Sentiment Prediction from Indian Political News Content

Ritwika Basak
Registration No. - 154237 of 2020-21

Examination Roll No. - MCA2360017

Supervisor: **Prof.Subhadip Basu**

Department of Computer Science & Engineering
Jadavpur University
Kolkata - 700032

This project report is submitted for the partial fulfillment of the degree of
Master of Computer Application

May 2023

Faculty Of Engineering And Technology
Jadavpur University

Certificate of Recommendation

This is to certify that the dissertation entitled "Sentiment Prediciton from Indian Political News Content" has been carried out by Ritwika Basak (University Registration No.: 154237 of 2020-21, Examination Roll No.:MCA2360017) under my guidance and supervision and be accepted in partial fulfilment of the requirement for the Degree of Master of Computer Application(MCA). The research results presented in the thesis have not been included in any other paper submitted for the award of any degree in any other University or Institute.

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

.....

Prof. Subhadip Basu(Thesis Supervisor)
Department of Computer Science and Engineering
Jadavpur University, Kolkata-32

Prof. Nandini Mukhopadhyay
Head, Department of Computer Science
and Engineering, Jadavpur University, Kolkata-32.

Prof. Ardhendu Ghoshal
Dean, Faculty of Engineering and Technology
Jadavpur University, Kolkata-32.

Faculty Of Engineering And Technology
Jadavpur University

Certificate of Approval

This is to certify that the thesis entitled "Sentiment prediction from indian political news content" is a bonafide record of work carried out by Ritwika Basak in partial fulfilment of the requirements for the award of the Degree of Master in Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of January 2023 to May 2023. It is understood that by this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approves the thesis only for the purpose it has been submitted..

_____

Signature of Examiner 1
DATE :

_____

Signature of Examiner 2
DATE :

Faculty Of Engineering And Technology
Jadavpur University

Declaration of Originality and Compliance of Academic Ethics

I hereby declare that this thesis entitled "Sentiment prediction from Indian political news content" contains a literature survey and original research work by the undersigned candidate as part of my Degree of Master of Computer Science. All information has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all materials and results that are not original to this work.

Name :Ritwika Basak
Exam Roll No :MCA2360017
Thesis Title: “Sentiment Prediction from Indian Political News Content”

_____

Signature with Date

ACKNOWLEDGEMENT

I would like to thank the holy trinity for helping me deploy all the right resources and shaping me into a better human being. I would like to express my deepest gratitude to my advisor, Prof. Subhadip Basu, Department of Computer Science and Engineering, Jadavpur University, for his admirable guidance, care, patience and for providing me with an excellent atmosphere for doing research. Our numerous scientific discussions and his many constructive comments have greatly improved this work. Without his enthusiasm, encouragement, support and endless optimism, this thesis would hardly have been continued. Most importantly, none of this would have been possible without the love and support of my family. I thank my parents , whose forbearance and whole-hearted support helped this endeavour succeed. This thesis would not have been completed without the inspiration and support of several wonderful individuals. I appreciate all of them for being part of this journey and making this thesis possible.

Name : Ritwika Basak
Exam Roll No : MCA2360017
Registration No:154237 of 2020-2021
Department of Computer Science Engineering,
Jadavpur University

Abstract

Twitter is a widespread micro-blogging, social media platform used for political campaigning, and in India, it is among the largest online platforms used for sharing political information and expressing ideas, and opinions about politics which gain popularity when liked by a large number of users. In today social media users create a huge amount of unstructured text in the form of messages, posts, and blogs. The aim of this paper is to predict the subject of political news and sentiment of that news by using a ml model which is trained by the 2019 twitter data on indian politics. In this research paper, we use 2019 Twitter data and find the polarity of every tweet. We collect 28,000 political headlines from 4 news websites .First we collect dataset from kaggle then we preprocess the tweet then using some machine learning model we predict the subject of tweet and sentiment of it .Lastly we use this model to predict the subject and sentiment of politics news.

The newspaper data was collected using Beautiful Soup python library which scrap the news from various web site. The sentiment analysis was performed using a dictionary-based lexicon approach in conjecture with a “bag-of-word” and “TF-IDF” and vector space model.

Table of contents

Chapter 1

Introduction

Social media is an internet-based form of communication. Social media platforms allow users to have discussions, share information and create web content. In the world social media is increasing rapidly with new innovations conversion of data to information. The Twitter data is particularly appropriate for our study since Twitter is widely used and popular microblogging service in the election prediction as well as sharing opinions regarding Political Parties.

In order to analyze opinions in tweets, we apply sentiment analysis, which is the research area aiming at detecting the people's attitude, emotions, or opinion about a given topic expressed in text where the word sentiment represents an attitude, view, or opinion caused by emotion and we detect the sentiment is positive or negative or neutral.It is important to perform appropriate preprocessing of such data in order to prepare it in the best possible way as input of sentiment analysis algorithms. Basic task in sentiment analysis is to classify text as being positive, negative, or neutral. This work presents the investigation of the different political preferences a user can have about the Indian electoral parties of 2019. The identification of real time political preferences over social media platforms, e.g. twitter is considerably useful in political campaigns.The tweets are cleaned and preprocessed then we find the sentiment of every tweet using VADER polarity score lastly we create a model to classify that the sentiment of the tweet and subject of the party. The tweets were collected during a period of three months from Jan to March 2019 as this was the peak time when sentiments and opinions were at an all-time high , since it was the time just preceding the 2019 Lok Sabha elections, which were held in April. Different features are extracted and various learning approaches are examined on the preprocessed data. Various machine learning models such as Support Vector Machines, Logistic Regression and Random Forest, naive bayes are being considered for the investigation and deep learning model.

We collect headlines form newspaper which are the biggest form of print media in India with more than 425 million daily readers according to the Indian Readership Survey (IRS). Politics is deeply associated with the electorate's perception and the media plays a pivotal role in shaping it. We are looking at events with a large quantity of streaming information which is dynamic. Public generates emotions, opinions, sentiments, evaluations, appraisals and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. General during our daily news coverage we observe heated correspondences between the opposition party and the ruling party where the ruling party is implicated of wrongdoing and later response from the government and it goes back and forth year-round.

This research work specifically studies the news articles generated by 4 news paper for a year 2022-23. The aim was to stream newspaper articles collected from authentic and authoritative sources created to predict the name of the political party about which the news is published and the sentiment of that news. This aims to create framework thought which opinion polling could be achieved using sentiment analysis.

Chapter 2

Literature Review

In this chapter, the following topics are discussed: data science, social media, Twitter, newspaper, Python, and related work on political sentiment analysis and subject prediction of the news.

2.1 Data Science

In this thesis, we work on the analysis of tweet data. And it is the part of data sciences (Hayashi, Chikio, 1998). Data sciences now a days is one of the fastest growing field in the world. The area studies how to extract the data from different disciplines and interact between each other like mathematics (statistics and algorithms), software engineering and data communication. Data science is divided into different parts such as data collection, knowledge extraction from data, data preparation (cleaning the data and transformation of the data), exploration of the data (what can be done with the gathered data and how to use it), modeling the extracted knowledge with effective tools (we used python), visualization and communication which can be one of the most trickiest part of the data since it is challenging in the thesis how to visualize and how to convey the data for other people ? And finally testing of that data through the tool. Data science also called interdisciplinary field that used scientist in methods, process, algorithms and systems said by (VasantDhar, 2013). The process of extracting meaningful information from the big raw data. Data science is the fourth "paradigm" of science that "everything of science is changing because of the impact of information technology" said in his article the name "fourth paradigm of science" by (Stewart Tansly, el, 2009). There are three different categories for data, first data analyst between the data communication and statistics, second data engineering between software engineer and mathematics and then the data scientist which analyzing every field. (Davenport Thomas H, el, Oct 2012) he said that the data scientist the attractive job of twenty first century.

2.2 Social Media

In this thesis the social media play the main role. The social media is digital world where people meets together without their presence. But Wikipedia write something like this “social media (noun) is website and application that enable users to create and share the content or participate in a social network” this is formal definition which everyone knows. The word social comes from society and society is living place where the human living the proper way with rules and regulation, make community interaction each other (buildings, roads and meeting clubs) which sometimes harmful for nature. In other word the place where human living and breeding together. Media is the band of communication (acquired and spread knowledge) between the people. Social media is digital platform where people communicate (sharing information or data) with each other to sharing their ideas for the benefit of new generation. There are many social networks like Twitter, Facebook, Instagram, Snapchat and etc. (Obar, el, 2015), where social media is referred as web 2.0 based interaction application. The social media influenced by 1840’s introduction of telegraph in USA, which connecting the country (the Daily Dot, 2016).

2.3 Twitter

Twitter is a social networking or a blogging platform that was founded in 2006 by Jack Dorsey, Biz Stone, Noah Glass and Even Williams (Twitter, 2016). Twitter is one of the biggest social media networks in the world. Twitter is the treasure trove of sentiments people around the world, since people update thousands of actions, opinions, on every topic on every second of the day. It is called one of the biggest psychological database which always being updated and we can analyze the millions of data through the machine learning. Twitter stands on good position in social media networks. Twitter was created in March 2006 founded by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams (Way back Machine, 2012). Twitter has 336 million active users and more the 100 million daily active users which posts every day more than 500 million posts which contains maximum 280 characters (Statista, 2018). Twitter has opened the most powerful API for developers which recognized as top 10 API of the world. Twitter has two type of accounts one for normal users and other one is developer accounts (using API). The normal users share and read the information (tweets) but the developer accounts have access to Twitter data through the API (Application program interface). In developer accounts data can 7 be collected through keys which is provided by Twitter.There are four types of keys, such as consumer key, consumer secret key, token key and token secret key. These keys are unique and different which are used in different

programming language to collect tweet data. Twitter is also a big hub for sharing political news . (Gilbertson, el, and 2011) said Twitter uses authentication for account security through the “SMS” service. Twitter is also a open source platform (twitter, 2013).

2.4 Political sentiment and print media

Political preferences of an individual are entirely based on one’s opinion. This bias can change the way that person perceives news information daily. However, this preference can be heavily influenced by news media and their coverage [1]. Harvard’s Lerner in a series of studies explains how emotions can influence decision-making their choices in her research [2]. Many psychologists have studied the relationship between various forms of emotional communication and its effect on an individual’s judgement [3]. Iyengar in her book pointed to the impact of news in framing the public’s preferential accountability for social, political and economic issues [4]. As mentioned before, politics is a practice of influencing people at an individual level and in this, all forms of media play a very important role as they often form opinions. Traditional media sources such as radio, television and newspaper as well as various social media platforms cover all kinds of political events. While social media platforms have gained significant traction, newspapers accessed via physical copies and digital lookups through smartphones remains a key figure in the spread of political news [5]. Newspapers usually embed writers’ bias and perspective in their articles but are largely controlled by the organizations they work for. Although journalists are refrained from using clearly positive or negative vocabulary. Studies observed that they resort to other means to express their opinion by placing the statement in more complex discourse or omitting facts . Reports in the UK suggest the “unashamedly partisan” nature of multiple news publications. Rueter reports of 2018 shows 51% of their respondents express concern over hyper-partisan content by Indian news publication . That said, print media plays a significant role in moulding the political sentiment of their readers. In a highly politicised context, a country with a history of communal violence and characterised by explosively growing access to news media and low trust in news, disinformation has emerged as a pressing issue in India. Reuters reported that 57% of their respondents said they are concerned about what is real and what is fake on the internet, a number comparable to levels of concern in Turkey and the United States . Growing numbers of cases have been sighted where new contents internationally feed to their belief of a particular sector and lobbyist to maximize financial gains from the commercialization of a viral news trend or through advertisements . One way of measuring the effect of newspaper coverage is to look at the voting intentions of readers just before elections. In most cases, it isn’t too much of a surprise. According to an

Ipsos/Mori poll in the UK, 65% of Telegraph readers said they would vote Tory in 2001 and 2005, while 67% of Mirror readers pledged to vote Labour in 2005. In the same year, 57% of Daily Mail readers said they intended to vote Tory while 22% promised to vote Labour and 14% Lib Dem . A similar study conducted by Brandenburg also found media bias 11 in the UK. He also stated Irish election of 2002 also witnessed negative sentiment projected towards political figures by certain news publications . Ahmad et al. also showed media bias as well as gender bias in the Irish election of 2011 . Indian General Election of 2014 and 2019 have also been of particular interest among the scholars and private agencies alike. Many studies have examined the political orientation associated with English newspapers leading up to the election. Padmaja et al., Barclay et al and Roy have all analysed the election using a print media and have been able to predict the outcome of elections with a high level of accuracy. Singh also followed a similar methodology to predict the Indian General Election of 2019 with success [6]. Although a vast amount of research has been conducted at national level politics in India covering national sentiment, the effect on the financial market and more . None have given attention to the more granular level and the very important State Assembly Election in the 29 States of India and question, could the results of one influence the results of another? Are they correlated? Is the conclusion cause? The lack of research on the impact of various state elections on the results of Indian General Election motivated me to study the political orientation in newspapers from different regions of the country leading up to both their respective state and national elections.

2.5 Sentiment Analysis

Sentiments analysis is the invented science of psychology and sociology and both are the scientific study of people emotions, relationships, opinions, and behaviors (wiki) defined by new Oxford American dictionary [6]. It is often used to capture opinion and sentiment of the public and media which plays a very important role and can decisively shape bias or prejudice [7]. Psychologist apply sentiments process through the hypothesis but data scientist apply through the data. In other words, it is the computational process which identifies and categories the opinions, thoughts and ideas through the text data. The sentiments analysis process also refer the NLP(Natural language processing). It is internal action process between human and computer.Machine learning techniques control the data processing by the use of machine learning algorithm and by classifying the linguistic data by representing them into vector form [8]. It also analyzes the treasure of natural language data. we are using VADER polarity score [9], if compound score is greater than equal to 0.5 then it is positive tweet and if it is less than -0.5 then it is negative tweet otherwise it is neutral.

2.6 Lexicon-Based Approach

The Lexicon based approach predict the sentiment using Vader aprroach [10].It obtains a score for each word in the sentence or document and annotates using the feature from the lexicon database that are present. It derives text polarity based on a set of words, each of which is annotated with the weight and extracts information that contributes to conclude overall sentiments to the text. Also, it is necessary to pre-process data before assigning the weight to the words.
Moreover, Lexicon dictionary or database contains the opinionated words that are classified with positive and negative word type, and the description of the word that occurs in current context. For each word in the document, it is assigned with numeric score, and average score is computed by summing up all the numeric scores and sentiment polarity is assigned to the document.

2.7 Machine Learning Approach

Machine Learning approach is widely seen in the literature on sentiment analysis [11]. Using this approach the words in the sentence are considered in form of vectors, and analyzed using different machine learning algorithms like Naïve Bayes, SVM, SVC,Logistic Regression ,Random Forest. The data is trained accordingly, which can be applied to machine learning algorithms.

Chapter 3

Methodology

In this thesis, both approaches have been combined, namely Lexicon-based and Machine learning for sentiment analysis on Twitter data. The algorithms were implemented for preprocessing of data set for filtering as well as reducing the noise from the data set [12]. Therefore, the core linguistic data processing algorithm using Natural Language Processing (NLP) has been developed and implemented, and assigned sentiment polarity to the tweets using lexicon-based approach. Finally, the data set is trained using machine learning algorithm [13]: Naïve Bayes and SVM ,SVC, XGBoost, Logistic Regression for measuring the accuracy of the training data set, and have compared results of both algorithms . The most abstract view of derived approach that combines the lexicon-based and machine learning for sentiment analysis is shown in Figure 1.

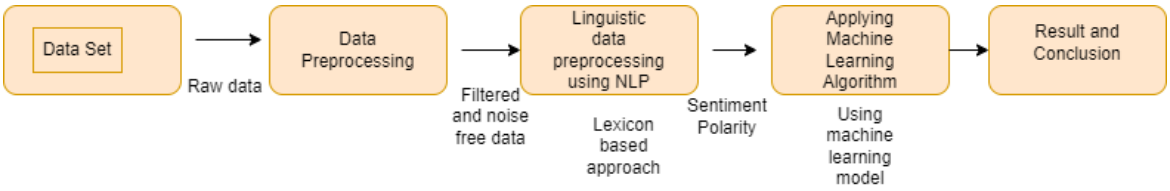


Fig. 3.1 Overview an approach for sentiment analysis

Chapter 4

DataSet

4.1 Data Collection

We have 46043 tweets in the dataset which are prominently significant to two national parties BJP(Bharatiya Janta Party), Congress and rest of the parties as Others. We collected 24,000 tweet of BJP and 20,000 tweet for Congress. For this study and analysis, only two major parties are taken into consideration and the rest of the parties are merged into Others, over which the analysis is performed collective in nature. Finally the field selected are 1.text:show the original tweet posted by user 2.subject:the name of the party which is mentioned by user in tweet , We take this dataset from Kaggle Indian Political Tweets 2019 (Feb to May).
To extract real-world sentiments expressed in the form of emotion and opinion, it was required to create a data set of indian political news which were relevant to our areas of research. To collect data we are using a python library BeautifulSoup [14]. Beautiful Soup [15] is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree.We are collecting 28,000 news from Times of India, Catchnews, Opindia, Indian Express.

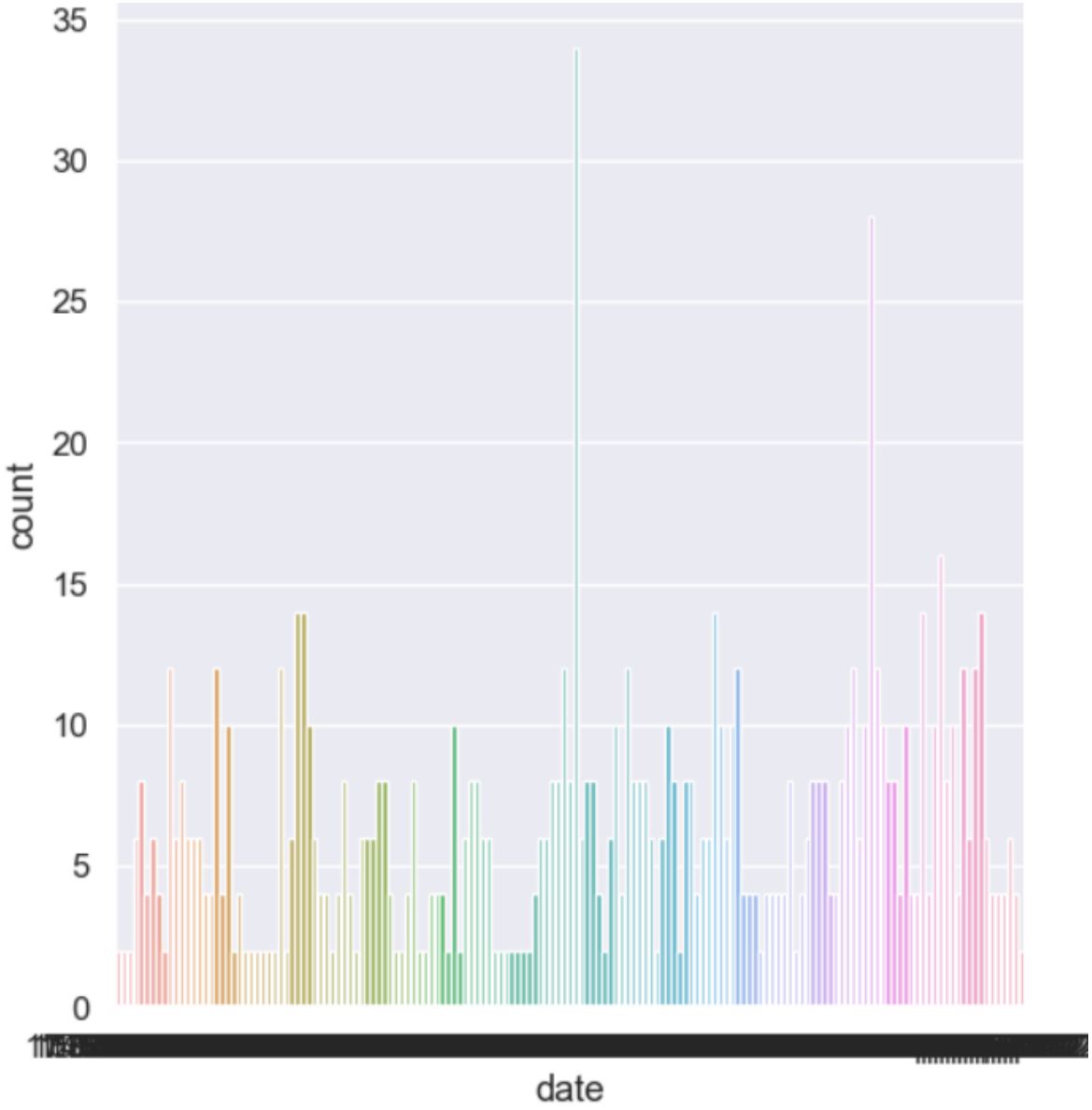


Fig. 4.1 News distribution of Times of India from apr 2022- apr 2023

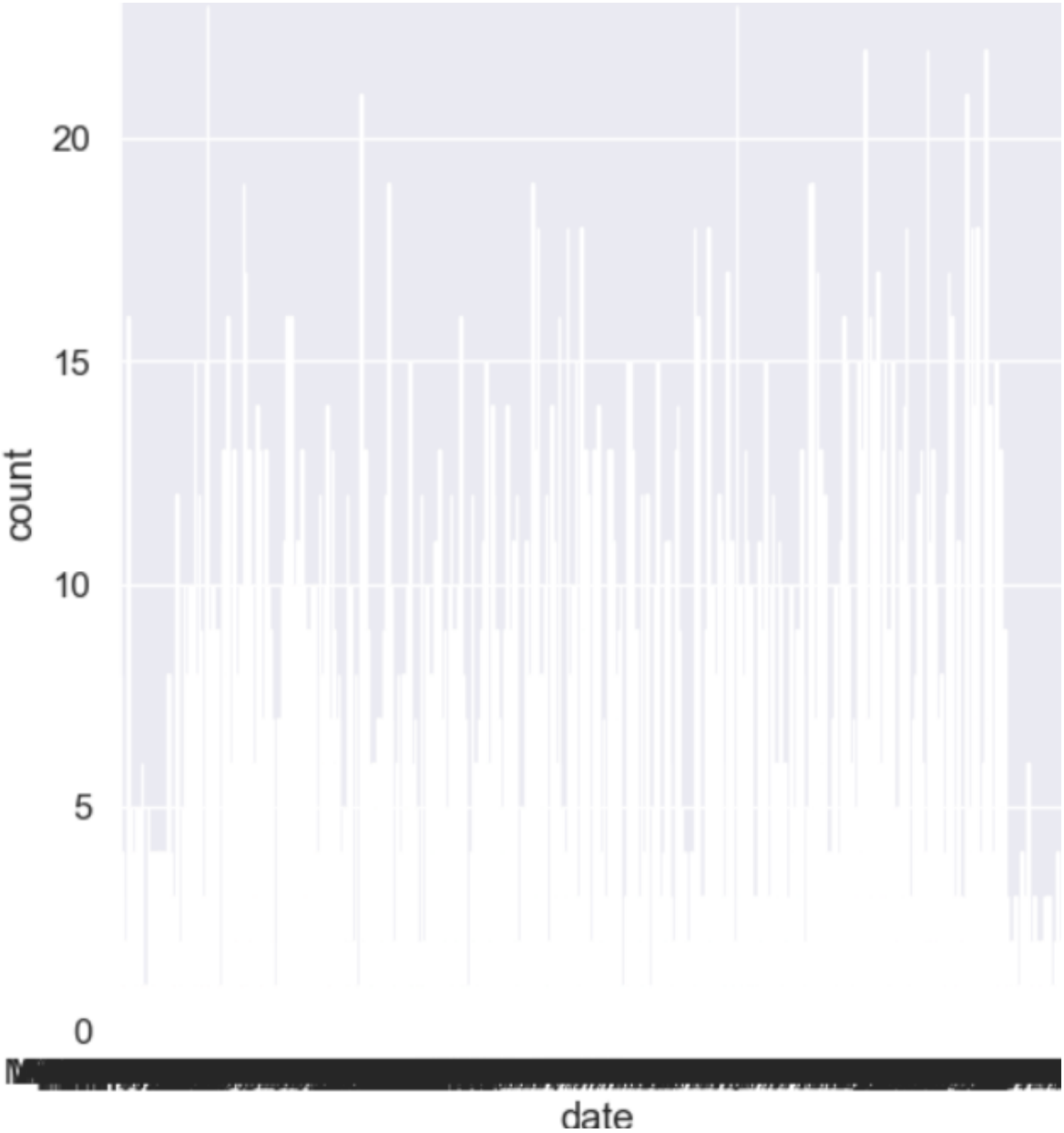


Fig. 4.2 News distribution of Indian Express from apr 2022- apr 2023

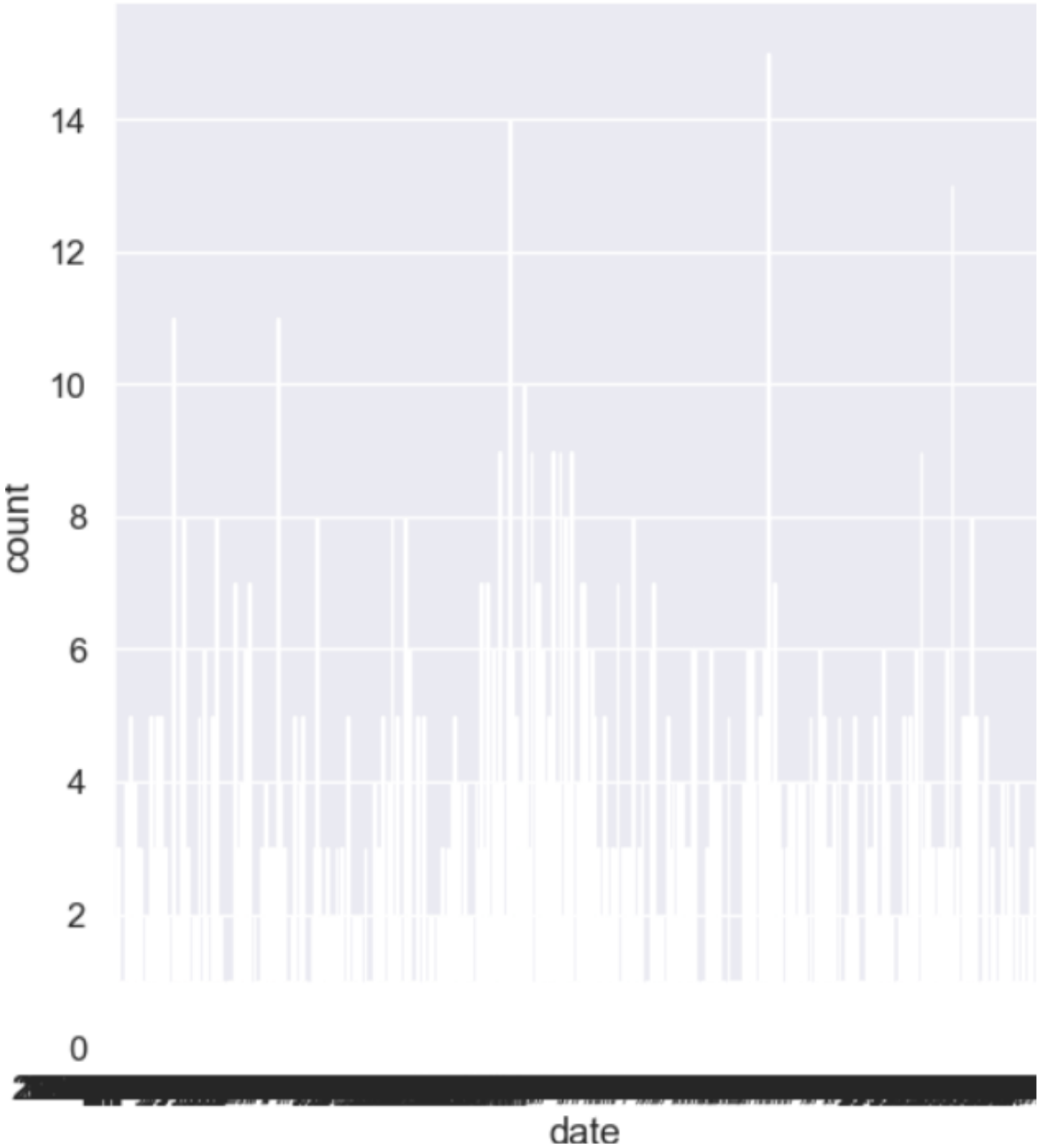


Fig. 4.3 News distribution of CatchNews from apr 2022- apr 2023

Chapter 5

Data Preprocessing

Data Preprocessing is the most important step of data mining which deals with the transformation and preparation of datasets for knowledge extraction [16][17]. There are several techniques involving in preprocessing. Some of them are cleaning, integrating, transforming, and reducing the dataset. This results in structured/clean data that are useful for modeling. Data collected or extracted from different sources are usually in their raw format which is not feasible for analysis, therefore, the raw data must first be cleaned before analysis. For all analysis projects, data cleaning takes about 70% of project work. It is cumbersome but extremely unavoidable. For this project work, data was extracted from February 2019 to May 2019. The uncleaned tweet was preprocessed to ensure the data is clean enough for model acceptance. We are only considering three column Full text, hashtag,Subject. The preprocessing is done of column text. Some of the preprocessing done on our raw data are stated below:

5.1 Remove the URLs from the text

To remove the urls from a text, a function as shown below is written in python and applied to the text[18].

```
def clean(text):
text = re.sub('https?://\S+|www\.\S+', '', text)
return text
    df['fulltext'] =df['fulltext'].apply(clean)
```

5.2 Remove HTML tags

To remove HTML tags from a text, a function as shown below is written in python and applied to the text [19].
def clean(text):
text=re.sub('<.*?>+', '', text)
return text
df['fulltext']=df['fulltext'].apply(clean)

5.3 Convert to lowercase

All the characters are converted to lowercase.The line of code below convert the characters into lowercase to avoid dulpication of words with different cases [20].
def clean(text):
text = str(text).lower()
return text
df['fulltext']=df['fulltext'].apply(clean)

5.4 Remove Punctuation

To remove punctuation form a text a function as shown below is written in python and applied to text [21]

```
text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
```

5.5 Removal Of Stop words

Stopwords are words that do not add meaning to a sentence and therefore can be ignored or removed without tampering with the meaning of the sentence [22].[reference stopwords]. Stopwords are found in most languages but for the purpose of this project work, stopwords in English are utilized. For sentimental analysis, Stop words are to be removed from the data to keep only the root words. Common stopwords include [i,me,my,myself,we,our,ours,ourselves,you,your]. A list of common stopwords (in english) can be found here. With these data, stopwords are removed by importing stopwords from the

corpus of nltk, the python module for natural language processing.
import nltk
from nltk.corpus import stopwords
stopwords = set(stopwords.words('english'))

5.6 Natural Language Toolkit(NLTK)

Using Python for performing operation on strings involves very simple functions for language processing tasks. To achieve an advanced functionality for processing linguistic data, Natural Language Toolkit (NLTK) available for Python is used. NLTK is a collection of modules and corpora, released under GPL open-source license, which permits student to learn and to perform research in NLP [23]. It has over 50 corpora and lexical resources like WordNet in combination with language processing libraries like work tokenization, classification and stemming, tagging, parsing and sematic rules for the analysis of text document, which will be discussed in detail (Bird et al. 2006). The key benefit of NLTK is that it is exclusively self-contained and has been praised by academic community [23]. Also, it not only gives access to methods and packages for common NLP tasks, but also provides the pre-processed and raw versions of standard corpora used in NLP literature and courses [23].

5.6.1 Word Tokenization

After filtering the noise from that dataset, all that was left were raw words in the sentences. These words individually have some meaning and may consist of emotion or sentiment expressed by the user in the tweet. In Natural Language processing, the process or steps for breaking down sentences into words and punctuations is known as Tokenization [23]. The goal for generating the list of words by separating the string is called Tokenizing sentence into words [24]. Here, to tokenize the words Natural Language Toolkit (NLTK) tokenize package is used. The choice for selecting tokenizer depends on the characteristic of data you are working on and the language. Here, to create a tokenizing method to tokenize the words using Tweet Tokenizer module for processing English language terms. The algorithm for word Tokenization using Tweet Tokenizer is shown in below
Processed data –> The Best World Cup Song So Far READY FOR BRAZIL Very GOOD World Cup Song Worldcup2014 Brazil2014
Word Tokenization –> ['The', 'Best', 'World', 'Cup', 'Song', 'So', 'Far', 'READY', 'FOR', 'BRAZIL', 'Very', 'GOOD', 'World', 'Cup', 'Song', 'World cup', '2014', 'Brazil', '2014']

5.6.2 Word Stemming

The stemming and lemmatizing of words are the approaches that produces the normalized form of a word [25] in the text. According to [26] word stemming is a technique that gets the root (base) of the word in the text. It normalize the word by removing the suffix from the word, which gives root meaning for the word. There are many stemming algorithms available openly to perform word stemming. In this approach of data pre-processing, the Snowball Stemmer algorithm is used for stripping suffix from the word to retrieve proper meaning from the text.Snowball stemmer is better than porter stemmer. Snowball Streamer is a real-time data processing framework developed by Twitter. It is designed to handle high-volume, high-velocity data streams efficiently. Snowball Streamer allows for the processing of data in real-time as it arrives, enabling organizations to gain valuable insights, make informed decisions, and respond quickly to changing conditions.

The framework is built on top of Apache Storm, which is a distributed, fault-tolerant system for processing streaming data. Snowball Streamer enhances the capabilities of Apache Storm with additional features and optimizations to handle large-scale data streams effectively.SnowBall Stemmer stems the word, character by character, and removes suffix and gives the base meaning to the word. Here, during the stemming process the word will stemmed and return the root meaning of the word.For instance, stemming the word ‘Caring‘ would return ‘Car‘.

Chapter 6

Data Exploration

In our twitter dataset data is collected from FEB 2019 to MAY 2019. After cleaning total row is 46043. In that dataset we have more than 24,000 bjp tweets and 20000 congress tweets .
Dataset Overview
The dataset consists of a collection of tweets related to Indian politics. It contains the following key fields:

- **Tweet ID**: Unique identifier for each tweet.
- **Timestamp**: Date and time when the tweet was posted.
- **Full Text**: The content of the tweet.
- **User Location**: where user belong to.
- **Retweets**: The number of retweets received by the tweet.
- **Favourites**: The number of times the tweet was favorited.
- **Hashtags**: Any hashtags included in the tweet.
- **Mentions**: Any user mentions within the tweet.

Dataset has 6072 unique user location and the most number of tweets are done from New Delhi, Mumbai, Washington DC, Bangalore,Hydrabad,Pune,United States.

There are multiple hashtags in the dataset.The most popular hashtags are `#LokSabhaElections2019`, `#MainBhiChowkidar`, `#TNWelcomesModi`, `#Modi`, `#BJP`, `#Congress`, `#GoBackModi` ,`#RafaleDeal` ,`#Modi2019Wave`, `#PulwamaAttack`, `#ChowkidarChorHai`

From the 5 news website, we can observe that all website produces neutral news most of the time.

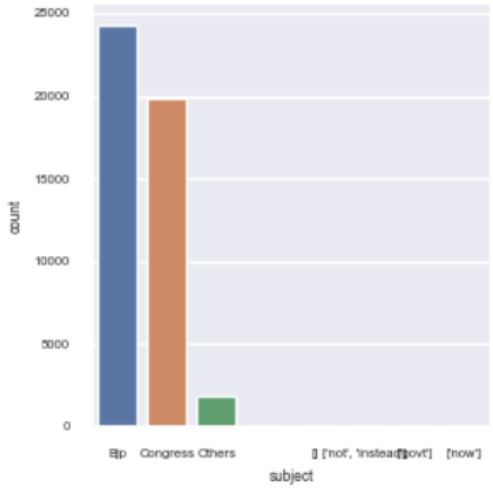


Fig. 6.1 Number of tweet of political party

Fig. 6.2 Wordcloud of BJP tweets hashtags

Fig. 6.3 Wordcloud of Cong tweets hashtags

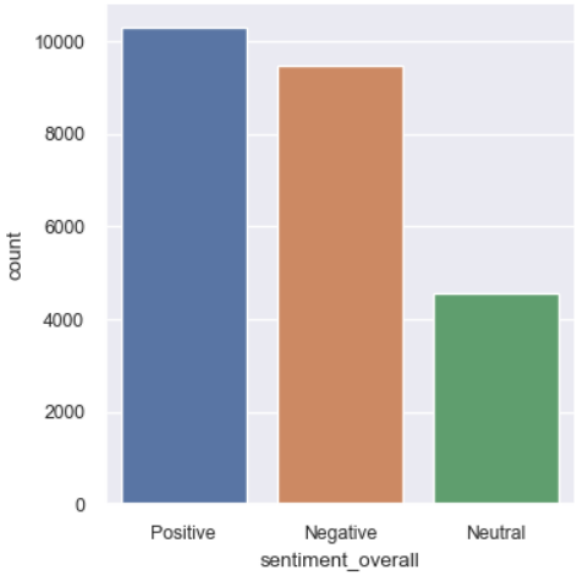


Fig. 6.4 BJP tweets sentiment

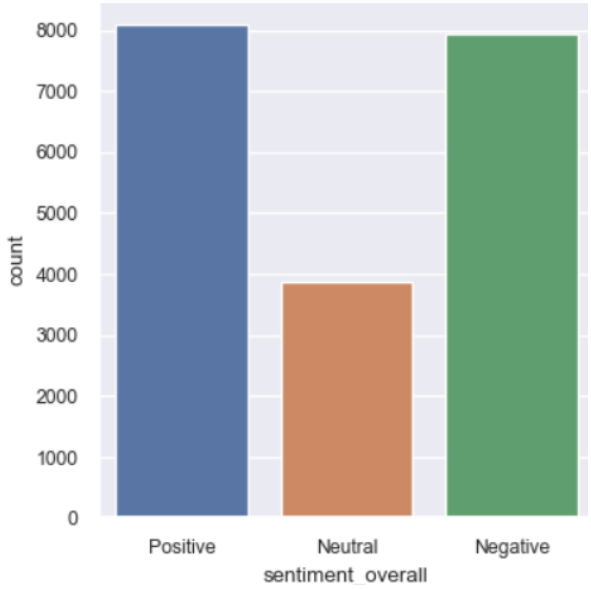


Fig. 6.5 Congress tweets sentiment

Fig. 6.6 Wordcloud of BJP positive tweet

Fig. 6.7 Wordcloud of BJP negative tweet

Fig. 6.8 Wordcloud of BJP negative tweet

Fig. 6.9 Wordcloud of Congress positive tweet

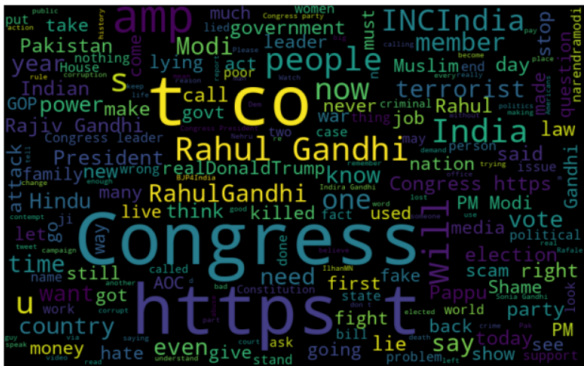


Fig. 6.10 Wordcloud of Congress negative tweet

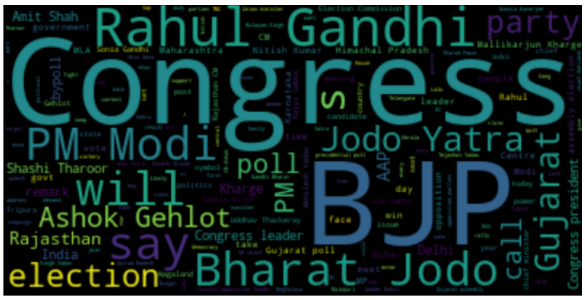


Fig. 6.11 Wordcloud of Times Of India news

Fig. 6.12 Wordcloud of Indian Express news

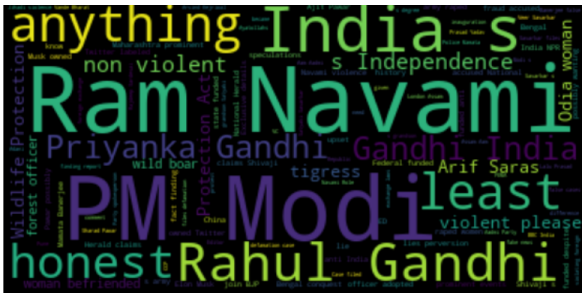


Fig. 6.13 Wordcloud of Opindia news

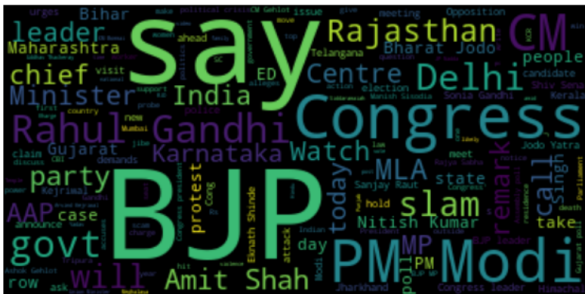


Fig. 6.14 Wordcloud of Catchnews news

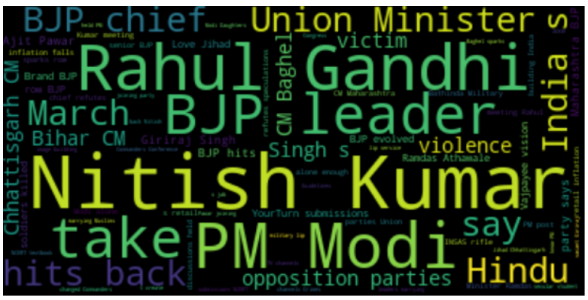
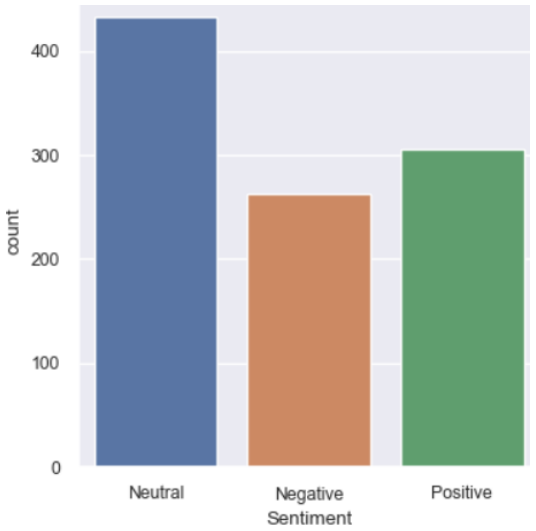
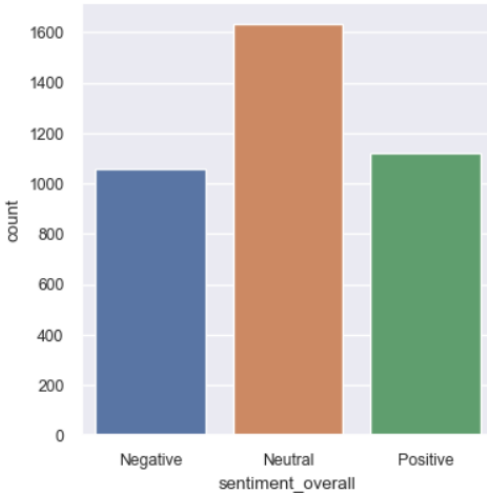


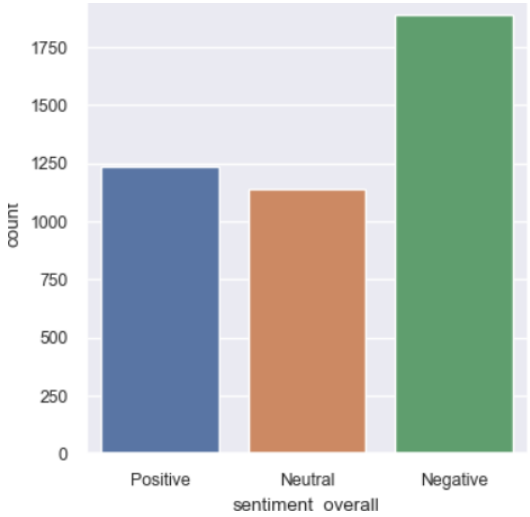
Fig. 6.15 Wordcloud of The print news



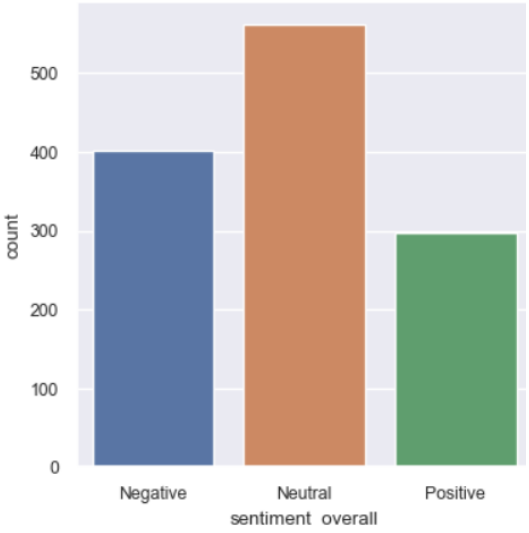
(a) Sentiment graph of Times Of India



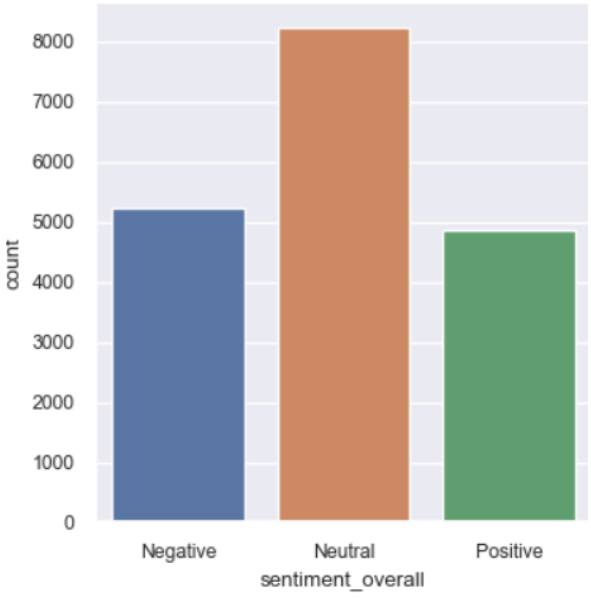
(b) Sentiment graph of Indian Express



(c) Sentiment graph of Opindia



(d) Sentiment graph of CatchNews



(e) Sentiment graph of The print

Fig. 6.16 Sentiment graph of different News paper

Chapter 7

Model Building And Experimental Results

This section covers the sentiment classification and subject prediction using 7 popular machine learning algorithms viz. Naive Bayes, Random Forest, and Support Vector Machine, Decision Tree, SGD, Logistic Regression, XGBoost. Since computers cannot process text data in its raw form, we must prepare the data before training the machine learning models. The text must be manually decomposed into a numerical format that the computer can understand [27]. Hence, we examine the results obtained using two natural language processing techniques such as Bag of Words and Term Frequency and Inverse document frequency approach [28]. Both BoW and TF-IDF are NLP techniques that help us convert tweet sentences into numeric vectors [29].

7.1 Feature Enginnering

Feature engineering in natural language processing (NLP) involves the process of transforming raw text data into numerical features that can be effectively used by machine learning algorithms. It plays a crucial role in extracting meaningful information and patterns from textual data.After selecting the relevant features from the datasets, some feature engineering are performed to transform the features of the datasets into Vectors and also creating new features from the dataset Some of these engineering features are discussed below:

7.1.1 Bag Of Words and term frequency - inverse document frequency

The Bag of Words model is a technique of extracting features from a text that can be used in modeling, like in our scenario for machine learning algorithms for tweet sentiment

classification. In simplistic terms, it is a group of words used to describe a sentence in a text with word count. It involves two things: first a list of well-known terms, and second metric for determining the presence of well-known terms. Another thing about BoW is the order in which they appear is discarded. The first step is to construct a vocabulary out of all the distinct words in our tweets Data frame. The next step is to list each of these distinct words and monitor their occurrence in every single tweet. Finally, you pass the matrix of numbers to the model for training purposes.

7.1.2 TF-IDF

The TF-IDF system outperforms the BoW approach because it is used to evaluate the importance of a word in a tweet [30]. When scoring word frequency, a common issue is that highly recurrent terms begin to dominate the text, but it may lack the 'informational content' required for the model to correctly differentiate. The IDF is a metric for measuring the significance of a word. We need the IDF value since just computing the TF isn't enough to appreciate the significance of words: The Eq. (1) shows the calculation of term frequency of the term t in document d. Term Frequency is a score dependent on the frequency in which a term appears in the document. Inverse Document Frequency is a metric for determining how rare a word is based on a document.

$$Tf(t,d) = \frac{N(t,d)}{T} \tag{7.1}$$

Here, TF(t,d) represents the term frequency of the term t in document d, N(t,d) is the number of times the term t appears in the document d, and T is the total number of terms in the document. Thus, for each document and word, a different TF(t,d) value will be assigned.

$$IDF(t) = \log \frac{N}{N(t)} \tag{7.2}$$

Equation (2) shows the calculation of IDF(t), which is the inverse document frequency of term t, N is the no, of documents, N(t) is the no. of documents with the term t.

$$TF - IDF = TF \cdot IDF \tag{7.3}$$

Equation (3) gives the calculation of TF-IDF.

7.2 Naive Bayes

The supervised learning algorithm, Naive Bayes is based on the Bayes' Theorem, which implies predictor independence [31]. In simple terms, a Naive Bayes classifier assumes that the presence of one function in a class has no bearing on the presence of any other feature. This allows one to comprehend what the Bayes theorem says. Often in machine learning, we need to select the best hypothesis(h) given the dataset (d). One of the simplest ways to choose a hypothesis is to use our previous knowledge of the situation. The Bayes' Theorem allows one to quantify the likelihood of a hypothesis given prior knowledge [32]. Bayes' Theorem is stated as:

$$P(h|d) = \frac{P(d|h) \cdot P(h)}{P(d)} \tag{7.4}$$

This equation gives the value for P(h/d), which is the probability of hypothesis h given the data d. This is called the posterior probability. P(d/h) is the probability of data d given that hypothesis h was true. P(h) is the probability of hypothesis h being true. This is called the prior probability of h. P(d) is the probability of the data. We should choose the hypothesis with the highest probability after determining the posterior probability for each hypothesis. The Maximum Posteriori Probability (MAP) hypothesis is used to describe this. We used the scikit-learn library to implement the Naive Bayes algorithm and before that, we converted the tweets to a matrix of token counts using count-vectorizer.

7.3 Decision Tree

In a tree-structured classifier, decision trees consist of internal nodes, which represent dataset attributes, branches that represent decision rules, and leaf nodes representing the outcome. In a decision tree, we have the decision node and the leaf node. Decision nodes are used to make decisions and have several branches, while leaf nodes are the result of those decisions and tell us whether the sentiment is positive, negative, or neutral and have no additional branches. Initially, our dataset consisting of tweets is considered as the root node or as the starting point to gain information.In decision tree algorithms, entropy is used to calculate the information gain for each attribute. The information gain measures the reduction in entropy achieved by splitting the dataset on a particular attribute [33]. The attribute with the highest information gain is chosen as the best attribute for the split [34].Entropy is calculated for a given dataset to measure the impurity or uncertainty in the class labels. In sentiment analysis, the class labels represent the sentiment categories (e.g., positive, negative,neutral). The entropy of a

dataset D with respect to sentiment labels can be calculated using the following formula:

$$E(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{7.5}$$

where n is the number of sentiment classes, and p_i is the probability of an instance belonging to the ith sentiment class. The probability p_i is calculated as the ratio of the number of instances in the class to the total number of instances in the dataset D.

7.4 Random Forest

Random Forest is a powerful ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and robustness in machine learning tasks [35]. It is widely used for classification and regression problems[36], offering a versatile and effective approach for a range of applications.Random Forest is constructed by building a collection of decision trees, where each tree is trained on a random subset of the original dataset [37]. The random subsets, known as bootstrap samples, are created by sampling with replacement. Additionally, a random subset of features is considered at each node during the tree construction process.Random Forest introduces randomness in two ways: random sampling of training data and random feature selection. These randomization techniques aim to introduce diversity among the individual decision trees, reducing the risk of overfitting and improving generalization.During prediction, each decision tree in the Random Forest independently produces a prediction. For classification tasks, the final prediction is determined through majority voting, where the class with the most votes across all trees is selected. For regression tasks, the final prediction is often the average or median of the predictions from individual trees.

7.5 Support Vector Machine

Support Vector Machine (SVM) is a widely used machine learning algorithm that is primarily used for classification tasks [38]. It is known for its ability to handle complex datasets and provide robust decision boundaries. SVM has gained popularity due to its solid theoretical foundation and effectiveness in various domains.SVM is based on the idea of finding an optimal hyperplane that separates different classes in the feature space. It aims to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class. The data points closest to the hyperplane are called support vectors, hence the name "Support Vector Machine."SVM can handle both linearly separable and non-linearly

separable datasets. In the case of linear separation, a linear hyperplane is used to separate the classes. For non-linear separation, SVM employs the kernel trick to map the input data into a higher-dimensional feature space where linear separation is possible.The margin in SVM represents the separation between classes. SVM aims to find the hyperplane that maximizes this margin, allowing for better generalization and robustness. The decision boundaries are defined by the support vectors and determine the classification of new data points.Kernel functions play a crucial role in SVM. They transform the input data into a higher-dimensional space, where non-linearly separable data can be linearly separated. Popular kernel functions include the linear kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel.SVM incorporates regularization to balance the margin maximization and the minimization of classification errors. The regularization parameter, often denoted as C, controls the trade-off between fitting the training data and the margin size. Other hyperparameters include the choice of kernel function and its associated parameters.

7.6 Logistic Regression

Logistic Regression is a popular statistical modeling technique used for binary classification tasks [39]. It is widely used in various domains, including healthcare, finance, and social sciences, due to its simplicity, interpretability, and effectiveness.Logistic Regression models the relationship between a dependent binary variable (target variable) and one or more independent variables (predictors) by estimating the probability of the binary outcome. It uses the logistic function (also known as the sigmoid function) to map the linear combination of predictors to a value between 0 and 1.The logistic function, represented as $\sigma(z)$, is defined as:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

where z is the linear combination of predictors and their corresponding coefficients.Logistic Regression estimates the coefficients of the independent variables through an optimization process, such as maximum likelihood estimation or gradient descent. The coefficients indicate the direction and strength of the relationship between the predictors and the log-odds of the binary outcome.Logistic Regression predicts the probability of the positive class (class 1) using the logistic function [40]. A threshold is then applied to convert the probabilities into binary predictions. Commonly, a threshold of 0.5 is used, where probabilities greater than or equal to 0.5 are classified as the positive class, and probabilities less than 0.5 are classified as the negative class.Logistic Regression models can be evaluated using various metrics, including accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics help assess the model's performance in

terms of correctly classifying the positive and negative instances.Logistic Regression can incorporate regularization techniques, such as L1 (Lasso) or L2 (Ridge) regularization, to prevent overfitting and improve model generalization. Regularization adds a penalty term to the loss function, discouraging large coefficient values and promoting simpler models.

7.7 Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is an optimization algorithm commonly used for training machine learning models, particularly in large-scale and online learning scenarios [41]. It is a variant of the gradient descent algorithm that updates the model parameters iteratively by considering a single or a subset of training examples at each step.Stochastic Gradient Descent differs from traditional gradient descent by considering a single training example or a small random subset of examples (known as mini-batches) at each iteration, rather than the entire training set. This random selection introduces noise but allows for faster updates and convergence, especially in large datasets.In each iteration of SGD, the model parameters are updated using the gradient of the loss function with respect to the selected training example(s). The parameters are adjusted in the opposite direction of the gradient, scaled by a learning rate that determines the step size.The learning rate is a hyperparameter in SGD that controls the magnitude of the parameter updates. A high learning rate can lead to rapid convergence but risks overshooting the optimal solution, while a low learning rate can result in slow convergence or getting stuck in local minima. Choosing an appropriate learning rate is crucial for SGD's effectiveness.SGD can be further categorized based on the size of the mini-batches used during parameter updates. Common choices are stochastic gradient descent (batch size = 1), mini-batch gradient descent (batch size > 1), and batch gradient descent (batch size equals the entire training set). The mini-batch size balances computational efficiency and stability of the parameter updates.

Table 7.1 Accuracies with different classifier of subject prediction

Model	**Bag-of-words**	**TF-IDF**
Logistic Regression	96.05	96
Naive Bayes	91.79	82
Random Forest	95.30	96
Decision Tree	99.50	96
SVC	99.50	96
SGD	99.50	96
XGBoost	97	97

Table 7.2 Accuracies with different classifier of sentiment prediction

Model	Bag-of-words	TF-IDF
Logistic Regression	81.66	86
Naive Bayes	72.00	63
Random Forest	72.97	85
Decision Tree	60.28	84
SVC	80.62	85
SGD	82.59	85
XGBoost	81.40	84

7.8 XGBOOST

XGBoost (Extreme Gradient Boosting) is a powerful and widely used machine learning algorithm that belongs to the family of gradient boosting methods. It is renowned for its efficiency, scalability, and state-of-the-art performance in various domains, including classification, regression, and ranking problems. key features of XGBoost:

- **Regularization**: XGBoost incorporates regularization techniques to prevent overfitting and improve generalization. It includes parameters for controlling the complexity of individual trees and the overall model.
- **Parallel Processing**:XGBoost efficiently leverages parallel processing capabilities to speed up training and prediction. It can take advantage of multiple CPU cores during training, making it highly scalable.
- **Tree Pruning**: XGBoost applies a technique called tree pruning to reduce the complexity of decision trees and avoid overfitting. Pruning removes branches that do not contribute significantly to improving the objective function.
- **Handling Missing Values**: XGBoost has built-in mechanisms to handle missing values in the input data, reducing the need for imputation techniques.
- **Cross-Validation**:XGBoost supports various cross-validation techniques to assess model performance and tune hyperparameters effectively.
- **Feature Importance**: XGBoost provides a feature importance score that indicates the relative importance of each input feature in the model's predictions.

Table 7.3 Precision,Recall,F1 Score in each class for all ML algorithm in subject prediction using Bag of words

Model	BJP precision	BJP recall	BJP F1 Score	Cong precision	Cong recall	Cong F1 Score	Others precision	Others recall	Others F1 score
Logistic regression	98	95	97	95	97	96	81	94	97
Naive Bayes	95	92	93	89	95	92	83	64	72
Random Forest	99	93	96	93	99	96	65	100	79
Decision Tree	100	100	100	99	100	100	98	99	98
SVC	99	94	97	95	98	96	74	100	85
SGD	98	95	96	95	97	96	78	1	87
XGBoost	99	96	97	96	98	97	81	99	89

7.9 Bert Model

7.9.1 For subject prediction

BERT (Bidirectional Encoder Representations from Transformers) is a highly influential natural language processing (NLP) model introduced by Devlin et al. in 2018 [42]. It has significantly advanced the field of NLP by achieving state-of-the-art results on various language processing tasks. The key innovation of BERT lies in its bidirectional transformer architecture, which enables it to capture contextual information from both the left and right contexts of a given word.

BERT undergoes a two-step training process. In the pretraining phase, BERT is pretrained on a large corpus of unlabeled text using two unsupervised tasks: masked language modeling (MLM) and next sentence prediction (NSP). The MLM task involves randomly masking 15% of the input tokens and training BERT to predict the original masked tokens based on their context. The NSP task involves determining whether two sentences appear consecutively in the original text or not.

After pretraining, BERT is fine-tuned on specific downstream tasks using labeled data. The model is initialized with the pretrained weights and then fine-tuned with task-specific layers added on top. The fine-tuning process involves minimizing the task-specific loss function, adapting BERT's learned representations to the target task.Here we have to predict three subject class BJP,Congress,Others. The accuracy of our model is 96.62%.

Table 7.4 Precision,Recall,F1 Score in each class for all ML algorithm in sentiment prediction using Bag of words

Model	**Pos precision**	**Pos recall**	**Pos F1 Score**	**Neu precision**	**Neu recall**	**Neu F1 Score**	**Neg precision**	**Neg recall**	**Neg F1 score**
Logistic regression	80	86	83	92	74	82	75	86	80
Naive Bayes	74	74	74	69	68	69	72	74	73
Random Forest	75	74	75	84	63	72	61	86	72
Decision Tree	65	65	65	53	49	51	59	61	60
SVC	80	83	82	90	74	81	74	85	79
SGD	81	88	84	95	73	83	75	88	81
XGBoost	80	86	83	95	72	82	73	88	80

Table 7.5 Performance evaluation in subject prediction using bag of words

Model	**Train score**	**Test score**
Logistic regression	97.08	96.05
Naive Bayes	92.01	91.79
Random Forest	96.05	95.30
Decision Tree	99.98	99.50
SVC	97.95	96.14
SGD	96.53	96.09
XGBoost	97.70	97

7.10 LSTM

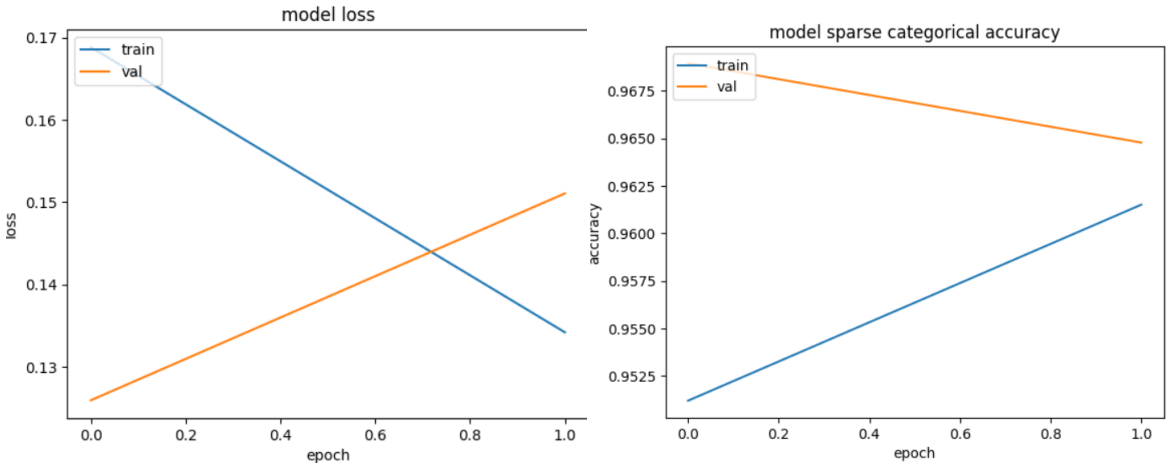
7.10.1 For sentiment analysis

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture introduced by Hochreiter and Schmidhuber in 1997 [43]. LSTM has proven to be highly effective in processing and modeling sequential data, overcoming the vanishing gradient problem that affects the training of traditional RNNs.

The LSTM architecture incorporates a memory cell and a set of gates into each recurrent unit, allowing it to capture long-term dependencies and store information over extended time intervals. The memory cell acts as a conveyor belt, while the gates control the flow of information within the LSTM unit.

Table 7.6 Performance evaluation in sentiment prediction using Bag-of-words

Model	Train score	Test score
Logistic regression	83.67	81.66
Naive Bayes	74.55	72.00
Random Forest	76.67	72.97
Decision Tree	99.01	60.22
SVC	90.91	80.62
SGD	83.51	82.59
XGBoost	84.77	81.40



(a) Loss graph in BERT model for sentiment prediction

(b) Accuracy graph in BERT model for sentiment prediction

LSTM consists of three main gates: the input gate, forget gate, and output gate. These gates manage the flow of information by regulating access to and modification of the memory cell. The equations governing the behavior of the LSTM unit are as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

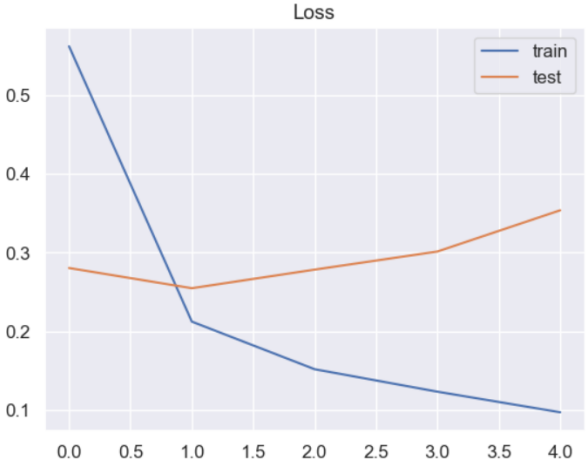
$$C_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot C_t$$

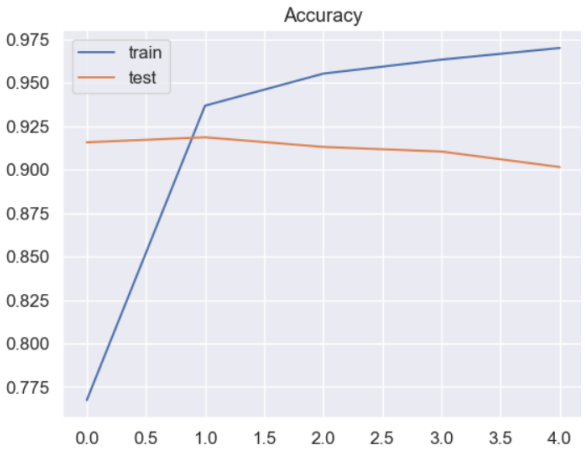
$$h_t = o_t \odot \tanh(C_t)$$

In the equations above, x_t represents the input at time step h_t is the hidden state/output at time step t C_t denotes the cell state at time step t , and W and b are the weight matrices and bias terms, respectively. The function σ represents the sigmoid activation function, and $\odot$ denotes element wise multiplication.

LSTM has been successfully applied in various domains such as natural language processing, speech recognition, and time series analysis. Its ability to capture long-term dependencies and mitigate gradient-related issues has made it a popular choice for modeling sequential data.we have three classes pos , neg , neu. The accuracy of the model is 89.8%.



(a) Loss graph in LSTM model for sentiment prediction



(b) Accuracy graph in LSTM model for sentiment prediction

From the performance evaluation of subject prediction we find that Decision Tree,SVC,SGD has the best performance of 99.50%**. For sentiment prediction, we find that LSTM have the best performance of** 89.8%**.**

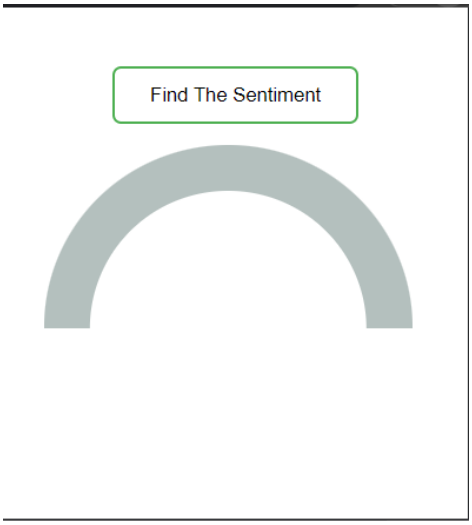
Lastly we test Times Of India's politics news . The performance evaluation of subject prediction and sentiment prediction is given in table 7.7 and 7.8.

Table 7.7 Political news performance evaluation in subject and sentiment prediction

Model	Sub prediction score	Sentiment prediction score
Logistic regression	69.19	62.3
Naive Bayes	66.60	57.49
Random Forest	68.10	60.4
Decision Tree	68.30	58.3
SVC	67.60	62.3
SGD	67.60	62.5
XGBoost	68	62.5

7.11 Application Discussion

We develop a Chrome extension aimed at analyzing the sentiment of political news articles. The main aim of this project is that it automatically assesses the sentiment of political news. So that we can easily understand the overall sentiment of the news of that paper. The chrome extension was designed to extract the text from the selected website .
The chrome extension consists of several components web-scraping, sentiment analysis using VADER sentiment function . Here in this extension we have only detecting the sentiment of the texts the subject prediction is not incorporated in this now.



(a) User interface of chrome extension

(b) Sentiment prediction of poltical news

Fig. 7.3 Different function of chrome extension

Chapter 8

Conclusion And Future Work

8.1 Conclusion

In this thesis we predict the subject and sentiment of political news using machine learning algorithm. For sentiment and subject prediction we removing noise or data filtering and preprocessing linguistic data using nlp techniques . We perform some operation to transform the text to a machine-understandable form .During this process the input tweets are filtered and processed to give more accurate data as well as reduce the size of dataset.We transform all tweets to lowercase and removing all html tags , punctuation marks , URL and then we tokenize the tweets then applying stemming and lematization the words obtained in this step are the base form of word which contains the root meaning for the given term. Hence, using derived algorithm the satisfied result is achieved that reduces the size of the dataset thereby, filtering unnecessary noise from the tweets and prepared tweets in the order perform further processing tasks.Then we analysis the tweet using machine learning algorithm Logistic regression, naive byas, random forest, decision tree, SVC, SGD, XGBoost, LSTM, BERT .

8.2 Future Work

For the future work on political sentiment analysis we should predict that the the tweet , news are negative or positive or neutral for subject (BJP,Congress) party . We could include these features on social media platforms for real time sentiment analysis so that a extreme negative posts can't create a violence and using these features we can also predict that news paper is biased or not.

References

[1] Jakob-Moritz Eberl, Markus Wagner, and Hajo G Boomgaarden. Are perceptions of candidate traits shaped by the media? the effects of three types of media bias. *The International Journal of Press/Politics*, 22(1):111–132, 2017.

[2] Christine Ma-Kellams and Jennifer Lerner. Trust your gut or think carefully? examining whether an intuitive, versus a systematic, mode of thought produces greater empathic accuracy. *Journal of personality and social psychology*, 111(5):674, 2016.

[3] Jennifer S Lerner and Dacher Keltner. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & emotion*, 14(4):473–493, 2000.

[4] Jean Chance. Shanto iyengar, is anyone responsible? how television frames political issues. chicago and london: University of chicago press, 1994. 208 pp. cloth, 22.95.paper, 11.95., 1996.

[5] Zeenab Aneez, TA Neyazi, Antonis Kalogeropoulos, and RK Nielsen. India digital news report 2019. *Reuters Institute for the Study of Journalism. Retrieved from https://reutersinstitute. politics. ox. ac. uk/sites/default/files/2019-03/India_DNR_FINAL. pdf*, 2019.

[6] Ayush Singhania. Political sentiment analysis: Investigating the impact of state elections on indian national elections using print media. 2020.

[7] Sylvie Graf, Pavla Linhartova, and Sabine Sczesny. The effects of news report valence and linguistic labels on prejudice against social minorities. *Media Psychology*, 23(2):215–243, 2020.

[8] Fredrik Olsson. A literature survey of active machine learning in the context of natural language processing. 2009.

[9] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[10] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008.

[11] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

[12] I Hemalatha, GP Saradhi Varma, and A Govardhan. Preprocessing the informal text for efficient sentiment analysis. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 1(2):58–61, 2012.

[13] I Hemalatha, GP Saradhi Varma, and A Govardhan. Sentiment analysis tool using machine learning algorithms. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(2):105–109, 2013.

[14] Leonard Richardson. Beautifulsoup: A python library for web scraping. *The Python Papers*, 3(1), 2008.

[15] S. Mukherjee, N. Banerjee, and S. Mandal. Exploring the feasibility of web scraping and sentiment analysis for real-time online reviews. In *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, 2019.

[16] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the international AAAI conference on web and social media*, volume 5, pages 538–541, 2011.

[17] Vitor Werner de Vargas, Jorge Arthur Schneider Aranda, Ricardo dos Santos Costa, Paulo Ricardo da Silva Pereira, and Jorge Luis Victória Barbosa. Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. *Knowledge and Information Systems*, 65(1):31–57, 2023.

[18] L. Priyatam, H. Yadav, and D. Chahal. Tweet normalization using edit distance metrics for url removal in twitter sentiment analysis. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015.

[19] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 4th edition, 2016.

[20] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[21] Yuhui Liu, Ting Zhang, and Jinghong Ma. Multi-perspective text sentiment analysis based on attention mechanism. *IEEE Access*, 8:123161–123174, 2020.

[22] Christopher D. Manning and Hinrich Schütze. The effects of stopword lists on information retrieval effectiveness. *Information Retrieval*, 1(1-2):143–160, 1999.

[23] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[24] Jacob Perkins. *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing Ltd, 2010.

[25] Michal Toman, Roman Tesar, and Karel Jezek. Influence of word normalization on text classification. In *Proceedings of InSciT 4*, pages 354–358, 2006.

[26] Eman MG Younis. Sentiment analysis and text mining for social media microblogs using open source tools: An empirical study. *International Journal of Computer Applications*, 112(5), 2015.

[27] S. Rawat, A. Rawat, D. Kumar, and A. S. Sabitha. Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights*, 1(2):100012, 2021.

[28] Y. Zhang, R. Jin, and Z. H. Zhou. Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

[29] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations*, pages 55–60, 2014.

[30] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing Management*, 39(1):45–65, 2003.

[31] F. Yang. An implementation of naive bayes classifier. In *2018 International conference on computational science and computational intelligence (CSCI)*, pages 301–306, 2018.

[32] S. Zervoudakis, E. Marakakis, H. Kondylakis, and S. Goumas. Opinionmine: A bayesian-based framework for opinion mining using twitter data. *Machine Learning with Applications*, 3:100018, 2021.

[33] P. H. Swain and H. Hauska. Decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3):142–147, 1977.

[34] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown. An introduction to decision tree modeling. *Journal of Chemometrics*, 18(6):275–285, 2004.

[35] G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):19–227, 2016.

[36] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[37] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[38] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[39] David W. Hosmer Jr and Stanley Lemeshow. *Applied logistic regression*. Wiley, 2004.

[40] Alan Agresti. *An introduction to categorical data analysis*. Wiley, 1996.

[41] Tong Zhang. *Stochastic Gradient Descent*. Chapman and Hall/CRC, 2021.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.

[43] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.