtags like #UPElection2022, and we searched tweets having the desired name such as Yogi Adityanath, Akhilesh Yadav. We collected the tweets in both English and Hindi languages. Total of 60,000 tweets are fetched, in which 30 thousand are in English and 30 thousand are in the Hindi language. In 30,000 tweets fifteen thousand tweets are related to one candidate and 15 thousand are related to the other candidate and the same case with Hindi language tweets.

	text	user	Subject
1	That's Enough Reason For Hindus Why We Suppor...	mohit42261248	yogi
2	That's Enough Reason For Hindus Why We Suppor...	shashan43460125	yogi
3	Samajwadi Party should be ready for another e...	BhagwatiParikh	yogi
4	GorakhpurAyecarry SaturdayMotivation😊Today th...	MukeshD77892696	yogi
5	rarely watch political stuff but yday watched...	LavaKm	yogi
6	BSP chief today attacked UP CM Yogi Adityanat...	htlucknow	yogi
7	Totally gobsmacked by what Boyapati and Balay...	agaddam	yogi
8	Totally gobsmacked by what Boyapati and Balay...	ckgaddam	yogi
9	That's Enough Reason For Hindus Why We Suppor...	Magu21885917	yogi
10	Received blessings at Akshardham in NewJersey...	UmeshGo38619224	yogi

Table 8: Sample of dataset

## 7.2 Pre-processing

Tweets that are collected contains lots of extra characters, words, a symbol such as emojis, hashtag, retweet, username, weblink these are not useful in opinion mining. So to clean the dataset and extract more insight from them we clean the data and remove its noise. It also reduces the size of the dataset and increases output accuracy. Following is the pseudo algorithm for data pre-processing.

```

import re
def clean_tweets(text):
    text = re.sub("RT @[\w]*:", "", text)
    text = re.sub("@[\w]*", "", text)
    text = re.sub("https?://[A-Za-z0-9./]*", "", text)
    text = re.sub("\n", "", text)
    text = re.sub("https?:\\/\S+", "", text)
    text = re.sub("#", "", text)
    return text
data['text'] = data['text'].apply(lambda x: clean_tweets(x))
def rem_stopwords_tokenize(data, name):

    def getting(sen):
        example_sent = sen

        filtered_sentence = []

        stop_words = set(stopwords.words('english'))

        word_tokens = word_tokenize(example_sent)

        filtered_sentence = [w for w in word_tokens if not w in stop_words]

        return filtered_sentence
    # Using "getting(sen)" function to append edited sentence to data
    x=[]
    for i in data[name].values:
        x.append(getting(i))
    data[name]=x

```

Table 9: Data Cleaning code

After removing link, stop words, extra space, username name, now we have clean dataset. Now we tokenize each sentence that is each sentence is splitted into its consecutive words. As shown in below figure.

```
[ ] token_tweet=tweet_df['text'].apply(lambda x:x.split())
    token_tweet.head()

0    [That's, Enough, Reason, For, Hindus, Why, We,...
1    [That's, Enough, Reason, For, Hindus, Why, We,...
2    [Samajwadi, Party, should, be, ready, for, ano...
3    [GorakhpurAyecarry, SaturdayMotivation🤔Today, ...
4    [rarely, watch, political, stuff, but, yday, w...
Name: text, dtype: object
```

Table 10: Tokenization

### 7.3 Sentiment Analysis

In this experiment, we have used two methodologies for English tweets and two methodologies for Hindi tweets. Former experiment done using senti-wordnet and textblob and latter Hindisentiwordnet and machine translation. Below all the methods are described.

### 7.4 IMPLEMENTATION FOR ENGLISH TWEETS

#### 7.4.1 Textblob

Textblob is a Python module for handling textual data. Textblob provides an API that allows you to access its functions and conduct NLP tasks with ease. The key rationale for using Textblob is that it functions similarly to a Python string, making it simple to use without having to worry about syntax. Textblob includes features such as part of speech, noun phrase, sentiment analysis, tokenization, word inflection, lemmatization, wordlist, spelling correction, translation or language recognition, and N-gram detection.

When we parsed our dataset into textblob code. We have got below pie chart for Yogi Adityanath and Akhilesh Yadav respectively. Here we can see that Adityanath got more positive tweets than Akhilesh Yadav.

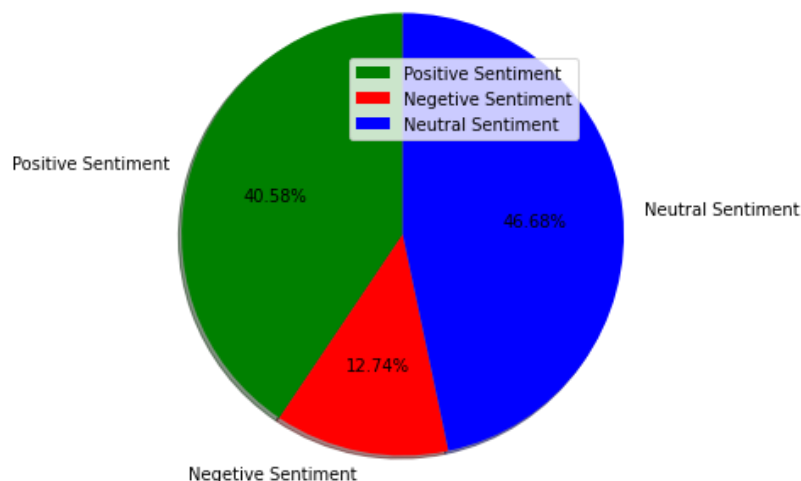


Figure 12: Yogi Adityanath using Textblob

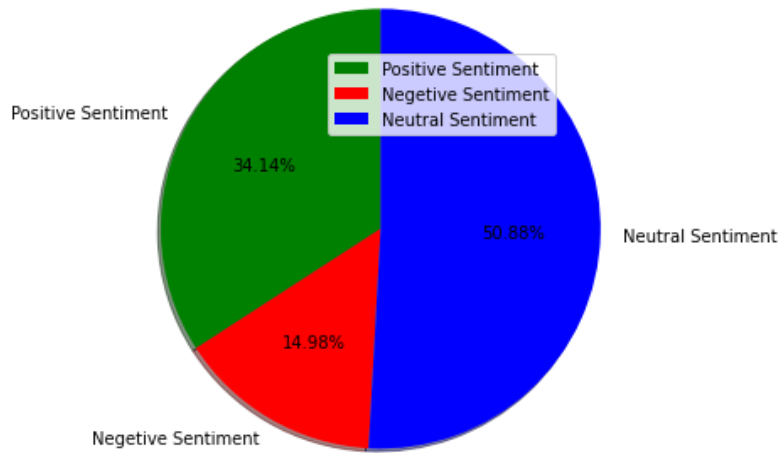


Figure 13: Akhilesh Yadav using Textblob

#### 7.4.2 Sentiwordnet

SentiWordNet is a database of user opinions obtained from the WordNet database, where each term is given a numerical score that includes the sentiment value for the word as well as information about the word. Each item in this dictionary relates to a set of words belonging to the same Part-of-Speech (POS) and having the same meaning (meaning). Each category has three numerical sentiment scores that represent how good, negative, or neutral the words inside it are. These scores range from 0.0 to 1.0, and the total for each category is 1.0

Example: POS Tagging	Text Data
Word Tokens	['Akhilesh', 'Yadav', 'going', 'to', 'win', 'the', 'UP', 'elections', 'in', '2022']
POS(Part-of Speech)-Tagged sentence	[('Akhilesh', 'NNP'), ('Yadav', 'NNP'), ('going', 'VB'), ('to', 'IN'), ('win', 'VB'), ('the', 'DT'), ('UP', 'JJ'), ('elections', 'NNS'), ('in', 'IN'), ('2022', 'CD')]

Table 11: POS tagging example

11614	mother National President Shri Akhilesh Yadav ...		
11615	Youth want serve country wearing uniform But r...		
11616	Mr. Akhilesh Yadav lead role Shri Krishna batt...		
11617	Akhilesh Yadav Ayodhya n't see Lord Shri Ram		
11618	Received blessing Mahant Shri Naga Maharaj Shr...		
		pos_tags	senti_score
0	[(Akhilesh, NNP), (Yadav, NNP), (contest, NN), ...		0.250
1	[(assembly, RB), (election, NN), (Akhilesh, NN...		-0.625
2	[(Exclusive, JJ), (chief, NN), (Akhilesh, NNP)...		-1.375
3	[(difference, NN), (akhilesh, JJ), (yadav, NN)...		-0.375
4	[(Akhilesh, NNP), (Yadav, NNP), (pride, NN), (...		0.500
...	...		...
11614	[(mother, JJ), (National, NNP), (President, NN...		-0.250
11615	[(Youth, NNP), (want, VBP), (serve, NN), (coun...		0.250
11616	[(Mr., NNP), (Akhilesh, NNP), (Yadav, NNP), (l...		0.250
11617	[(Akhilesh, NNP), (Yadav, NNP), (Ayodhya, NNP)...		0.000
11618	[(Received, VBN), (blessing, NN), (Mahant, NNP...		0.000
[11619 rows x 5 columns]>			

Table 12: POS tagging in dataset

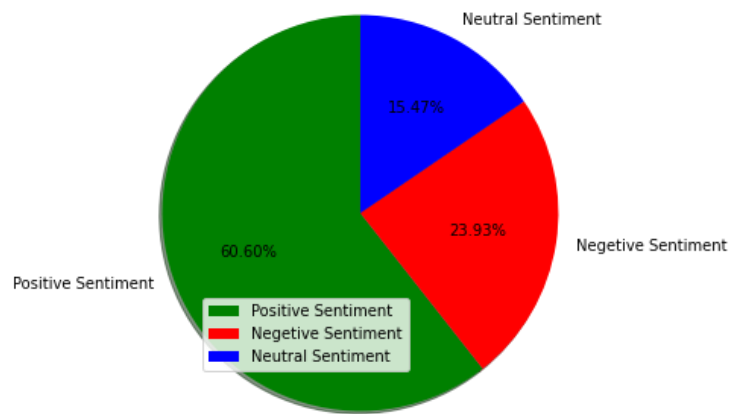


Figure 14: Adityanath using SWN

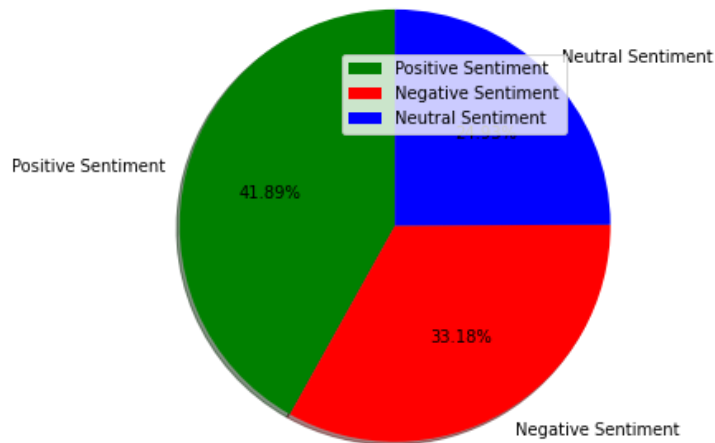


Figure 15: Akhilesh Yadav using SWN

## 7.5 IMPLEMENTATION FOR HINDI TWEETS

### 7.5.1 Hindi-Sentiwordnet(HSWN)

This is one of the most significant and widely used Hindi language resources. The Hindi Wordnet was created in 2002 and contains terms relating to various elements of speech such as adjectives, adverbs, nouns, and verbs that are related to other words via syntactic and semantic relationships. When we pass both dataset after cleaning them, then first it tokenize the words and it searches into a text file which consist words with its corresponding senti score. It apply senti score to each token and sum them for each sentence.

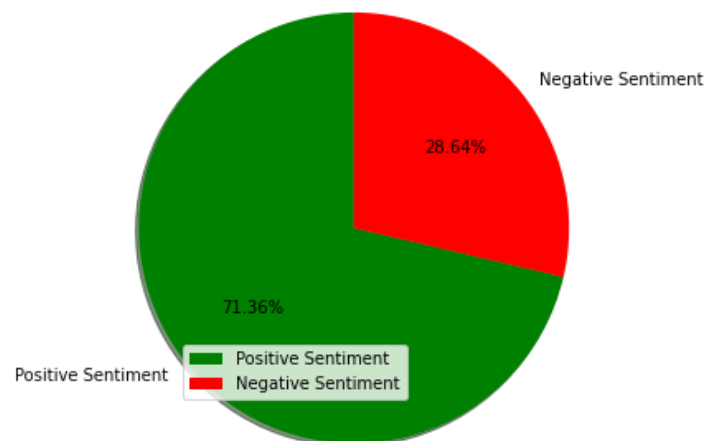


Figure 16: Adityanath using HSWN



	text	Tweet in English	After_lemmatization	pos_tags	senti_score	Overall Sentiment
0	अखिलेश यादव मैनपुरी के करहल विधानसभा से चुनाव ...	Akhilesh Yadav contest Karhal assembly constit...	Akhilesh Yadav contest Karhal assembly constit...	[(Akhilesh, NNP), (Yadav, NNP), (contest, NN),...]	0.250	Positive
1	यूपी विधानसभा चुनाव : अखिलेश यादव ने की टीवी च...	UP assembly election : Akhilesh Yadav demanded...	assembly election Akhilesh Yadav demanded ban ...	[(assembly, RB), (election, NN), (Akhilesh, NN)...]	-0.625	Negative
2	Exclusive: सपा प्रमुख अखिलेश यादव () ने महंगाई...	Exclusive : SP chief Akhilesh Yadav blames BJP...	Exclusive chief Akhilesh Yadav blame BJP infla...	[(Exclusive, JJ), (chief, NN), (Akhilesh, NNP),...]	-1.375	Negative
3	अखिलेश यादव और योगी जी में यही अंतर है	difference akhilesh yadav yogi ji	difference akhilesh yadav yogi	[(difference, NN), (akhilesh, JJ), (yadav, NN),...]	-0.375	Negative
4	अखिलेश यादव का गुमान मिट्टी में मिल गया अखिलेश...	Akhilesh Yadav 's pride found soil Akhilesh Ya...	Akhilesh Yadav pride found soil Akhilesh Yadav...	[(Akhilesh, NNP), (Yadav, NNP), (pride, NN), (...)]	0.500	Positive
5	अखिलेश यादव और योगी जी में यही अंतर है	difference akhilesh yadav yogi ji	difference akhilesh yadav yogi	[(difference, NN), (akhilesh, JJ), (yadav, NN),...]	-0.375	Negative
6	उन्नाव की रेप पीड़िता ने अपनी मां आशा सिंह के ...	Unnao rape victim thanked Akhilesh Yadav Mayaw...	Unnao rape victim thanked Akhilesh Yadav Mayaw...	[(Unnao, NNP), (rape, NN), (victim, NN), (than...)]	0.000	Neutral
7	Exclusive: अपर्णा यादव के बीजेपी में जाने पर अख...	Exclusive : On Aparna Yadav joining BJP , Akhi...	Exclusive Aparna Yadav joining BJP Akhilesh sa...	[(Exclusive, JJ), (Aparna, NNP), (Yadav, NNP),...]	0.125	Positive
8	श्री अखिलेश यादव जी समाजवादी पार्टी	Shri Akhilesh Yadav ji , release	Shri Akhilesh Yadav release list	[(Shri, NNP), (Akhilesh, ...)]	0.125	Positive

Table 13: picture of final dataset with senti score

## 7.5.2 Machine Translation

There is a lack of corpus and wordnet resources for the Hindi language. So we use the machine translation technique. In this technique, we use translation technology to transform the subjective lexicon from source to target language. Google translator API is used for translation. The Hindi documents are passed through a translation system to translate it into English, and then an English language trained classifier classifies the text. In this, we assume that the sentiment of the sentence will not be changed after translation. This is the code for translation-

```

def conv_eng(df):

    data['Tweet in English'] = ""
    for i in range(0,200):
        count = 0
        try:
            new_text = translator.translate(data["text"][i], des='en').text
            data["Tweet in English"][i] = new_text
        except:
            count = count + 1
    print("No of exceptions : ", count)

```

Table 14: Translation code Hindi to English

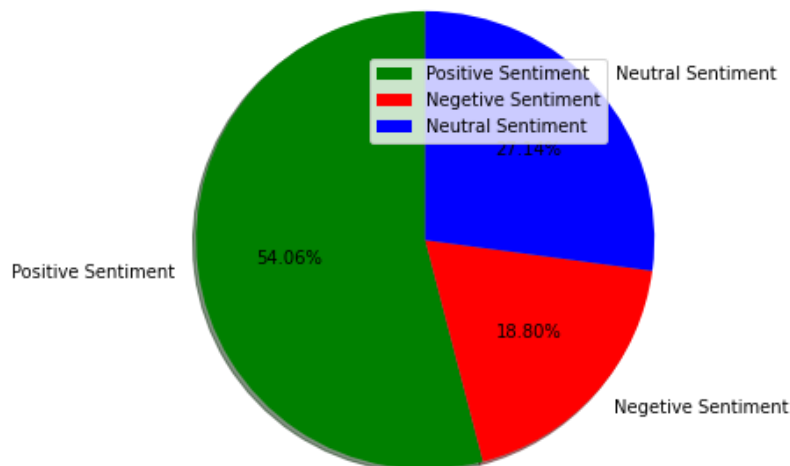


Figure 18: Adityanath using Machine Translation

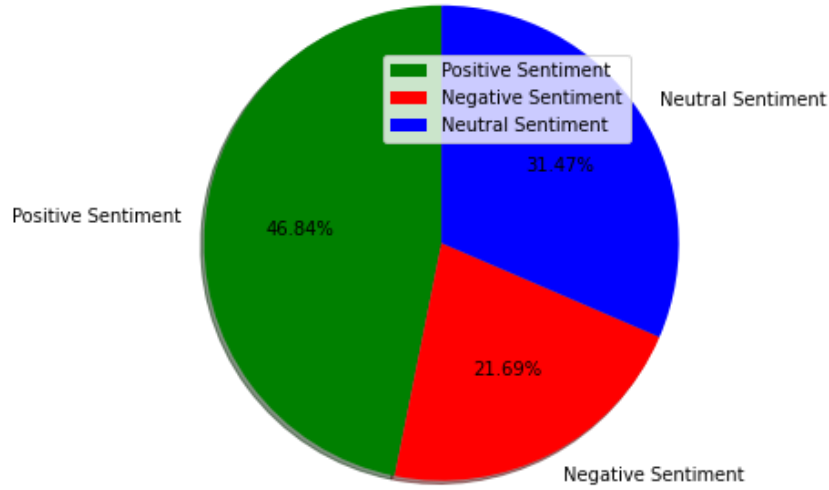


Figure 19: Akhilesh Yadav using Machine Translation

As we can see here, in HSWN method we got high accuracy for the same dataset. For Adityanath using HSWN we got approx 71 percent and by using machine translation we got 54 percent accuracy. So it is clear that Hindi sentiwordnet gives more accuracy than machine translation.

	text	Tweet in English	After_lemmatization
0	अखिलेश यादव मैनपुरी के करहल विधानसभा से चुनाव ...	Akhilesh Yadav contest Karhal assembly constit...	Akhilesh Yadav contest Karhal assembly constit...
1	यूपी_विधानसभा_चुनाव : अखिलेश यादव ने की टीवी च...	UP assembly election : Akhilesh Yadav demanded...	assembly election Akhilesh Yadav demanded ban ...
2	Exclusive: सपा प्रमुख अखिलेश यादव () ने मंहगाई...	Exclusive : SP chief Akhilesh Yadav blames BJP...	Exclusive chief Akhilesh Yadav blame BJP infla...
3	अखिलेश यादव और योगी जी में यही अंतर है	difference akhilesh yadav yogi ji	difference akhilesh yadav yogi
4	अखिलेश यादव का गुमान मिट्टी में मिल गया अखिलेश...	Akhilesh Yadav 's pride found soil Akhilesh Ya...	Akhilesh Yadav pride found soil Akhilesh Yadav...
5	अखिलेश यादव और योगी जी में यही अंतर है	difference akhilesh yadav yogi ji	difference akhilesh yadav yogi
6	उन्नाव की रेप पीड़िता ने अपनी मां आशा सिंह के ...	Unnao rape victim thanked Akhilesh Yadav Mayaw...	Unnao rape victim thanked Akhilesh Yadav Mayaw...
7	Exclusive:अपर्णा यादव के बीजेपी में जाने पर अख...	Exclusive : On Aparna Yadav joining BJP , Akhi...	Exclusive Aparna Yadav joining BJP Akhilesh sa...
8	श्री अखिलेश यादव जी समाजवादी पार्टी गठबंधन के ...	Shri Akhilesh Yadav ji , release list candidat...	Shri Akhilesh Yadav release list candidate Sam...
9	22 लाख रोजगार,300 यूनिट मुफ्त बिजली पुरानी प...	22 lakh jobs , 300 units free electricity , ol...	lakh job unit free electricity old pension sch...
10	इतिहास गवाह है जब जब किसी अहंकारी (अखिलेश यादव...	History witness time destruction egoist ( Akhi...	History witness time destruction egoist Akhile...
11	उत्तरप्रदेशरोडवेजराजकीयकरो राष्ट्रीय अध्यक्ष ,...	Uttar Pradesh Roadways Rajkikaro National Pres...	Uttar Pradesh Roadways Rajkikaro National Pres...
12	अखिलेश यादव, प्रियंका गांधी, योगी आदित्यनाथ, स...	Akhilesh Yadav , Priyanka Gandhi , Yogi Aditya...	Akhilesh Yadav Priyanka Gandhi Yogi Adityanath...

Table 15: Translated dataset

# Chapter 8

## Result

By distinguishing dataset sentences and labeling them into positive and negative. We separated the positive tweets and negative tweets. Two-word clouds are created: word cloud for positive tweets and word cloud for negative tweets to visualize that which words are attached to positive and negative sentiments in general. Word clouds are created for both Yogi Adityanath and Akhilesh Yadav.

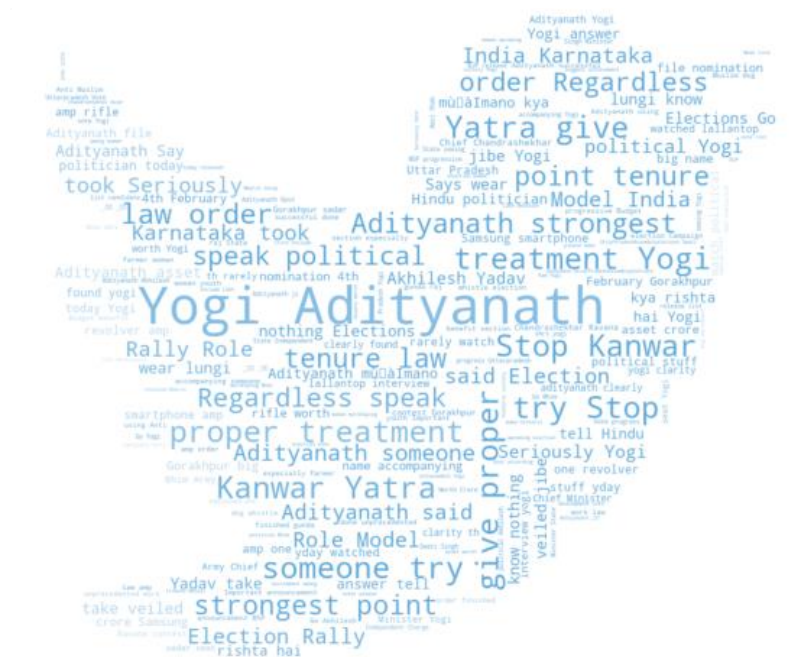


Figure 20: Positive words for Adityanath

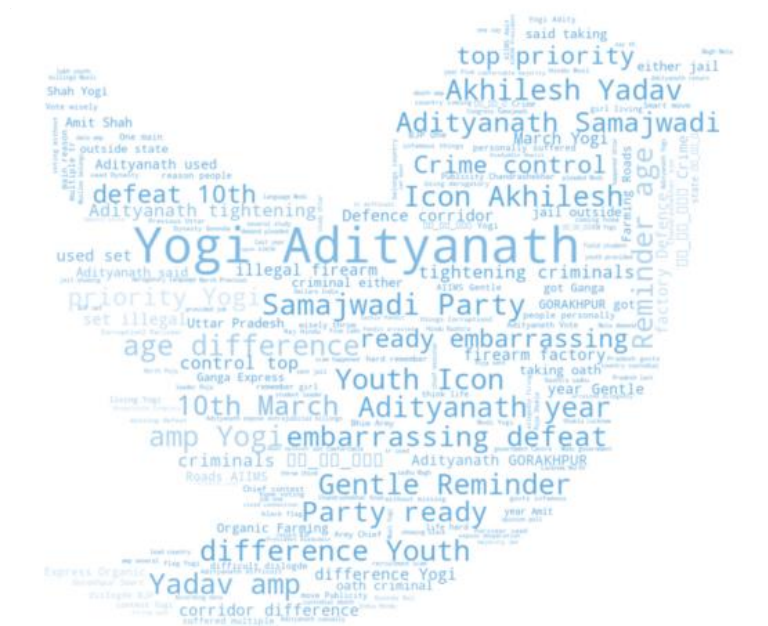


Figure 21: Negative words for Adityanath



For all the English tweets we see that percentage of positive tweets and public opinion are with Yogi Adityanath in both the methods Sentiwordnet and Textblob. wordnet gives more accurate results because it uses a corpus in which words are assigned senti-score.

Method	Yogi Adityanath		Akhilesh Yadav	
	Positive	Negative	Positive	Negative
<b>Sentiwordnet</b>	60.60	23.23	45.09	25.6
<b>Textblob</b>	40.58	12.74	34.14	14.96

Table 16: Result for English tweets

For all the Hindi tweets as the first method we used HSWN and it gave a positive score of 71.36 for Yogi and 54.78 for Akhilesh Yadav. As the second method, we first translate the Hindi tweets into English and then apply the wordnet classifier which was used for English tweets.

Method	Yogi Adityanath		Akhilesh Yadav	
	Positive	Negative	Positive	Negative
<b>HSWN</b>	71.36	28.64	65.09	34.91
<b>Machine Translation</b>	54.78	17.56	46.84	21.69

Table 17: Result for Hindi tweets

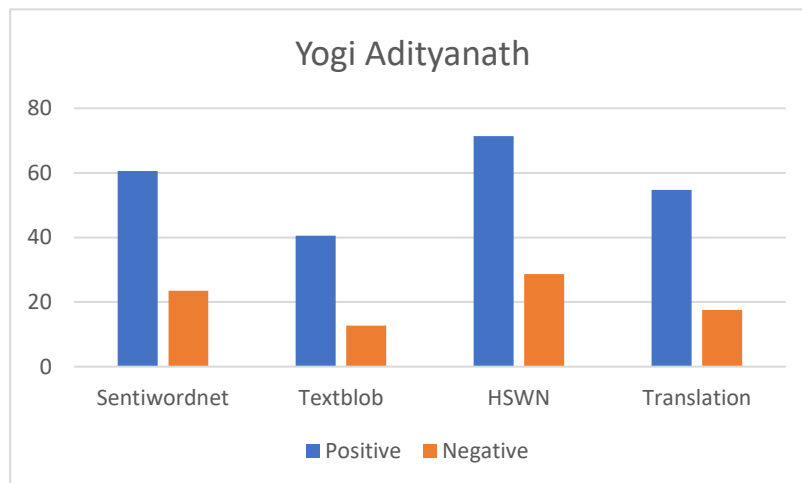


Table 18: Comparison chart English tweets

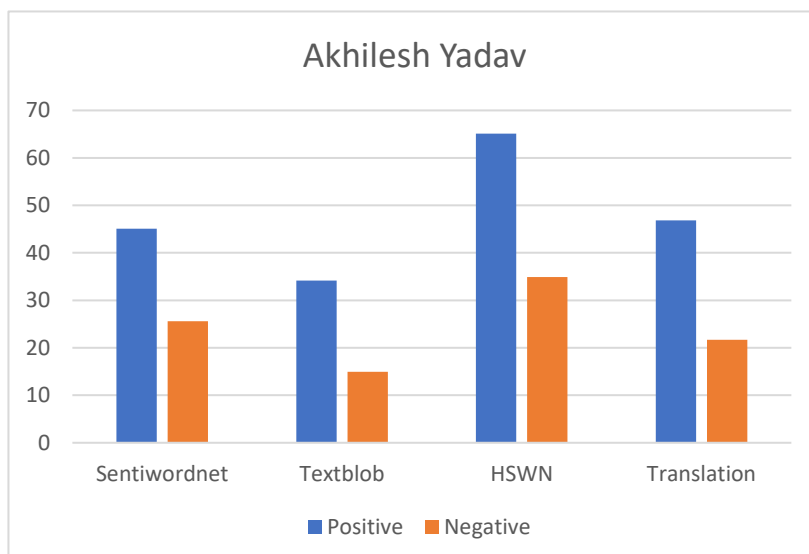


Table 19: Comparison chart Hindi tweets

## Chapter 9

# Conclusion and Future Work

---

### 9.1 CONCLUSION

This report presents and summarizes the currently used sentiment analysis and classification technique. It consists of a review paper that tells which technology and tools they have used. These papers contribute to a variety of SA-related topics that employ SA approaches in a variety of real-world applications. To filter the emojis, stop-words and other type of redundancy from the textual Twitter data, it is necessary to perform sequence of pre-processing steps. During this process the input tweets are filtered and processed to give more accurate data as well as reduce the size of dataset.

Sentiment analysis is one of the best approaches to observing and knowing public opinion and perspective. SA uses machine learning models to predict the outcome. By adopting a hybrid approach we can increase accuracy.

In the implementation result it came out that between two candidate Akhilesh Yadav and Yogi Adityanath, Yogi Adityanath have more positive tweets which mean that people have a positive perspective about him. As this research is done on 20<sup>th</sup> February and the actual result came on 10<sup>th</sup> March. So We can see that in the original vote count BJP has won 273 and SP has won 125 seats of 403 total seats.

We have collected 30000 tweets that contains hashtag #UPElection2022. When we analysed that and extracted subject Yogi Adityanath and Akhilesh Yadav. We can see in below scatter plot, we got more positive reactions for Adityanath than Akhilesh Yadav. It means that tweets which contain #UPElection2022 have more positive sentiment for Yogi's previous work (2017-2022) tenure.

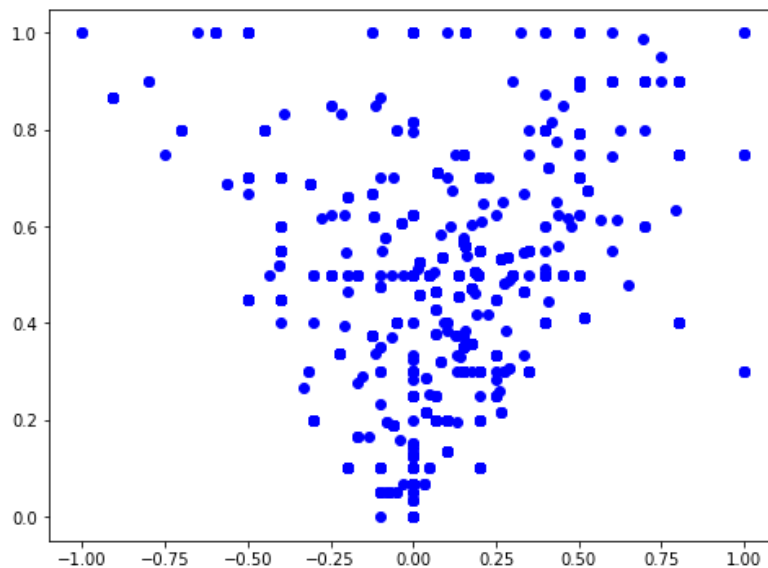
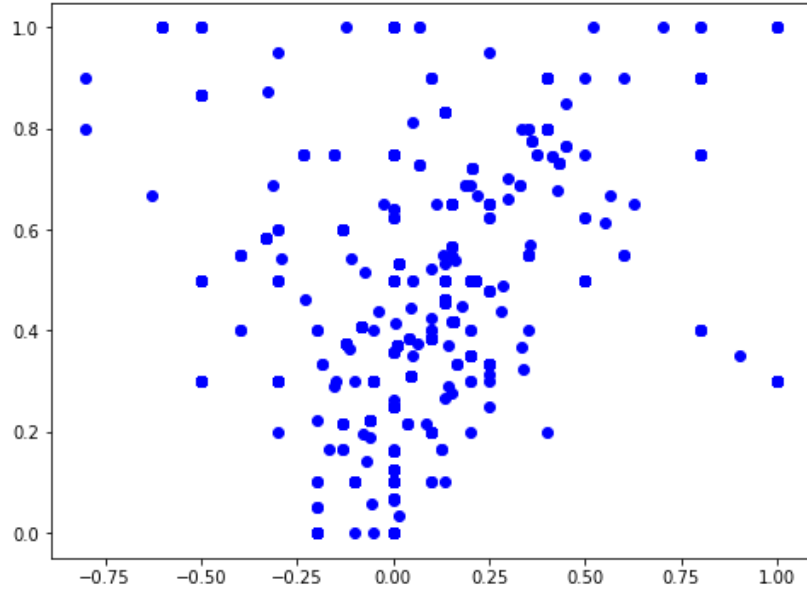


Figure 24: Scatter plot for Adityanath





*Figure 25: Scatter plot for Akhilesh Yadav*

## 9.2 FUTURE WORK

Work on detecting sarcasm, which has yet to be addressed satisfactorily, can be done in the future. To enhance accuracy, better algorithms may be devised. The lexical coverage in Hindi should be expanded, as it is currently restricted. Multilingual, image, sarcasm, and emoji detection can be a future model.

In context to Indian Language, earlier work done for sentiment analysis has been on Bengali and Hindi, rest all the languages are unexplored. The nature of Indian languages varies a great deal in terms of the script, representation level and linguistic characteristics etc. So, there is a large amount of work that needs to be done to understand the behaviour of Indian languages and perform the analysis of same accordingly.

## Chapter 10

### Bibliography

---

- [1] Younis, Eman MG. (2015) "Sentiment analysis and text mining for social media microblogs using open source tools: an empirical study." *International Journal of Computer Applications*, 112(5) :0975 – 8887.
- [2] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* 5.4 (2014): 1093-1113.
- [3] Agarwal, Apoorv, et al. "Sentiment analysis of Twitter data." *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics
- [4] P. D. Turney, semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting on Assoc. for Computational Linguistics*, Philadelphia, PA.
- [5] B. Yang and C. Cardie, \Context-aware learning for sentence-level sentiment analysis with posterior regularization." in *Proc. 52nd Annu. Meeting on Assoc. for Computational Linguistics*, Baltimore,MD, 2014, pp
- [6] M. Hu and B. Liu, \Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Seattle, WA, 2004
- [7] Sentiment analysis algorithms and applications: A survey Walaa Medhat a,\*, Ahmed Hassan b, Hoda Korashy.
- [8] Imran, Muhammad, et al. "Processing social media messages in mass emergency: A survey." *ACM Computing Surveys (CSUR)* 47.4 (2015): 67
- [9] Yu Liang-Chih, Wu Jheng-Long, Chang Pei-Chann, Chu Hsuan-Shou. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowl-Based Syst* 2013;41:89–97.
- [10] Michael Hagenau, Michael Liebmann, Dirk Neumann. Automated news reading: stock price prediction based on financial news using context-capturing features. *Decis Supp Syst*; 2013.
- [11] Tao Xu, Peng Qinke, Cheng Yinzhaoh. Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *KnowlBased Syst* 2012;35:279–89.
- [12] A. Farzindar and D. Inkpen, \Natural language processing for social media," *Synthesis Lectures on Human Lang. Tech.*, vol. 8, no. 2, pp. 1{166, Sept. 2015.
- [13] C. Strapparava and R. Mihalcea, \Learning to identify emotions in text," in *Proc. the 2008 ACM Symp. Appl. Comput.*, Cear, Brazil, 2008, pp. 1556{1560.
- [14] Kasper, W. & Vela, M., "Sentiment analysis for hotel reviews", proceedings of the computational linguistics-applications, Jacharanka Conference, 2011.
- [15] Hemalatha, I., Dr GP Saradhi Varma, and A. Govardhan. "Case Study on Online Reviews Sentiment Analysis Using Machine Learning Algorithms." *International Journal of Innovative Research in Computer and Communication Engineering* 2.2 (2014):3182-3188
- [16] Bandgar, B. M., and Binod Kumar. "Real time extraction and processing of social tweets." *International Journal of Computer Science and Engineering, EISSN* 2347-2693 (2015): 1-6.
- [17] Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2019). Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication*, 29(3), 252-273



- [18] Maks Isa, Vossen Piek. A lexicon model for deep sentiment analysis and opinion mining applications. *Decis Support Syst* 2012;53:680–8.
- [19] Amitava Das, Sivaji Bandopadaya, SentiWordnet for indian language, *Proceedings of the 8th Workshop on Asian Language Resources*, pages 5663, Beijing, China, 21-22 August 2010.
- [20] Aditya Joshi, Balamurali AR, Pushpak Bhattacharya, A fall back strategy for sentiment analysis in hindi, *Proceedings of ICON 2010: 8th International Conference on Natural Language Processing*
- [21] Piyush Arora, Akshat Bakliwal, Vasudev Verma, Hindi Subjective Lexicon Generation using WordNet Graph Traversal, *IJCLA VOL. 3, NO. 1, JAN-JUN 2012*.
- [22] Namita Mittal, Basant Aggarwal, Garvit Chouhan, Nitin Bania, Prateek Pareek, Sentiment Analysis of Hindi Review based on based on Negation and Discourse Relation, *International Joint Conference on Natural Language*
- [23] Piyush Arora, Sentiment Analysis For Hindi Language, MS by Research in Computer Science, Master Thesis, IIT Hyderabad, 2013.
- [24] D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 675–682, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [25] Bo, P., & Lillian, L., “Opinion mining and sentiment analysis”, *Journal Foundations and Trends in Information Retrieval*, Vol. 2, 2008.
- [26] Younis, Eman MG. "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study." *International Journal of Computer Applications* 112.5 (2015)s
- [27] Liu, Haibin, et al. "BioLemmatizer: a lemmatization tool for morphological processing of biomedical text." *Journal of biomedical semantics* 3.1 (2012):
- [28] Bhattacharyya, Pushpak, et al. "Facilitating multi-lingual sense annotation: Human mediated lemmatizer." *Global WordNet Conference*. 2014
- [29] C. Hutto and E. Gilbert, \A parsimonious rule-based model for sentiment analysis of social media text," in *8th Int. Conf. Weblogs and Social Media*, Ann Arbor, MI, 2014
- [30] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." *Icwsn* 11 (2011): 538-541.
- [31] Ohana, Bruno, and Brendan Tierney. "Sentiment classification of reviews using SentiWordNet." *9th. IT & T Conference*. 2009
- [32] Yin, Z., Rong, J., and Zhi-Hua, Z., “Understanding Bag-of-Words Model: A Statistical Framework”, *International Journal of Machine Learning and Cybernetics*, 2010.
- [33] Fernández-Gavilanes, Milagros, et al. "Unsupervised method for sentiment analysis in online texts." *Expert Systems with Applications* 58 (2016): 57-75
- [34] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.
- [35] Murphy, Kevin P. "Naive bayes classifiers." *University of British Columbia* (2006)
- [36] D. Narayan, D. Chakrabarti, P. Pande, and P. Bhattacharyya. An experience in building the indo wordnet - a wordnet for hindi. In *First International Conference on Global WordNet*, 2002

## REFERENCE FROM TABLE

- [1] Twitter Sentiment Analysis Aliza Sarlan<sup>1</sup>, Chayanit Nadam<sup>2</sup>, Shuib Basri<sup>3</sup>
- [2] Sentiment analysis algorithms and applications: A survey Walaa Medhat, Ahmed Hassan, Hoda Korashy
- [3] Sentiment Analysis to Predict Election Results Using Python Ms. Farha Nausheen, Ms. Sayyada Hajera Begum
- [4] Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree Ferdin Joe John Joseph
- [5] Sentimental Analysis of Twitter Data with respect to General Elections in India Ankita Sharma, Udayan Ghoseb USICT, Guru Gobind Singh Indraprastha University, New Delhi, India
- [6] A Practical Approach to Sentiment Analysis of Hindi Tweets, Yakshi Sharma\*, Veenu Mangaty and Mandeep Kaurz
- [7] Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections Prabhsimran Singha, Yogesh K. Dwivedib, Karanjeet Singh Kahlonc, Annie Pathaniaa, Ravinder Singh Sawhney
- [8] A survey on sentiment analysis challenges Doaa Mohey El-Din Mohamed Hussein
- [9] Prediction of Indonesia Presidential Election Results for the 2019-2024 Period Using Twitter Sentiment Analysis Dinar Ajeng Kristiyanti, Normah, Akhmad Hairul Umam
- [10] Twitter Sentiment Analysis using Various Classification Algorithms Ajay Deshwal<sup>1</sup>, Sudhir Kumar Sharma
- [11] A Survey of Sentiment Analysis from Social Media Data, Koyel Chakraborty , Siddhartha Bhatt

