National College of Ireland

Real-Time Sentiment Analysis and Interactive Dashboard Deployment for Social Media Discourse

MSc Research Project
Data Analytics

Jaiprakash Chandraker
Student ID: 22213546

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jaiprakash Chandraker
Student ID:	22213546
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	10/08/2024
Project Title:	Real-Time Sentiment Analysis and Interactive Dashboard Deployment for Social Media Discourse
Word Count:	7500
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	21st August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Real-Time Sentiment Analysis and Interactive Dashboard Deployment for Social Media Discourse

Jaiprakash Chandraker
22213546

Abstract

This research project focuses on the development and implementation of a real-time sentiment analysis system and an interactive dashboard for monitoring social media discourse. the goal of the research is to identify the sentiment of people toward major political leaders in India, with a focus on Arvind Kejriwal, Narendra Modi, and Rahul Gandhi based on data gathered from Twitter. this study Employs the VADER sentiment analysis tool to classify the sentiments into positive, negative, and neutral. and a machine learning model based on a support vector machine (SVM) with TF-IDF feature extraction. The obtained data is then visualized using an interactive dashboard aimed to be a user-friendly and accessible means of understanding sentiment trends. Therefore, this study is important to fill the gap by providing real-time insights into public opinion on social media offering valuable information for political strategists, researchers, and policymakers.

1 Introduction

Today, social media networks act as highly influential tools for forming public opinion and determining the direction of political processes. Commission (2020) Social media like Twitter, Facebook, and Instagram have now become some of the most important platforms through which politicians, political parties, and citizens carry out interaction, expression of opinion, organization for meetings and engagements, and discussion.Center (2020) This movement from ordinary media to social media not only enabled political messages to get to the public but also brought new factors and ideas to the table when it came to polls and sentiments.Intyaswati and Fairuzza (2023) Social media is a massive data reservoir in terms of volume and coverage because people express their opinions on many issues and incidents in real time. However, the constantly increasing daily activity and the resulting information production on these sites require the creation of more effective instruments and mechanisms for data analysis. Shifting from the conventional approaches of opinion and polling-based analysis, political analysis is now getting support from computational structures that can work in real-time with large data sets. of such technique, there is sentiment analysis in which the emotional sentiment behind a text is analyzed. There are positive, negative, and neutral types of text, and this type of analysis can be used to gain insight into the large population. This study focuses on applying sentiment analysis to tweets related to key political figures in India, including the major political personalities, namely, Arvind Kejriwal from AAP, Narendra Modi from BJP, and Rahul Gandhi from INC.

1.1 Motivation

The rationale for this research is based on the imperative of comprehending the nature and meaning behind political discussions on social media. Therefore, through the sentiment of the tweets, to gain insight on public opinion, which is very critical for political strategists, researchers, and even policymakers. the purpose of this research is to fill the gap in the approaches to political analysis by creating a real-time sentiment analysis system and a dynamic dashboard for visualization. In particular, the study focuses on the sentiment surrounding three prominent Indian political leaders, Figures like Arvind Kejriwal, Narendra Modi, and Rahul Gandhi. Such leaders usually receive a lot of attention on social media and learning the tweets that are about them may bring information about the sentiment that the public holds about the leaders and their actions and policies.

1.2 Objectives

Specifically, the main purpose of this thesis is to design and implement an interactive dashboard and a sentiment analysis system to analyze the public feelings on social media platforms in real-time. undefined

- Classify Sentiments Implement the VADER sentiment analysis tool to categorize the analyzed tweets, concerning Arvind Kejriwal, Narendra Modi, and Rahul Gandhi as positive, negative, and neutral.

- Compare Sentiments Across Parties Identify and contrast the sentiment in the numerous tweets that are linked with AAP, BJP, and INC.

- Evaluate Model Performance Evaluate the effectiveness of a model based on support vector machines implemented with TF-IDF feature extraction for classifying tweet sentiments.

- Develop an Interactive Dashboard Develop an easy-to-use data visualization tool in the form of a dashboard that presents sentiment analysis findings at a glance, thus allowing the end-users to easily analyze the trends and sentiment reflected in social media platforms.

1.3 Research Question

The research question guiding this study is "How can public sentiment on particular topics or events be efficiently monitored and visualized on social media platforms through the development of an interactive dashboard and a real-time sentiment analysis system?"

This question pertains to the approaches needed to deal with the large amount of data that emanates from social networks and how real-time analytics and visualization are crucial for tracking public sentiments.

1.4 Structure of the Thesis

This thesis begins with an Introduction outlining the study's background, objectives, and methodology. The Literature Review explores prior research in sentiment analysis, focusing on political discourse and tools like VADER and machine learning models. The

Data Collection and Preprocessing chapter details methods for gathering and cleaning tweet data. Sentiment Analysis discusses using VADER for classifying sentiments toward political leaders. The Machine Learning Model section describes the SVM model training, TF-IDF vectorization, and evaluation metrics. The Results and Discussion chapter analyzes findings, comparing sentiments across political figures and discussing the study's strengths and limitations. The Conclusion and Recommendations summarize key insights and suggest areas for future research.

1.5 Significance of the Study

This study is significant for several reasons

- Advancement of Social Media Analytics The findings of this study concerning the real-time sentiment analysis system and the interactive dashboard benefit the development of social media analytics literature. The concepts introduced in this research are useful for areas apart from political analysis, for example, in marketing, customer service, and public relations.

- Insight into Public Opinion The study gives an understanding of the perception people of India have about some of the influential politicians. In other words, the analysis of tweets about Arvind Kejriwal, Narendra Modi, and Rahul Gandhi allows for gaining a better understanding of their image in the Twitter community's perspective.

- Real-time Monitoring and Visualization The solution that has been proposed and implemented in the scope of this work is the interactive dashboard which helps to gain insights on current trends in social media based on people's sentiment. This feature will be especially beneficial for political analysts, groups concerned with the study of the electorate, and policymakers who often require the latest perception and actions taken by the public.

- Methodological Contributions The study shows that on the analysis of the data with the help of social media platforms, the result can be generated more accurately with the help of sentiment analysis tools like VADER and machine learning models such as SVM. The method applied in this research study can be taken as a guide to similar research in social media analysis especially in automatic sentiment analysis.

2 Related Work

2.1 Social Media Twitter Sentiment Analysis

The paper D'Avanzo et al. (2017) presents an experimental approach of using Google search queries and tweets' social data for investigating sentiments and emotions towards emerging trends. The central part of the framework rests upon the method based on Bayesian machine learning and deep natural language processing to identify emotions and sentiment orientations of geolocated tweets on Twitter. Specifically, the framework was tested using the study that involved monitoring tasks in the consumer electronics, healthcare, and political domains. The findings indicate that the proposed evaluation methodology is useful in assessing the social media sentiments and emotions about Google

Trends. This shows that the framework has the ability to capture valuable opinions of the public at a given period of time. the paper under discussion offers an innovative approach that helps to combine Google search query information and Twitter sentiments. In terms of developments, the efficiency and value of the framework lie in its ability to guide the decision-making process in response to social media users' reactions to trends.

It was revealed that sentiment analysis is highly useful for capturing the public mood and its response to critical events, studied by Albaldawi and Almuttairi (2020)such as the COVID-19 pandemic. Given the proposed model's ability to analyze Twitter data, this paper establishes that AI and NLP can be applied to extract desirable information from social media platforms. The work under discussion focuses on four classification models, namely Logistic Regression, MLP, SVM, and Random Forest to investigate sentiment in the context of the social media platform, Twitter, and make predictions on the data. These models are tested under the Apache Spark framework following the assessment using Precision, Recall, F-measure, and confusion matrix. The following study's purpose is to establish which of these models works well enough to analyze sentiment in the Twitter data of students. By this, the researchers were able to obtain an accuracy of 0.71 when the MAPE was applied to the scenario with the record level equal to 8000. Also, the area of hyperparameter tuning for the purpose of fine-tuning the models and enhancing the accuracy values falls under the basket of the study. through the study, the performance of other classification models in sentiment analysis is revealed with a special focus on the right model for such a job. The results achieved within this study could be used as references for future research in the sphere of sentiment analysis and big data processing.

The research paper by Khan et al. (2020) aimed at employing sentiment analysis of Twitter data regarding the COVID-19 pandemic. Concerning the results of the study, it was determined that pre-processing of the data using the regex and a pre-trainer worked well in the improvement of the pre-processing of the algorithm alongside the simplification of the data. The use of the model when halving the data by splitting it then training the model and using it with a classifier lessened the time complexity of the analysis process. Analyzing the results, it was noted that active users posted more tweets than other sentiments in consideration of government decisions during the pandemic. Even when the infections, cases, and death tolls were rising, the mental strength of the populace was unchanged. The content analysis performed over three months in the Indian sub-continent revealed the trend of positive, negative, and neutral concerning cases. The model underwent training for pandemic-related sentiment analysis exclusively concerning the COVID-19 outbreak. The researchers advised increasing its size for examining responses toward other events in the world such as #blacklivesmatter, #Brexit, and #IndiaChinaFaceoff. This advice included the integration of a service that enabled the users to upload the data set on their own and refine the search results using keywords concerning the event as this would enable the improvement of the algorithm to be made in a more dynamic manner.

This paper aims to outline the use of Twitter to alert an organization and detect events such as earthquakes in real-time. Albayrak and Gray-Roncal (2019) They consider Twitter users as social sensing entities, where tweets related to events can be considered as

sensing data. For the study, a classifier is employed and for this, a support vector machine (SVM) is adopted in order to classify the tweets as either of the positive or negative sentiment. Three groups of features are considered for each tweet: features based on statistical characteristics of the documents, words, and phrases frequencies, and features based on the context of words' usage. This paper presents a spatiotemporal event detection model; this paper uses Kalman filtering and particle filtering to estimate the position of events. In their work, the authors noted that it was possible to estimate the coordinates of events based on the tweets connected with their locations. It also speaks of the application of a more complex algorithm of query expansion as a way of enhancing the recall of the tweets. Regarding the outcomes and the degree of precision, the paper presents the evaluation of the classification model with the help of Support Vector Machines. High classification accuracy of tweets in relation to earthquakes was attained by them. In the same manner, the actual location of lab residents was well estimated by the researchers in applying the particle filter approach which proves the usability of the developed spatiotemporal model. in particular, the paper describes a new method for event detection based on the analysis of the information published in the Twitter environment and demonstrates how social networks can be used as one of the sources of actual real-time information on events such as, for example, earthquakes.

The authors Sakaki et al. (2010) propose a framework for sentiment analysis on the social media platform, Twitter, that entails particular topics relating to the tweets. The model used in the study includes the feature set acquisition and using the information gained together with topics extraction for the Bag-of-Words approach with N-Gram identification and Parts of Speech linguistic tags. These features are then used in combination with a Decision Tree as a method of training and testing the software. This research shows great potential: the classifier reached a rather high training accuracy of 92.62%. The fact that it is very high means that the model developed to analyze the sentiment of tweets is effective based on the issues related to the tweets. Furthermore, the integration of the user interface with the help of Pyplot and Tkinter allows the performing sentiment analysis of the tweets concerning certain topics in real-time, thus the practical value of the framework. this paper offers a significant contribution to the research on sentiment analysis on the social media platform Twitter. the topic-based additional model helps to provide the reader with a more elaborate view of sentiment count in connection with specific topics expressed in tweets. The classifier achieves very high training accuracy, thus making it very reliable and viable for use in solving real-world sentiment analysis problems.

The paperAgarwal et al. (2023) contains a detailed discussion of a methodology used in the classification of sentiment data, specifically data acquired from social media platforms, such as Twitter and YouTube – using the BERT model. The authors highlight that BERT is very efficient in handling the casual language of users provided it is pre-trained so that it can distinguish between complicated and multiple sentiments expressed in the user-created content without retraining. The proposed method focuses on the generation of features and visualizations, which could be valuable for businesses by shedding light on people's attitudes toward their offerings. Furthermore, the study also focuses on the issues of data acquisition and data cleaning revealing how influential and crucial it is

to get clean data from social media APIs. The objective of the social science research is to propose an intuitive dashboard through which the consumers' sentiment analysis will be feasible in real-time, with an emphasis on the pictures of the consumer sentiment analysis. Besides, it increases the efficiency of the sentiment analysis as well as optimizes user satisfaction because it offers meaningful information processed from the comments left on social media accounts.

2.2 Machine Learning Methods

Customer reviews of mobile phones bought from Amazon are captured throughout the paper to establish the polarity of the reviews. This is useful in establishing the attitude of consumers towards different products. The study Singla et al. (2017) applies a method known as Naïve Bayes, a Support Vector Machine (SVM), and a Decision Tree to classify the reviews. The scholarly research essay on customer product reviews' sentiment analysis, employing the use of machine learning, underlines the significance of identifying customer sentiments in the context of e-commerce. Since the emphasis is made on grouping a large database of more than 4,000,000 reviews collected from Amazon, the methodology uses Naïve Bayes, Support Vector Machine, and Decision Tree algorithms to distinguish between positive and negative reviews. Thus, based on the outcome of 10 Fold Cross Validation, the study exposes the mean accuracy of each type of model. From the above table, Naïve Bayes attained an accuracy of 64. 57% to 68.60% The classification scenario with lower accuracy was Naive Bayes displaying only 57% The classification with the highest accuracy ranged from 77% to 82%, with the maximum accuracy marked at 8-1. The overall summary of the findings indicates that modest improvements in diagnostic effectiveness can be obtained by using an increase of 25% readers and reducing the number of false positives and false negatives to half. 77%. On the other hand, accuracy levels that were depicted with the help of the Decision Tree ranged from 67.68% and 81.25%. As shown by these outcomes, the type of question asked plays a favorable role in identifying the sentiment of customers' feedback, favoring the use of the SVM model as a classifier for sentiment in the e-commerce context. It not only helps consumers to get correct information from other consumers through product reviews but also helps the business world accumulate knowledge of consumer behavior to make better decisions in the effective completion of the world's marketplace.

To determine the attitudes toward e-cigarettes, the authors Martinez et al. (2019) of the study used a sample of tweets and analyzed the sentiment about the visibility of stigma, harm, addictiveness, and efficacy as a quit aid. Estimating the results through the SMART dashboard, they found out that the majority of the tweets are positive in terms of e-cigarettes. Stigma appeared to be a core component of confirming messages as well as rejecting ones. The study also took note of the emergence of semi-biotic accounts in the discussion. To carry out the study, the authors gathered and processed the Twitter data by utilizing sentiments to qualify the tweets as positive, negative, neutral, and ambivalent as well as nonsensical. In comparing attitudes towards e-cigarettes with the official health census, as well as with data collected throughout other studies, they aimed to better assess the public's perception of them. SMART dashboard was used in monitoring and visualizing trends in e-cigarette sentiment in real-time with the aim of developing a plan in public health interventions. Thus, the study also underscored the

significance of using social media analytics in the surveillance of health concerns. The results indicate that the majority of Twitter users discussed e-cigarettes in a positive light (68%) and that stigma situates itself fundamentally in the discursive construction of the social objects in positive tweets as well as in the exclusion and marginalization of e-cigarette smokers in negative tweets. Furthermore, the analysis pointed towards the activity of the hybrid human-bot accounts in the topic. Hence, it runs and compares the sentiment of tweets with the health records and offers rich information to public health workers. The approach used was the sentiment analysis that helped categorize the tweets and give real-time trend visualization. With regard to the results which show 68% positivity towards e-cigarettes, it is crucial to continuously analyze social media to know the people's perceptions and to create the right program for countering the social ills affecting the population.

Recent studies Mathew et al. (2022) have explored the use of Bi-LSTM models on Twitter-based live-streaming SNS data for sentiment analysis. An important comparison was also made between the Bi-LSTM model and the machine learning algorithms like Naïve Bayes and linear regression. The results evidenced that the Bi-LSTM model had better performance compared with those traditional methods; the accuracy rate of the Bi-LSTM model ranged from 80% to 85%.

the study dealt with analyzing the sentiments of the US election polling on Twitter in real-time. This study underlined the need to undertake a current analysis of social media to counter cyberbullying and evaluate the trending cases. The authors also sampled a possible method to categorize social media conversations according to the sentiments from positive, negative, or even neutral. The data was obtained from Twitter, the official news feed, and Google after which the data was pre-processed, and the Bi-LSTM model was used on the data collected. Therefore, the findings showed that the Bi-LSTM model had higher accuracy than the other methods.

The research paper Miah et al. (2024) aims at analyzing sentiments in Arabic, Chinese, French, and Italian translated into English. Due to the nature of the specific problem, the study uses an ensemble model based on transformers for the two parts as well as a large language model (LLM). The approach entails converting the entire content of these languages into English with the help of neural machine translation tools such as LibreTranslate and Google Translate. The translated sentences are then analyzed for sentiment using pre-trained models: The datasets are namely Twitter-Roberta-Base-Sentiment-Latest, bert-base-multilingual-uncased-sentiment, and GPT-3 from OpenAI. It is observed that the fusion of the individual pre-trained models results in an ensemble model, it can achieve an accuracy of over 86% when applied to translated sentences of the movie reviews in sentiment analysis test, and it performed better than the LLM and the individual pre-trained models. This research aims to contribute to the field by showing the possibility and the efficiency of translating text and using an ensemble model for sentiment analysis.

Past work explored the use of numerous methods for the sentiment analysis problem based on data from social media including logistic regression, MLP, and BERT. These studies stress the methods for large dataset storage and frequent model refinement for

accurate and precise sentiment classification while they do not address the issues related to hyperparameters with concern to the SVM implementation. Also, little prior research has been done on how to visualize sentiments in near real-time. However, in detail, this particular project chooses solely the SVM model for conducting sentiment analysis and sets itself apart by creating a live sentiment analysis dashboard. This dynamic interface will be used to track and map current fluctuations in the level of public interest.

3 Methodology

This chapter provides information on the research method that has been put in place to achieve the overall goal of the study with regard to the design and development of the sentiment analysis system and the interactive dashboard. The research process is designed to meet the objectives of the study by integrating theoretical and practical sections that relate to the research question. Some of the information stated in this chapter includes data collection techniques, data preparation, tools and techniques used for sentiment analysis, machine learning techniques used, and the creation of the dashboard. The study mainly seeks to employ a model of sentiments on the tweets by Indian politicians using natural language processing. It involves categorizing opinions with the help of filters to either be positive, negative, or neutral in text messages. This study will focus on the tweets of the top leaders of India's political parties, such as the Aam Aadmi Party (AAP), the Bhartiya Janata Party (BJP), and the Indian National Congress (INC).

3.1 Data Collection

The data for the analysis is gathered from Twitter and it includes more than 200,000 tweets downloaded from the Twitter API. This study focused on the leaders of three parties namely AAP (Arvind Kejriwal), BJP (Narendra Modi), and INC (Rahul Gandhi), tweets which were acquired from their official Twitter handles. The peculiarities of this dataset are that it contains the tweet ID, the date of the creation, the username, and the text of the tweet. Complete data allows for considering various features of public emotions towards these politicians' concerns in the context of their tweets. It enables carrying out an in-depth sentiment analysis coupled with visualizations on an interactive dashboard.

3.1.1 Data Collection Procedure

- Keyword and Hashtag Identification: Thus, specific keywords and hashtags associated with each political leader were determined. For example, ArvindKejriwal, NarendraModi, RahulGandhi, etc have been used to get the related tweets only.
- Using Twitter API: Tweets that contain the above-mentioned keywords and hashtags were collected using the Twitter API. To do this, the author created a developer account on Twitter to get API keys and then used these keys to make requests for data concerning the tweets.
- Data Storage: The tweets gathered were saved in CSV files containing attributes like the text of the tweet, the time it was posted, details of the user tweeting, and more. This type of structure made subsequent data preprocessing and analysis much easier to conduct.

3.2 Data Preprocessing

When the data is obtained from an API, it is often in the form of raw and unstructured information that requires thorough processing to be useful. The raw data must be cleaned and refined to ensure that it is accurate and precise. The steps involved in preprocessing include removing duplicates, managing missing values, correcting mistakes, and putting the data into a usable format. preprocessing refers to the standardization of data into one format. For instance, there may be normalization of texts or even changing them wholly into lower case letters or even deletion of uncommon characters and breaking down sentences into words.

3.2.1 Data Cleaning

This entire process of converting raw data into a usable format is essential for accurate and effective analysis.

3.2.2 Dates Format

Dates are formatted consistently to YYYY-MM-DD for simplifying sorting, filtering, and analysis of dates in the dataset.

3.2.3 Removing Links

Only URLs contained in tweets were eliminated in this dataset because it does not provide enough information for research.

3.2.4 Removing Duplicates and Null Values

Duplicate records should be identified and removed from a duplicated record index if necessary so that there is no redundancy. this way every single entry will be unique throughout this database. In addition, the lines having missing values were removed to avoid problems or errors caused as a result of keeping everything incomplete. Thus, it means that it will offer a reliable database devoid of extraneous elements that are ready for further analysis.

3.3 Text Preprocessing

During the process of pre-processing the text, a number of important steps were taken to enhance cleanliness and bring about uniformity in text data. Lowercasing, stopword removal, punctuation mark removal, deletion of repeating characters, as well as number and emoji elimination are some of these processes.

3.3.1 Lowercasing

All the characters in the data set were converted to lowercase in order to ensure that there exists uniformity across it. Such an act makes it easy to consider “Happy” and “happy” variants of the same word this makes analysis easier through simplification.

```
stopwordlist = ['a', 'about', 'above', 'after', 'again', 'ain', 'all', 'am', 'an',
             'and','any','are', 'as', 'at', 'be', 'because', 'been', 'before',
             'being', 'below', 'between','both', 'by', 'can', 'd', 'did', 'do',
             'does', 'doing', 'down', 'during', 'each','few', 'for', 'from',
             'further', 'had', 'has', 'have', 'having', 'he', 'her', 'here',
             'hers', 'herself', 'him', 'himself', 'his', 'how', 'i', 'if', 'in',
             'into','is', 'it', 'its', 'itself', 'just', 'll', 'm', 'ma',
             'me', 'more', 'most','my', 'myself', 'now', 'o', 'of', 'on', 'once',
             'only', 'or', 'other', 'our', 'ours','ourselves', 'out', 'own', 're','s', 'same', 'she', "shes", 'should', "shouldve",'so',
             't', 'than', 'that', "thatll", 'the', 'their', 'theirs', 'them',
             'themselves', 'then', 'there', 'these', 'they', 'this', 'those',
             'through', 'to', 'too','under', 'until', 'up', 've', 'very', 'was',
             'we', 'were', 'what', 'when', 'where','which','while', 'who', 'whom',
             'why', 'will', 'with', 'won', 'y', 'you', "youd","youll", "youre",
             "youve", 'your', 'yours', 'yourself', 'yourselves']
```

Figure 1: Stop words list

3.3.2 Removing Stopwords

Some common words like “the,” “is,” and “in,” are usually removed since they frequently appear within language but do not contribute much towards sentiment analysis. Such an action helps to concentrate more on relevant words only. (Figure 1: Stop words list)

3.3.3 Removing Punctuations

The removal of all punctuation marks serves as a way of purifying the data and making them less complicated.

3.3.4 Removing Repeating Characters

By doing this, such words with too many duplicated letters were placed back into their normal forms. for instance, “soooo” became “so.” This helps achieve coherence within our texts and increases precision.

3.3.5 Removing Emojis

the author applied some seriously slick algorithms to get those emojis out of there. While the funny faces can definitely convey emotion, removing them helps us only pay attention to textual information, making our pre-processing more efficient.

3.4 Sentiment Labeling

For sentiment labeling, the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool was used to categorize each tweet into the following three groups.

3.4.1 Positive

These are tweets that express favorable or happy sentiments. These tweets include words or phrases that have a positive tone as well as approval.

3.4.2 Negative

Tweets that convey an unfavorable or unhappy feeling are negative. Such tweets include words and phrases that are negative and carry with them disapproval.

3.4.3 Neutral

Tweets that have neutral sentiments. tweets such as these don't lean strongly towards any of these two extremes – they tend more towards being objective in tone and indifferent concerning valence.

3.4.4 Vader Sentiment Function

VADER is particularly well-suited for analyzing social media text due to its sensitivity to both the polarity (positive/negative) and intensity (strength) of sentiments expressed in short texts. A comprehensive lexicon of words with assigned sentiment scores is used along with contextual rules, the valence of words employed together with sentence modifiers like capitalization and exclamation marks. This tool can interpret informal yet contextually rich social media communications, making it effective for accurately labeling the sentiment of tweets. VADER's capability to manage brevity, as well as diversity in terms of language.

3.4.5 Machine Learning Model

3.4.6 Selection of Model

The selection of a model for further sentiment classification is a Support Vector Machine (SVM). SVM is another supervised learning algorithm extensively used for classification. It was chosen because of its capacity to respond well to large amounts of data with dimensions as well as its performance in text categorization.

3.5 Data Visualization and Distribution Plots

Visual tools are important for textual data summarization and opening areas of inquiry regarding sentiment trend analysis from a global aspect. First, word clouds give a quick overview of the most common words associated with each sentiment category while distribution plots present a quantitative perspective on their prevalence within the dataset under consideration. Consequently, this combination of different forms of representation contributes towards a better understanding of public sentiment as expressed via tweets. These visualizations consist of word clouds and distribution plots for positive, negative, and neutral tweets. (Figure 2: Distribution of data), (Figure 3,4: word clouds)

3.5.1 Positive Tweets

To visualize the most commonly used words in positive tweets, word clouds were created. The words that are the most in positive tweets are enlarged in these templates.

3.5.2 Negative Tweets

In addition, there were examples of negativity. hence similar constructions as above had been presented highlighting the commonest terms found through their usage by people expressing bad feelings respectively. Larger font sizes mark greater occurrences within them indicating frequent appearances among negatory posts.

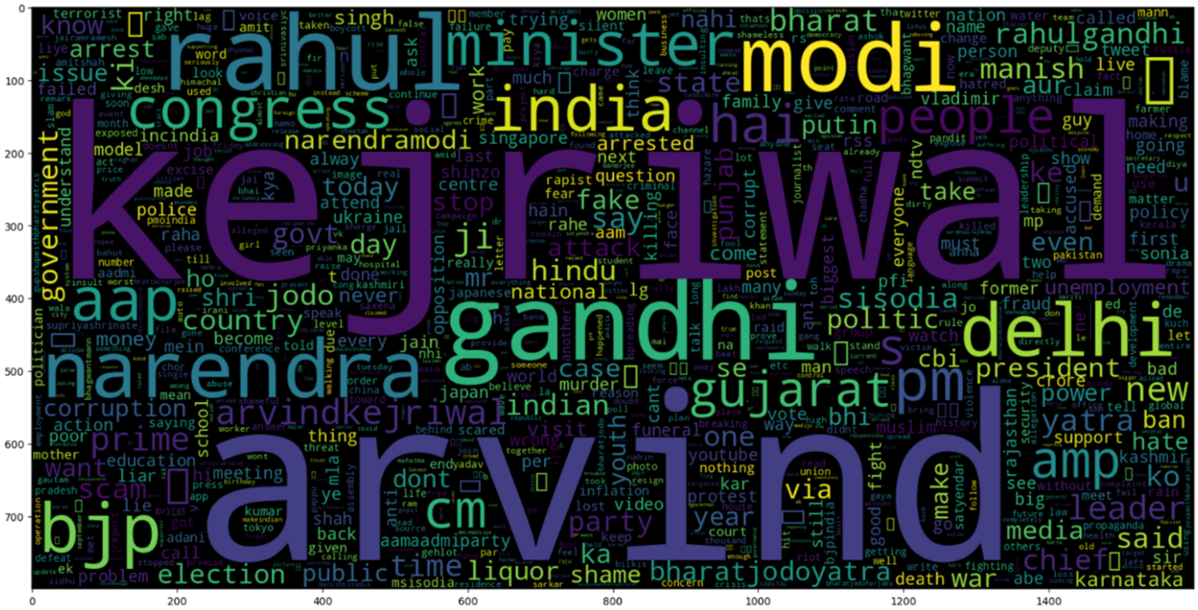

Figure 2: word clouds

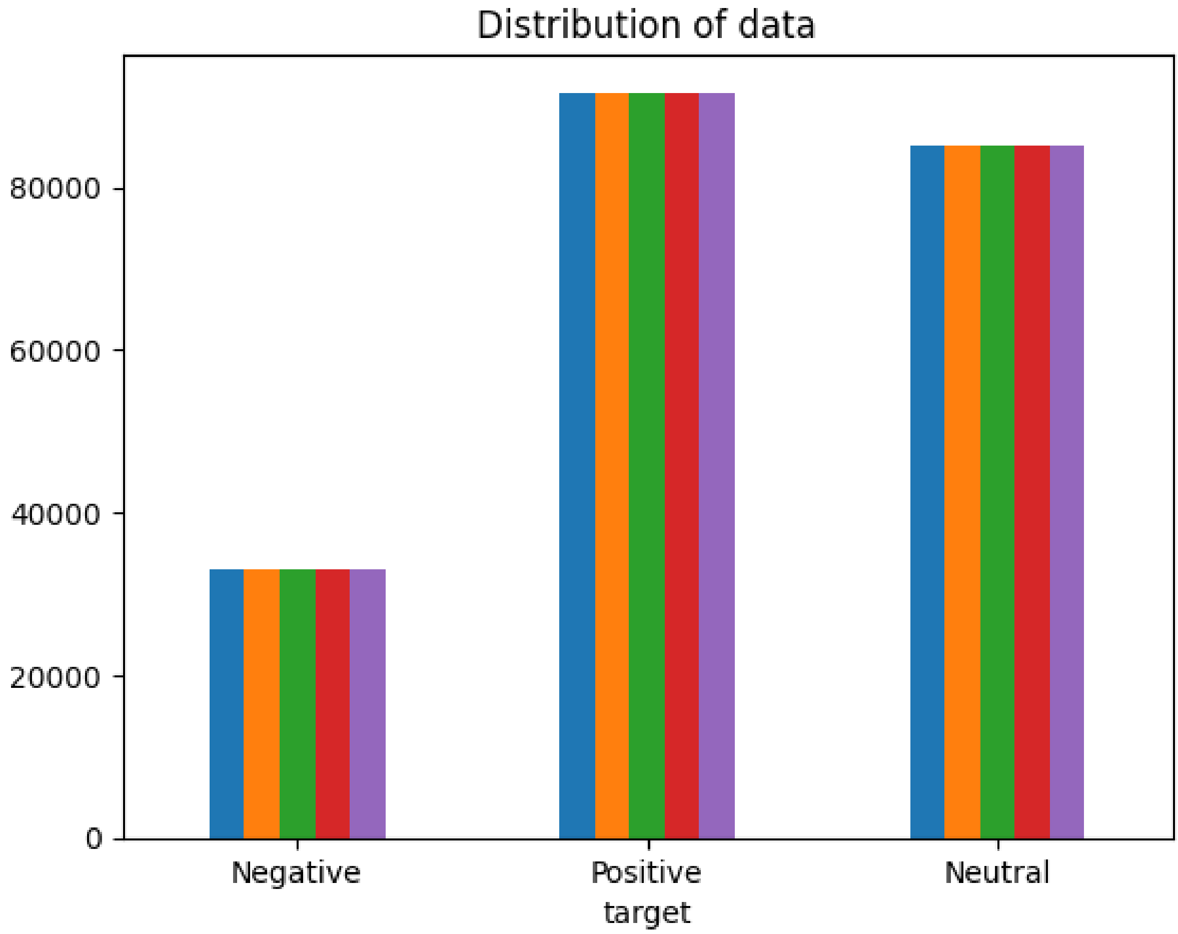

Figure 3: Distribution of data

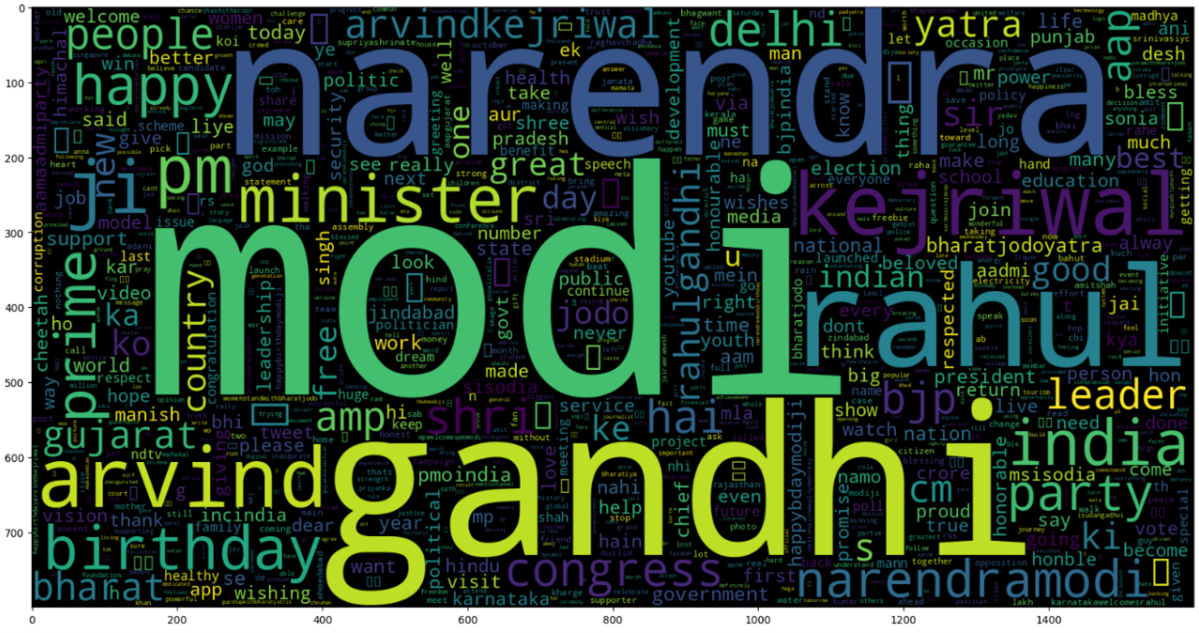

Figure 4: word clouds

```
import sklearn
from sklearn.feature_extraction.text import TfidfVectorizer

vectoriser = TfidfVectorizer(ngram_range=(1,2), max_features=500000)
vectoriser.fit(X_train)
print('No. of feature_words: ', len(vectoriser.get_feature_names_out()))
```

```
No. of feature_words:  276994
```

Figure 5: Feature Extraction

3.5.3 Neutral Tweets

The procedures adopted here enabled one to find out what sort of expressions are frequently referenced when such posts lack a strong leaning towards either end of the spectrum i.e., too much positivity or extreme negativity.

3.6 Feature Extraction

The textual data was converted to numerical features using the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer. This technique produces an importance score for each word in a document by assigning higher weights to infrequent word occurrences in specific document contexts. this makes sure that common words do not have much weight while rare ones that are meaningful can easily pass these thresholds hence contributing more significantly during the analysis and interpretation of results. From the words, the author came up with 276,994 significant features. These features are vital for training the model and evaluating performance metrics like the confusion matrix and ROC curve, thus ensuring a robust and accurate sentiment analysis system.(Figure 5: Feature Extraction)

3.7 Development of Interactive Dashboard

The interactive dashboard was designed to provide a user-friendly interface for exploring the sentiment analysis results by using Power BI. (Figure 7: Dashboard Screenshot)

4 Design Specification

This chapter gives an overview of the design characteristics and components of the project, including the system design and data flow, these designs act as a basic blueprint for the system implementation work package largely due to demonstrating how various components will fit in the overall system. The architecture of the real-time sentiment analysis project is believed to integrate and link all components for the purpose of organized data acquisition, computation, sentiment analysis, and display. The data flow process starts from the Data Collection where all the tweets concerning any of the political personalities (Arvind Kejriwal, Narendra Modi, and Rahul Gandhi) are harvested from the Twitter platform by accessing the official Twitter API. The obtained data is conditioned to a structured form for further work with it.

The second stage identified is the Data Preprocessing stage where cleaning of the raw tweet data is done. The removal of URLs, mentions, hashtags, special characters, and stop words, then tokenization, and finally lemmatization are actions done in this step to prepare the data for sentiment analysis. Subsequently, the cleaner data is stored so that it can be used for analysis at a later time.

In the Sentiment Analysis phase, two techniques are employed: Namely, VADER (valence-aware dictionary and sentiment Reasoner) and a Support Vector Machine (SVM) model. VADER gives initial polarities which are the base to apply the SVM model on the set of tweets with their associated labels. This trained SVM model then offers the sentiment labels on the new tweets.

Next, in Model Training and Evaluation, the data set is divided into training and the testing dataset, the model is trained using the SVM model and evaluated with parameters like accuracy, precision, recall, and AUC. This way the accuracy of the model is guaranteed and the model is ready to achieve a good sentiment classification.

Next, the Dashboard Development and Deployment stages are relevant to the concept of developing and managing the interactive dashboard the analyzing the sentiment. The real-time dashboard of the product is created with Power BI with components such as trends, real-time updates, and sentiment distribution graphs. Their architecture makes sure there is a constant stream of data from the collection, processing, and visualization aspects to the users, giving the public sentiment insights. (Figure 6: System Architecture and Data Flow Chart)

5 Implementation

The implementation phase is concerned with converting the design specifications into a working system. In this chapter, the authors describe the procedure of the interactive dashboard and real-time sentiment analysis system. Among the activities within the implementation are establishing data collection, preparing the data, conducting the sentiment analysis, Data management, and dashboard.

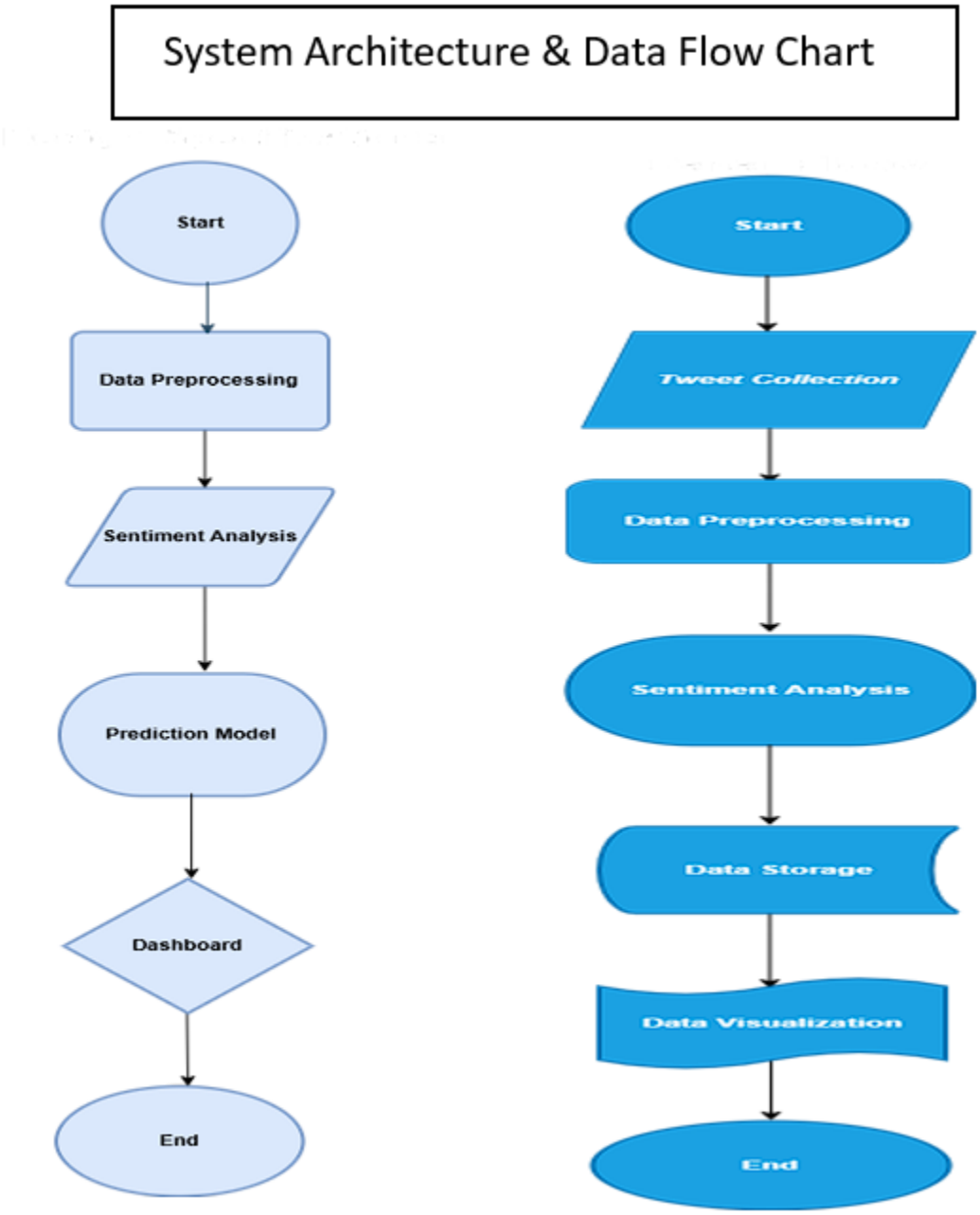

Figure 6: System Architecture and Data Flow Chart

5.1 Data Collection

5.1.1 Setting up the Twitter API

The first procedure in data collection is to establish a connection with the Twitter API. This entails applying to get the Twitter Developer account so that the API key and tokens needed can be accessed. With these credentials, the Tweepy library in Python then allows for authenticating the request for a tweet in real-time.

5.1.2 Collecting Tweets

Tweets are collected based on specific keywords related to the political figures in our study: Arvind Kejriwal is one of the prominent leaders; Narendra Modi is the current Prime Minister of India; Rahul Gandhi is an example of a young politician. Using Tweepy's streaming API the tweets that contain these keywords are collected. This makes it possible to collect the tweet data on a continuous and real-time basis to capture the most recent and relevant public sentiment data.

5.2 Data Preprocessing

Data pre-processing is crucial so that the raw tweet data contains no redundant information and is ready for sentiment analysis. The preprocessing steps include:

5.2.1 Cleaning the Data

The acquired tweets are preprocessed to ensure that the irrelevant information is removed from them. This includes:

- Removing URLs Deleting them to make it easier for the text not to have any connection to other works.
- Removing Special Characters Refining the text by removing any punctuation mark or symbol that is unrelated to sentiment analysis.
- Removing Stopwords Ignoring stop-words/standard words (e.g., "and", "the", "is", etc.) having no sentiment value.

5.3 Sentiment Analysis

The cleaned tweets are analyzed to determine their sentiment using two approaches: developed in this paper VADER and a Support Vector Machine (SVM).

5.3.1 VADER Sentiment Analysis

VADER is a sentiment analysis tool based on affective words' lexical database and rules to classify emotions that are especially characteristic of social media. It assigns an integrated value to every single tweet, which reveals the sentiment polarity.

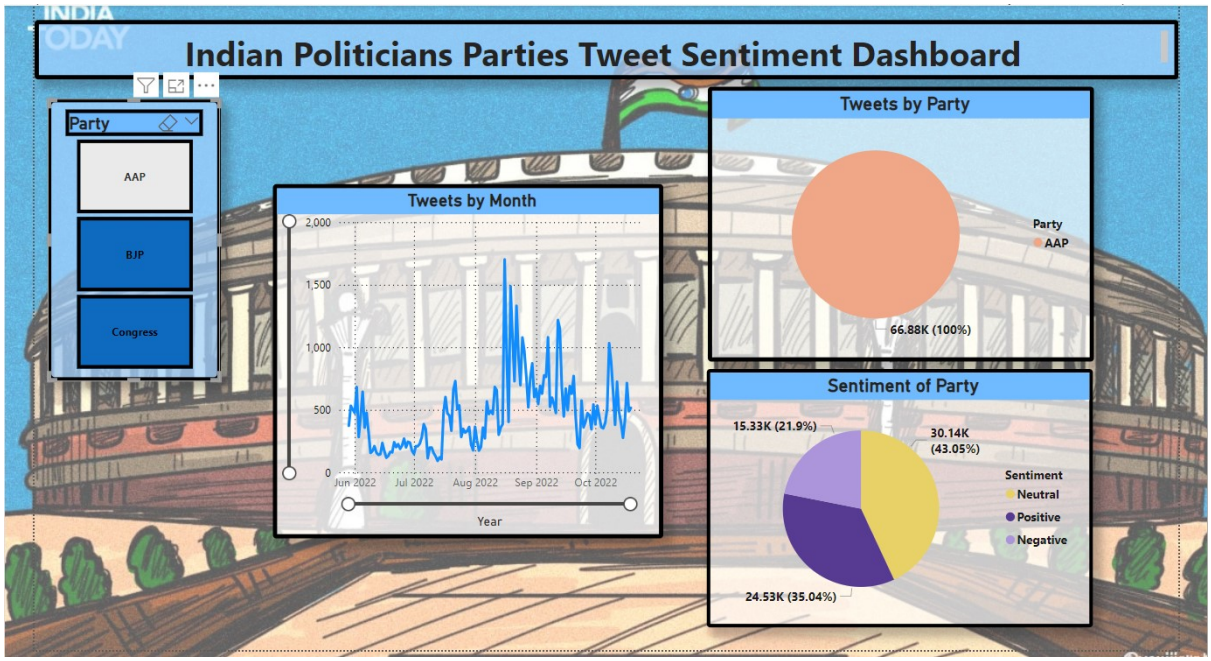

Figure 7: Dashboard Screenshot

5.3.2 SVM Model

To further improve the information provided by the sentiment analysis, a machine-learning implementation of the SVM algorithm is introduced. The steps include.

- Vectorization: Text data preprocessing – Use of TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer to transform the text data into numerical format.
- Model Training: Training the SVM model on the training dataset in which every single tweet is pre-labeled either as positive, negative, or neutral.
- Prediction: To predict the sentiment of the new tweets which has not been part of the training data set Using the trained model.

5.4 Dashboard Development

The interactive dashboard is created using Power BI which gives the user a live visual of sentiments. A Dashboard is developed corresponding to this front-end interface. The visualization of the dashboard concept entails the determination of formats such as Graphs, charts, and real-time updates on the dashboard layout. (Figure 7: Dashboard Screenshot)

6 Evaluation

The evaluation phase is an integral part of the development of any system because it checks whether the introduced solution satisfies the objectives of the project and effectively works. The SVM classifying model's performance on sentiment analysis can be summed up by many measures such as confusion matrix, classification report, and ROC-AUC curve. The high precision, recall, and F1-scores obtained in all categories of

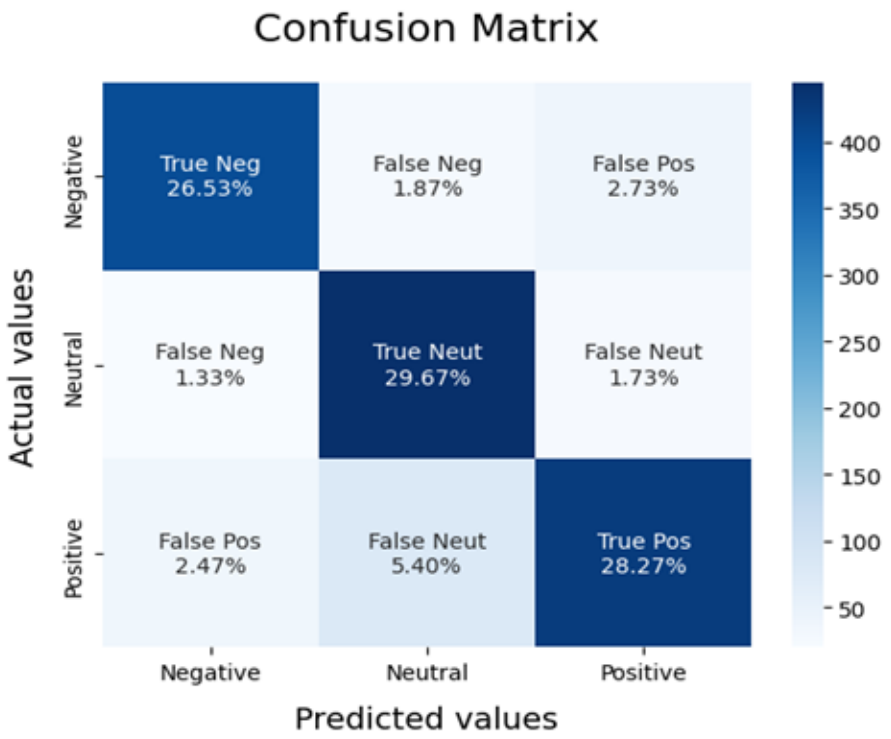

Figure 8: Confusion Matrix

sentiments and a strong AUC value indicate that this model classifies tweets in a correct way into positive, negative, or neutral. In other words, it shows its efficiency in analyzing and interpreting sentiments found in social network data.

6.1 Confusion Matrix

The confusion matrix gives the complete picture of the model accuracy by presenting the values of true positive, true negative, false positive, and false negative of the given classes. For the Negative class, the model gets 26.53% of the time with a small fraction being classified as Neutral with 1. 87% and Positive at 2. 73%. As for the Neutral class, the proposed model's result gives an excellent performance to identify as true neutral with 29.67% with some that are classified as Negative by 1.33% and Positive by 1.73%. The Positive class raises more issues with a true positive rate of 28.27%, but it has higher misclassification rates with 2.47%, and 5.40% of the positive instances are being wrongly classified as Negative and Neutral. Based on the results derived from the confusion matrix the following observation can be made. The model performed at its best in identifying the Neutral instances as reflected by the highest True Neutral rate of 29.67%. However, it performs poorly with the Positive class wherein the highest false positive rate is observed when the Positive instances are classified as Neutral at 5.40%. This means that, although the model is quite good at classifying Negative and Neutral instances, the situation with Neutral and Positive classes still requires further enhancement of the model. (Figure 6: Confusion Matrix)

	precision	recall	f1-score	support
Negative	0.87	0.85	0.86	467
Neutral	0.80	0.91	0.85	491
Positive	0.86	0.78	0.82	542
accuracy			0.84	1500
macro avg	0.85	0.85	0.85	1500
weighted avg	0.85	0.84	0.84	1500

Figure 9: Classification Report

6.2 Classification Report

The classification report provides a detailed breakdown of the model's performance across three classes Negative, Neutral, and Positive. Precision for the Negative class is 0.87, showing that of the Total number of instances predicted as Negative, 87 of them are Negative. The recall for this class is 0.85, which shows that out of 100 actual Negative cases, 85 are correctly identified by the model. The F1-score of 0.86, which is a balanced considered using the harmonic means of the calculation of both precision and recall. but the recall rate is 0.91 showing that the model is more effective in predicting the Neutral instances. The positive class has a precision of 0. 86. which is a recall of 0.78 summarizing, the proposed approach achieved an accuracy of 78%, and an F1-score of 0.82, This proves that although the model differentiates the classes with a fairly good degree of accuracy, it fails to capture a much greater percentage of Positive instances compared to Negative and Neutral classes.

The score of accuracy of the model is equal to 0.84 It is interpreted to mean that 84 out of the total predictions are accurate. which is the unweighted average of precision, recall, and F-1 score, is 0.85 It appears to be 85 in all the parameters proving that it has done well in all the classes for instance each class is incorporated in the average performance figures depicted under precision, recall, and F1 scores are also fairly high and move in the range of 0.84 to 0.85. These averages show that this model is good and will work well for all classes, even though the recall for the Positive class could be improved.(Figure 7: Classification Report)

6.3 ROC Curve

The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) for the classifier between two classes. the AUC curve is 0.86, which shows that the separability is quite high. This level of separation is in good measure between the positive class and the negative class. the model will have about 86% probability of sorting right between the two.

The ROC curve demonstrates the true positive rate is high even when the false positive rate is low, which indicates good performance. An AUC of 0.86 means that this model is significantly better than a model that just randomly guesses, which in turn means that

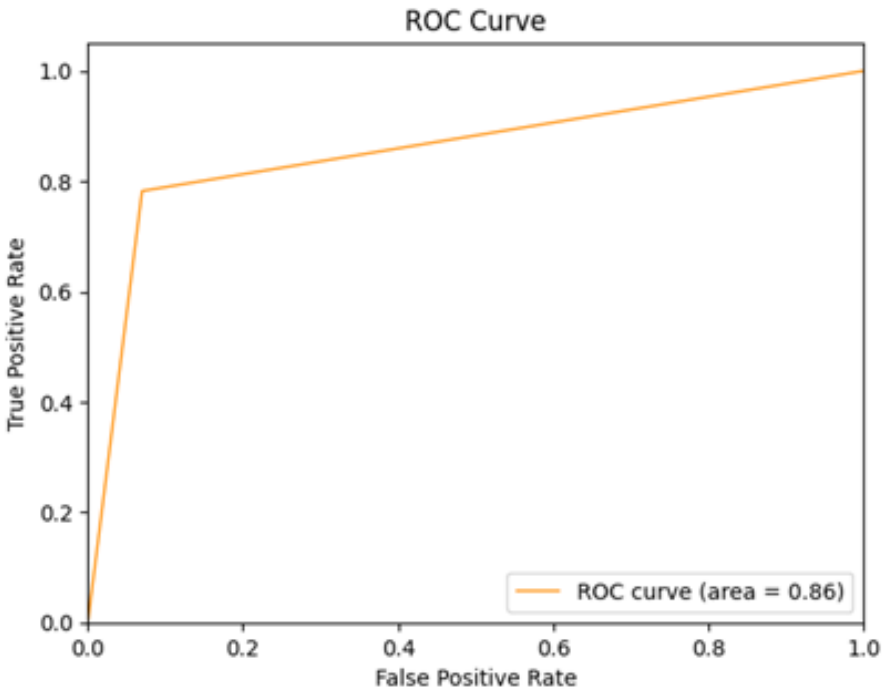

Figure 10: ROC Curve

the AUC of 0.5. would indicate no discriminative ability. The AUC value is closer to 1, the model is better, and an AUC of 0.86 shows strong model performance. (Figure 8: ROC Curve)

The results of the classification report, the confusion matrix, and the ROC curve collectively indicate that the model performs well overall, especially in the Neutral instances. Nevertheless, there is a notable area for improvement in the recall of the Positive class and reducing the misclassification between Neutral and Positive instances. The ROC curve also similarly shows that the model is remarkably accurate with an AUC of 0.86, proving a very high discriminative ability. Therefore, the above findings can be used in the formulation of successive improvements to refine and improve the model's precision and dependability.

7 Conclusion and Future Work

In conclusion, this research addressed the problem statement and developed a real-time sentiment analysis system and an interactive dashboard for tracking the social media discourse about leading Indian political personalities. The sentiments were accurately classified using the VADER tool and the support vector machine model and for the given data, the high discriminant ability was observed with the AUC of 0.86. However, there are still opportunities to further enhance the model with better training in the positive sentiment class and decreasing the tendency of misclassification of instances into the classes of neutrality and positivity. Future research should aim at extending the validity and accuracy of the existing model, which could be achieved through improved algorithms of the ML method or by including more variables in the analysis. In addition, it can be redeployed and monetized as a brand analytics tool for real-time monitoring of political sentiment, which may benefit a political campaign, media partners, or the social sciences.

Here, several significant trends for future research can be identified, including cross-lingual sentiment analysis and the effects of multimedia content on sentiment distributions.

References

Agarwal, M., Chaudhary, P. K., Singh, S. K. and Vij, C. (2023). Sentiment analysis dashboard for socia media comments using bert, *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, IEEE, pp. 292–298.

Albaldawi, W. S. and Almuttairi, R. M. (2020). Comparative study of classification algorithms to analyze and predict a twitter sentiment in apache spark, *IOP Conference Series: Materials Science and Engineering*, Vol. 928, IOP Publishing, p. 032045.

Albayrak, M. D. and Gray-Roncal, W. (2019). Data mining and sentiment analysis of real-time twitter messages for monitoring and predicting events, *2019 IEEE Integrated STEM Education Conference (ISEC)*, IEEE, pp. 42–43.

Center, P. R. (2020). Americans, politics and social media. Accessed: 2024-07-25.
URL: *https://www.pewresearch.org/internet/2020/07/29/americans-politics-and-social-media/*

Commission, E. (2020). Social media influences our political behaviour and puts pressure on our democracies, new report finds. Accessed: 2024-07-25.
URL: *https://joint-research-centre.ec.europa.eu/publications/social-media-influences-our-political-behaviour-and-puts-pressure-our-democracies-new-report-finds$_e$n*

D'Avanzo, E., Pilato, G. and Lytras, M. (2017). Using twitter sentiment and emotions analysis of google trends for decisions making, *Program* **51**(3): 322–350.

Intyaswati, D. and Fairuzza, M. T. (2023). The influence of social media on online political participation among college students: Mediation of political talks, *Southern Communication Journal* **88**(3): 257–265.

Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., Mittal, A. et al. (2020). Social media analysis with ai: sentiment analysis techniques for the analysis of twitter covid-19 data, *J. Crit. Rev* **7**(9): 2761–2774.

Martinez, L. S., Tsou, M.-H. and Spitzberg, B. H. (2019). A case study in belief surveillance, sentiment analysis, and identification of informational targets for e-cigarettes interventions, *Proceedings of the 10th International Conference on Social Media and Society*, pp. 15–23.

Mathew, J. K., Deepika, R., Sharanyaa, S., Therasa, M. et al. (2022). Live streaming data analysis using distributed stochastic bi-lstm model, *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, IEEE, pp. 1–4.

Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S. and Mridha, M. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm, *Scientific Reports* **14**(1): 9603.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors, *Proceedings of the 19th international conference on World wide web*, pp. 851–860.

Singla, Z., Randhawa, S. and Jain, S. (2017). Sentiment analysis of customer product reviews using machine learning, *2017 International Conference on Intelligent Computing and Control (I2C2)*, pp. 1–5.

Real-Time Sentiment Analysis and Interactive Dashboard Deployment for Social Media Discourse

MSc Research Project

Data Analytics

Jaiprakash Chandraker

Student ID: 22213546

School of Computing

National College of Ireland

Supervisor: Vladimir Milosavljevic

Q1. What is the state of the art in this specific task? It's critical for research and could justify your choice of a particular modelling approach.

Answer:

The state of the art in sentiment analysis, particularly in the context of social media data like Twitter, includes several advanced techniques and models. Traditional methods involve lexicon-based approaches, such as VADER (valence-aware dictionary and sentiment Reasoner), which are effective for general sentiment analysis tasks. However, more recent advancements have seen the use of machine learning models, such as Support Vector Machines (SVM), and deep learning models, such as Recurrent Neural Networks (RNN) and Transformer-based models like BERT (Bidirectional Encoder Representations from Transformers).

The choice of a particular modeling approach often depends on the balance between computational efficiency and the accuracy required for the task. For example, while deep learning models like BERT offer high accuracy, they require significant computational resources and large datasets. In contrast, SVMs are a robust choice for smaller datasets and can offer competitive performance with less computational overhead. Given these considerations, SVM is a well-established method for sentiment classification, offering a good trade-off between accuracy and resource consumption, making it suitable for the real-time sentiment analysis task in this project.

Q2. Why did you choose SVM over other modelling approaches?

Answer:

The choice of Support Vector Machine (SVM) over other modeling approaches was likely driven by several factors. SVM is known for its effectiveness in binary and multi-class classification problems, particularly when the feature space is not overly complex. It performs well on smaller datasets, which is often the case with specific-topic sentiment analysis like the one conducted in this project.

Additionally, SVMs are less prone to overfitting compared to more complex models like deep neural networks, especially when the dataset is limited in size. The ability to handle high-dimensional spaces and the margin-based classification approach of SVMs can result in good generalization performance. This makes SVM a suitable choice for the sentiment analysis task, where the goal is to accurately classify tweets as positive, negative, or neutral without requiring the extensive computational resources associated with deep learning models.

Q3. What sort of hyperparameter tuning was done for the SVM? If not, why?

Answer:

I did not perform specific hyperparameter tuning for the SVM model in this project. Because the default parameters were found to be sufficient for achieving satisfactory performance, nevertheless, it is generally recommended to conduct hyperparameter tuning using techniques such as grid search or cross-validation to ensure that the SVM model is performing optimally.

However, in typical SVM implementations for sentiment analysis, hyperparameter tuning is crucial for optimizing model performance. Common hyperparameters to tune include the regularization parameter C, which controls the trade-off between maximizing the margin and minimizing classification error, and the kernel type (e.g., linear, polynomial, radial basis function) which affects how the decision boundary is formed.

Q4. Data collection methodology: Besides the name keywords, what other hashtags were used to collect this dataset? Was it collected during a specific time period?

Answer:

For the data collection phase of this project, the primary focus was on capturing tweets that mentioned the names of prominent Indian political figures Arvind Kejriwal, Narendra Modi, and Rahul Gandhi—as keywords to filter tweets (Twitter (X) handler @narendramodi, @ArvindKejriwal, @RahulGandhi). While collecting data I used hashtags, hashtags like #Elections2024, #BJP, #INCIndia, #AAP,

Although the dataset was not initially constrained by a specific time frame during the collection process, for the purposes of this analysis, only data from the year 2022 was utilized. This decision was made to maintain temporal relevance and ensure that the sentiment analysis accurately reflected the public discourse surrounding these figures during a period that was most pertinent to the project's objectives. This temporal focus on 2022 was deemed appropriate due to the significant political events and discussions during that year, which likely influenced public sentiment and made the data especially relevant for analysis.