TOPIC-BASED ANALYSIS OF SOCIAL MEDIA POSTS USING RNN AND LSTM

ZHU QIAN

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/23)

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name :

Student's Matric No. : Academic Session :

Date of Birth : UTM Email :

Choose an item. Title : TITLE IN CAPITAL LETTERS
TITLE IN CAPITAL LETTERS
TITLE IN CAPITAL LETTERS

I declare that this Choose an item. is classified as:

□ **OPEN ACCESS** I agree that my report to be published as a hard copy or made available through online open access.

□ **RESTRICTED** Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☒ **CONFIDENTIAL** Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name
Date :

Approved by Supervisor(s)

Signature of Supervisor I:

Full Name of Supervisor I
NOOR HAZARINA HASHIM

Date :

Signature of Supervisor II

Full Name of Supervisor II
MOHD ZULI JAAFAR

Date :

TOPIC-BASED ANALYSIS OF SOCIAL MEDIA POSTS USING RNN AND LSTM

ZHU QIAN

A project report submitted in fulfilment of the
requirements for the award of the degree of
Master of Data Science

School of Computing
Faculty of Computing
Universiti Teknologi Malaysia

OCTOBER 2024

DECLARATION

I declare that this project report entitled *"Topic-Based Analysis of Social Media Posts Using RNN and LSTM"* is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	....
Name	:	
Date	:	17 JANUARY 2025

ACKNOWLEDGEMENT

In preparing this thesis, I am grateful to Professor Dr. Mohd Shariff Nabi Baksh for his support in providing guidance and methods. With the help of public scholarship and previous research, the study was able to start off on a strong foundation. Although the study faced unexpected challenges, with the support of the professor and classmates, the solutions were eventually found.

I am also indebted to Universiti Teknologi Malaysia (UTM) for supporting my study.

I would also like to express my sincere gratitude to the people who helped me on different occasions. Their comments and suggestions were really useful..

ABSTRACT

This study focuses on analysing public sentiment and reactions regarding the 2024 American election based on social media platforms. By examining the most influential platforms related to political posts, the study aims to extract real trends in voter behaviour during the election. With advanced analytical methods, unlike traditional social media analysis, this study introduces deep machine learning techniques and utilizes advanced artificial intelligence algorithms. Following rigorous data collection protocols from social media platforms, the study cleans the data and trains models to ensure high performance and accurate analysis. To achieve the objectives, the study incorporates the Support Vector Machine (SVM) algorithm, Support Vector Classifier (SVC) kernels, and the Term Frequency-Inverse Document Frequency (TF-IDF) matrix, building on previous research findings. Traditional media data analysis often suffers from biases and limitations, which can lead to negative responses from organizations and impact the ability to respond effectively to certain events. By improving the analysis system through a more robust research framework, this study aims to provide a more accurate reflection of public sentiment. This study uses Exploratory Data Analysis (EDA) to support the prediction model. It compares the expected output with the model's results in order to adjust the model's parameters. During this process, the structure of the development system is explored to find the balance between performance and accuracy. With the help of open-source libraries, the system generates charts from the model results, including distributions and word clouds. Visualization provides more intuitive insights into the relationships within the dataset for researchers. The study also presents initial insights and acknowledges its limitations, while the future work section offers recommendations for further development.

(Note: Students are allowed to use either single or one-and-a-half spacing for the abstract, as long as it fits within one page. The chosen spacing style must be consistent across both the English and Malay sections.)

ABSTRAK

Kajian ini memberi tumpuan kepada menganalisis sentimen dan reaksi orang ramai mengenai pilihan raya Amerika 2024 berdasarkan platform media sosial. Dengan meneliti platform paling berpengaruh yang berkaitan dengan jawatan politik, kajian ini bertujuan untuk mengekstrak arah aliran sebenar dalam tingkah laku pengundi semasa pilihan raya. Dengan kaedah analisis lanjutan, tidak seperti analisis media sosial tradisional, kajian ini memperkenalkan teknik pembelajaran mesin yang mendalam dan menggunakan algoritma kecerdasan buatan lanjutan. Mengikuti protokol pengumpulan data yang ketat dari platform media sosial, kajian itu membersihkan data dan melatih model untuk memastikan prestasi tinggi dan analisis yang tepat. Untuk mencapai objektif, kajian ini menggabungkan algoritma Mesin Vektor Sokongan (SVM), kernel Pengelas Vektor Sokongan (SVC), dan matriks Kekerapan Jangka-Sbalik Frekuensi Dokumen (TF-IDF), berdasarkan penemuan penyelidikan terdahulu. Analisis data media tradisional sering mengalami berat sebelah dan batasan, yang boleh membawa kepada tindak balas negatif daripada organisasi dan memberi kesan kepada keupayaan untuk bertindak balas secara berkesan kepada peristiwa tertentu. Dengan menambah baik sistem analisis melalui rangka kerja penyelidikan yang lebih mantap, kajian ini bertujuan untuk memberikan gambaran yang lebih tepat tentang sentimen awam. Kajian ini menggunakan Analisis Data Penerokaan (EDA) untuk menyokong model ramalan. Ia membandingkan output yang dijangkakan dengan keputusan model untuk melaraskan parameter model. Semasa proses ini, struktur sistem pembangunan diterokai untuk mencari keseimbangan antara prestasi dan ketepatan. Dengan bantuan perpustakaan sumber terbuka, sistem menjana carta daripada hasil model, termasuk pengedaran dan awan perkataan. Visualisasi memberikan cerapan yang lebih intuitif tentang hubungan dalam set data untuk penyelidik. Kajian ini juga membentangkan pandangan awal dan mengakui batasannya, manakala bahagian kerja masa depan menawarkan cadangan untuk pembangunan selanjutnya.

(Note: Students are allowed to use either single or one-and-a-half spacing for the abstract, as long as it fits within one page. The chosen spacing style must be consistent across both the English and Malay sections.)

TABLE OF CONTENTS

LIST OF TABLES

TABLE NO.	TITLE	PAGE

LIST OF FIGURES

FIGURE NO. **TITLE** **PAGE**

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
GA	-	Genetic Algorithm
PSO	-	Particle Swarm Optimization
MTS	-	Mahalanobis Taguchi System
MD	-	Mahalanobis Distance
TM	-	Taguchi Method
UTM	-	Universiti Teknologi Malaysia
XML	-	Extensible Markup Language
ANN	-	Artificial Neural Network
GA	-	Genetic Algorithm
PSO	-	Particle Swarm Optimization

LIST OF SYMBOLS

δ	-	Minimal error
D, d	-	Diameter
F	-	Force
v	-	Velocity
p	-	Pressure
I	-	Moment of Inersia
r	-	Radius
Re	-	Reynold Number

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Unlike traditional media, social media serves as a more open platform. However, the rise of fake users and internet trolls has become a significant issue that threatens data quality. The platform is often inundated with irrelevant data, while the internet is filled with non-compliant, emotional, repetitive, and meaningless content. Furthermore, some users distort their own views or misrepresent the perspectives of others.

Meanwhile, web surveys struggle to verify whether respondents are real individuals, and the results are easy to tamper with. In contrast, social media provides a more reliable and genuine source of data.

Real public reactions to various topics provide insights that can help enterprises and governments make more informed decisions in response to real-world changes and challenges. At the same time, these organizations can reduce costs associated with traditional questionnaires, benefiting their financial standing.

1.2 Problem Background

The weaknesses of the current working model highlight the need for new data sources and a more efficient, sophisticated approach. Utilizing data science and machine learning techniques to analyze topic-based social media posts will help enterprises and governments identify valuable data to address these challenges.

1.3 Problem Statement

Social media posts are presented in various data formats such as text and emojis often mixed with irrelevant content and influenced by emotions. Traditional survey methods frequently produce false or unreliable results, which often deviate significantly from the actual situation.

1.4 Research Goal

Collecting and filtering social media X posts by topic, and using machine learning techniques to analysis them, yields valuable and authentic data about the topic.

1.4.1 Research Objectives

The objectives of the research are:

(a) To identify significant relationships between the content of posts and the topic.

(b) To build and develop analytical models that capture the topic inclination of posts.

(c) To measure public reactions to the topic by summarizing the analysis results.

(d) To define the best parameter estimate.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

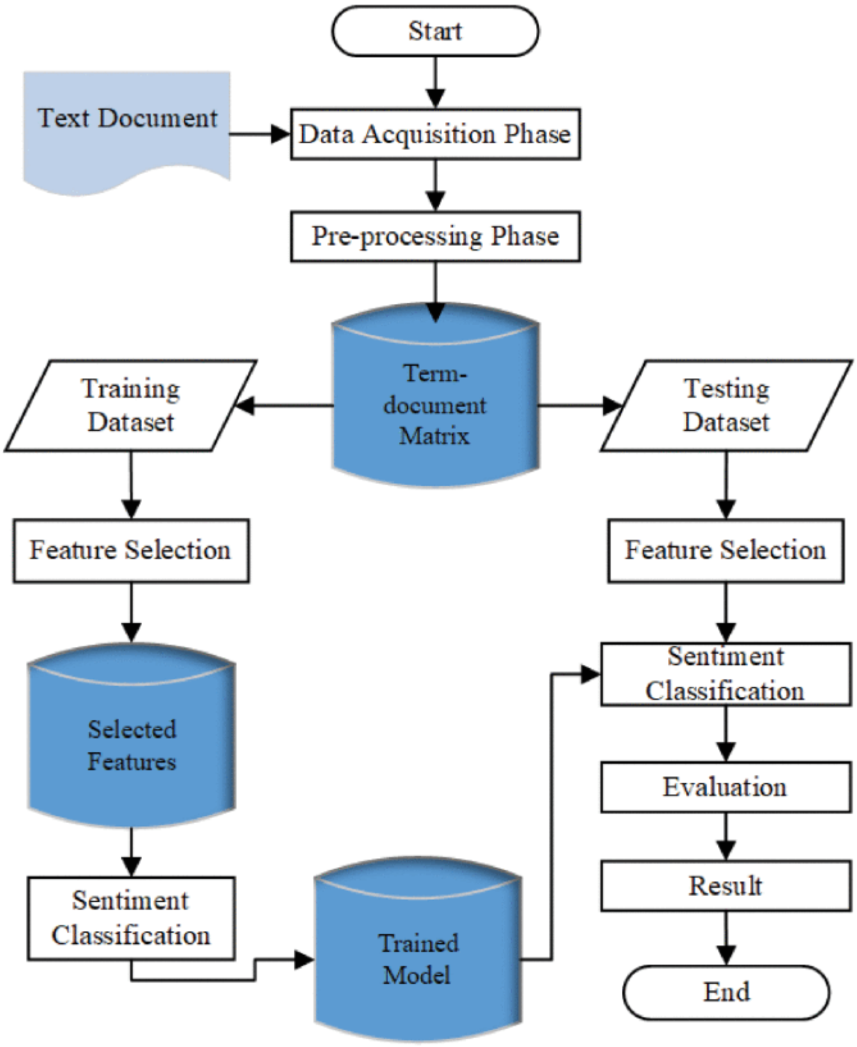
This chapter will review historic methods and challenges involved in analysing social media data. Secondly, it examines how new technologies, and here we are talking particularly about the use of deep learning, could advance the quality and speed of analytical processes. The key area observed in this chapter is the (LSTM) memory network, which has become one of the most broadly employed strategies for investigating several kinds of social media content. Lastly, the 45th section talks about the present situation and future framework that has to do with LSTM applications.

2.2 Traditional Approaches to Text and Topic Analysis

As the popularity of monitoring the contents of social media data has increased drastically in recent years, the data sources have made a leap from solely text-oriented details to a myriad of data types including text, image, media, emojis, sound, hashtag, and mixed data. In the first stage, researchers designed the perfect means to apply social network data and, consequently, extract valuable knowledge from them into a structured format. The main difficulty lied in the huge and changing scale of data sources. Some of the techniques and models developed by the researchers to solve these issues are as follows.

The most popular approaches include the Bag-of-Words (BoW) model, Latent Dirichlet Allocation (LDA), and Term Frequency-Inverse Document Frequency (TF-IDF), which takes on more sophisticated natural language processing (NLP) algorithms. Firstly, a model is created for the BoW analysis of the text through introducing the technique called 'Bag-of-Words'. The technique is considered one of the simplest and the most frequently applied ones because other text analysis methods were later derived from it. BoW is an approach to text representation that views text as a group of just individual words, ignoring textual structure and how frequently a word occurs in the original content. In this manner, it can glean semantic and syntactic relations among the most pertinent terms within a document. However, the model has its own imperfections and faces such challenges when it comes to dealing with actual social media data.

Here we are not discussing a social media content composed of separate sentences, but which in turn is made up of both useful and senseless responses from different users/persons. In the same manner, transforming paragraphs into bags of words is also an insurmountable task, and in many situations, even impossible. To adjust such novel issues as a result of real life challenges, an innovative technique was developed by the researchers, which is represented by the novel text analysis: Term Frequency-Inverse Document Frequency (TF-IDF). A basic method known as TF-IDF is undertaken where the frequency of terms in a document is calculated that also takes into account of their relative importance in the whole corpus.



To conclude, there is a difference between BOG and TF-IDF models. In the BOG model, the frequency of meaningful words is not considered, whereas in TF-IDF model, the frequency of meaningful words is considered as well as the way they are distributed throughout the corpus, which makes the representation of the text more precise. Such an increase overcomes the restrictions which the bag-of-words model represents, hence the analyst can do more than just the document by document correlation of the text, bring up a more nuanced and better-internalized understanding of the issues under consideration.

2.3 Topic-Based Analysis Using Deep Learning

The analysis of TF-IDF keywords is a powerful tool for the extraction of dominant words in the text based on social media. It helps not only for tracking the most popular topics but also the significant keywords. Therefore, it is widely used in social networks monitoring systems. Similarly to BoW, the other approach, TF-IDF has its own limitations. Indeed, the BoW model is merely superficial, and it involves no deeper insights in context, so at times, it does not manage to appreciate data which could be very important. In spite of the fact that TF-IDF can be more effective than BoW in other terms, on the contrary, it shows lower accuracy, which is an obstacle for forecasting similar trends.

Social media have raised the necessity for studying text analysis, and this is a need that has been recognized by researchers. However, they do realize that currently text analysis will not be sufficient to keep up with rapidly changing social media that is developing both in complexity and structure. These developments in techniques, key of which are the deep learning, have been quickly employed in up-to-date identification platforms. Meanwhile, social media metadata is becoming more and more intricate, providing a pattern of not only texts but also images, hashtags, emojis, and other multimedia, and so this one can be regarded as a classed data. This transition means that standard methods of data analysis are now obsolete, and we have to come up with unique ways to process such data. It was pointed out at the previous part of this chapter that a more standard model, such as the Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), is unsuited for a variety of data types. Despite this, original texts are still analyzing, but the process is time-consuming, leading to resource use.

The constraints of traditional methods remain a problem, but new analytics methods like BoW (Bag-of-Words) and TF-IDF (Term Frequency-Inverse Document Frequency) contribute much to research and represent the primary tools of basic research. Current approaches, particularly the emergence of deep learning methods, revolve around the issue of handling noise and filtering out the error and detecting UN-informal content from social media. Their goal is to help guide humans to consume small passages rather than big chunks of meaningless text. Therefore, using structured and formal data will be of better quality, which are prone to the principles of quantitative analysis in the research platform. This system prevents the loss of meaning to a certain level and makes the allocation of related topics much easier.

The information extraction models of deep learning perform better on understanding deep subjects and estranged cases than plain methods of learning. Nowadays, the most critical issue is the design of tools to resolve intricate and untraveled tasks in the area of social media metadata. The two techniques that have turned out as the most vital approaches in the area of Deep Learning for this task are Recurrent Neural Networks (RNNs) and Long-Short Term Memory (LSTM) networks. Science, as it was introduced in this chapter, REVEALING of meaningful relationships within diverse contexts remains a significant challenge for any analytical method. Memory storage, computing function, and the ability to write analyses on the fly all are the essential elements of this system. Improved performing skills in terms of memory, processing, and editing of data are several of the pros of the Graphics Processing Units (GPUs) that have been developed.

With these Optimizations now Graphics Processors (using Deep Learning techniques as their core technology) became highly efficient and versatile platforms. Deep learning functions deal with words partnerships, full sentences, complete documents, bonds that reach to deep levels in the data. The most important trainable feature of deep learning is the ability the irrelevant and non-meaningful data that are made usually in this type of media. Looking to the future, the deep learning strategies will be able to learn complex hierarchical features from theses raw files, Bringing sequence A more precise requirement for enterprises employing online social media.

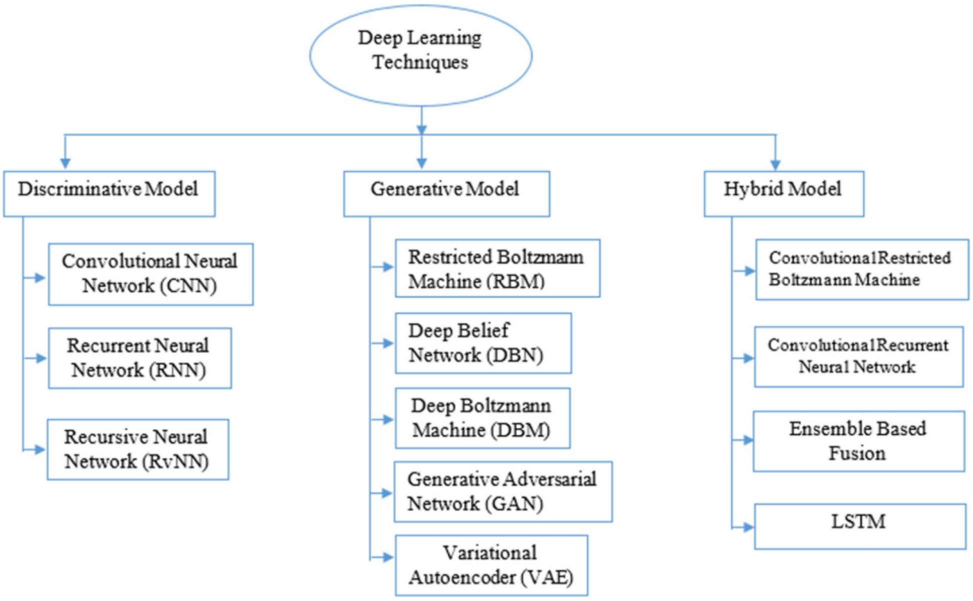
This leads to the inference that the better the functioning of the system, the more accurate its results will be. This is a remarkable extension of routine data processing methods. For the topic-based analysis, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) work better in most cases. The primary role of CNNs is the extraction for the purpose of features out of text. These systems have been used in this area and have widely applied. They stand out in the sense of identifying factors distinctive to a certain topic. In contrast, RNNs, especially LSTMs, are mainly used to discern meaning out of running text as they focus on capturing sequential connections useful for maintaining context in social media postings. The method used shall be designed so that researchers can look into different types of information in order to study them.

2.4 Introduction to Long Short-Term Memory Networks (LSTMs)

There is a variety of methods of categorization that permit choosing the appropriate one depending on the specifics of the data. Here's the section that's going to deliver a presentation of Long Short-Term Memory (LSTM) networks that are being among the most recent applications for topic-based social media recognition systems.

LSTMs present an advantage in terms that they are prudent and all important tools in creating very dependable modulation frameworks. In 1997, "The analysis of Long Short-Term Memory (LSTM) networks" came to be. Despite the obstacles which the hardware technology presented at that time, LSTMs were rather regarded as a theoretical concept but in a very little way as LSTMs commenced to advance quickly and were being applied in the field to effect real-world application. The period extended in years 2015 to 2016, and I can witness that AI or artificial intelligence found its way to the fore most and paved the way to its wide acceptance when, advantages in hardware, particularly Graphics Processing units (GPUs), enabled the bursting into the scene of AI, specifically deep learning.

Thus, in the course of developing AI, deep learning eventually achieved the computational ability necessary for it to be used with complex systems. This approach is now applied in real-world areas. Concurrent to this, Google made its own unique Artificial Intelligence (AI) platform which aimed at handling challenges like the parallel computing in terms of software. Google researchers showed significant improvements through their work on deep neural networks (DNNs), which were crucial for progress in improving LSTM networks. This undertaking was a joint effort with Jeff Dean and, as a result, many goals were achieved. Combining graphics technology processing units - GPUs with deep learning technology, Google sped up procedure related to these tasks. This result is encouraging as the new method is cosmodrome very far behind conventional systems in their speed. These technique improvements enabled Google to derive effective algorithms that could learn highly sophisticated representations of input data sets and thus it made significant contributions to the development of Artificial Intelligence applications.


Deep Learning Techniques
Discriminative Model
Generative Model
Hybrid Model
Convolutional Neural Network (CNN)
Recurrent Neural Network (RNN)
Recursive Neural Network (RvNN)
Restricted Boltzmann Machine (RBM)
Deep Belief Network (DBN)
Deep Boltzmann Machine (DBM)
Generative Adversarial Network (GAN)
Variational Autoencoder (VAE)
Convolutional Restricted Boltzmann Machine
Convolutional Recurrent Neural Network
Ensemble Based Fusion
LSTM

The work that has been carried out to develop on these prior achieved levels makes LSTMs much more feasible to be implemented to deal with some challenging issues that have been specified for the use case. When you come to the LSTM basis, the architecture is divided into four major parts includes the cell state, forget, input, and output gates. These constituents provide deep learning systems with the basic functionality necessary to meet their objectives.

The Cell State is vital in the information capture and preservation of the input data sequences long-run association. It acts as a memory module for the network, which keeps the knowledge over prolonged time by storing and analyzing the relationships between interconnections of a vast number of contexts. Apart from that, it ensures steady data flow through the system and eliminates the inaccurate. Such an ability allows the unit to handle complicated patterns, not only those shown above but also incoming streams from natural language, for instance.

Yet social media does not consist solely of natural language, but a large proportion of its volume is still based on text that may be translated into a natural language form. The crucial impact of Cell State is to make the model more descriptive, in only a single word to clarify more and have higher precision. Components of the doors except of the gates are relatively easier to interpret. Three gates: Forget, Input, and Output provide a decision at the current time step based on the hidden state at the previous time step as well as the target. The rectifying connections, with different activation functions, calculate the coefficients for the Forget, input, and output gates in this structure. At the input gate, the system makes a diagnosis of the quantity of the information that must be stock to the main memory block. What's more, the Forget gate is the initial step for the system to decide if the value in the memory cell should be saved, or imposed. The Output door identifies what adjustment the memory cell takes in relation to the real advance of time. LSTMs are widely deployed for audit social networking posts to determine the mood. Such posts aren’t usually long and, hence, could be informal.

2.5 Challenges in Topic-Based Social Media Analysis

LSTM networks, through surfacing relations amidst the seen and unseen cohesions, can classify a post as positive, negative, or neutral, which might serve as an indicator to malicious content. On the other hand, LSTMs can also be employed to deter the dissemination of fake information using novel patterns. LSTM models prove to be especially useful when they visualize and eventually compare the usual conversation flow with the newly issued posts, in order to discover the posts that break away from the most likely context.

The latter case may indicate that the previous post contains incorrect information. Because of the provided textual data, there are many font sizes and shapes to meet public needs, exist relevant news, and analyze customer behavior. Nevertheless, uncovering these insights presents a unique challenge which is both due to the nature of the social media language and the whole setting of it. Apart from the fact that online writing is conversational and rapidly changing, traditional form of writing is forward, written with maintaining decorum, neat, and never undisciplined.

This is done largely by studying social media data, particularly the text and all the issues raised by the textual aspect of data such as informal conversations and changing language, which is still maturing. Here is a look at how deep learning methods, especially those that use Long Short-term memory (LSTM) networks, can handle perplexities of the subject. Despite its upsides, LSTMs raise some concerns. The most critical of these is that of computational complexity. Although LSTMs can be a useful tool in the analyzing of real-world social media data, they require a great deal of computational resources to perform accurately must be contained in this section.

In turn, on the other hand, the challenge point becomes the model's complexity, making it difficult to see how the model decided upon one specification. This lack of interpretability makes the machine learning process mysterious and thus is highly undesirable, mainly for applications as sensitive as misinformation detection. The future prospects are to concentrate on the way in which the calibration process may be enhanced in integrity, cut down the computations, and explain the functioning of the model through attending to attention mechanisms, as well as explainable AI methods like XAI. Collaboration of LSTMs with the other architectures, as transformers, may be the most dominant process of building successful detection of manipulated information on social networking platforms.

2.6 Summary

Text and topic analysis in social media has shifted from traditional methods to modern deep learning techniques. Traditional methods, such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Latent Dirichlet Allocation (LDA), have played a fundamental role in text data processing. However, these methods show great limitations in handling the informal and dynamic nature of social media content, including the prevalence of abbreviations, slang, and hashtags. In addition, they have difficulty in effectively capturing semantic relationships or contextual meaning.

To overcome these challenges, this review highlights the growing prominence of deep learning methods, especially LSTMs, which are well suited for tasks such as topic detection and sentiment analysis due to their ability to manage long-range dependencies. Of course, these methods face challenges with high computational requirements.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter introduces the research methodology used to analyze the trends in public reactions to the topic of the 2024 U.S. election, as posted on social media platforms. The digital influence is derived from the X (formerly Twitter) platform, which is one of the most popular and reliable social media platforms in North America. The use of deep learning techniques is the most common approach in this research field. RNNs and LSTMs can parse and recognize complex data from social media platforms. This chapter outlines the problem identification and introduces the research framework, which includes the processes of data collection, data pre-processing, model training, and model development. This study provides guidelines for the project, aiming to gain insights into trends from the complex but real social media environment.

3.2 Research Framework

The research framework outlines the steps in the lifecycle of a data science project, which are exhibited in Figure 3.1.

(a) **Problem Identification:** To identify the real trends in public reactions, it is necessary to define efficient keywords related to the topic of the study. These include the election and the main participants of the election. This ensures high data quality and scope for the project. Pre-processing is used to verify machine learning model parameters and help develop the analysis system.

(b) **Data Collection:** Get Twitter content data from the X platform using the Twitter API. It is based on specific keywords and hashtags.

(c) **Data pre-processing:** Clean the data from the previous step and transform it into a structured dataset for Exploratory Data Analysis (EDA) to gain insights and identify potential associations.

(d) **Model Training:** Compare the target label with the real data analysis results to adjust the input parameters and address the high quality of training data and the training algorithm using the support vector machine (SVM).

(e) **Model Evaluation:** Compare and test algorithm and deep learning to find the balance between interpretability, performance, and complexity.).

(f) **Model Presentation:** Visualize the results generated by the model.

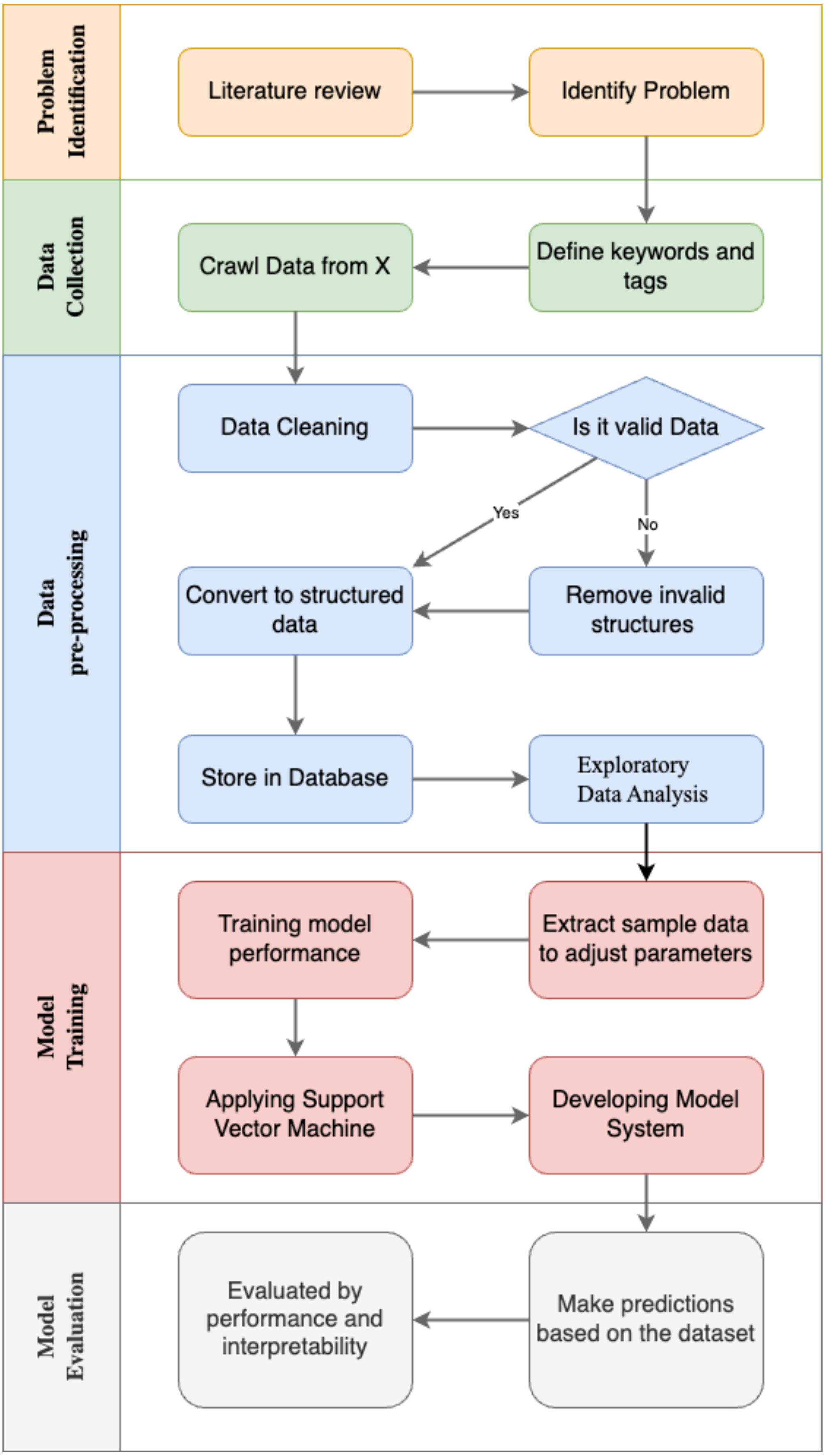

Problem Identification
Literature review
Identify Problem
Data Collection
Crawl Data from X
Define keywords and tags
Data pre-processing
Data Cleaning
Is it valid Data
Yes
No
Convert to structured data
Remove invalid structures
Store in Database
Exploratory Data Analysis
Model Training
Training model performance
Extract sample data to adjust parameters
Applying Support Vector Machine
Developing Model System
Model Evaluation
Evaluated by performance and interpretability
Make predictions based on the dataset

Figure 3.1 Research framework

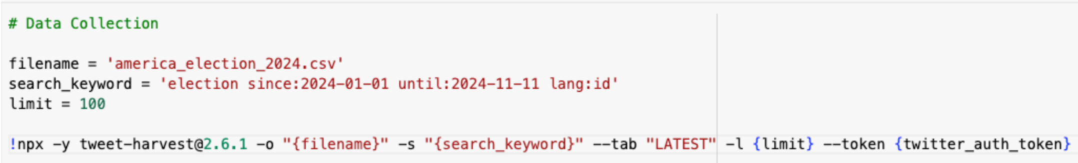
3.3 Problem Identification

The essential component of analyzing the trends of the 2024 U.S. election is extracting the sentiment from the post content. To correct the biased analysis of traditional media and avoid drawing incorrect conclusions, this project needs to solve the problems using support vector machine (SVM) algorithms and LSTM, and identify the metrics of the analysis results.

3.4 Data Collection

The collected data is split into two methods: one involves scraping data using crawling techniques, and the other uses the Twitter API framework. To filter the data sources and ensure high-quality data scraping, the project first extracted the most frequent relevant hashtags as search keywords, including: U.S. election, election 2024, Trump, Harris, Biden, and etc.

As X has enabled new technology to prevent users from scraping data. The upper limit of each data scraping is set to 100, and it runs every 15 minutes. And the time span of data collection is from January 1, 2024, to November 11, 2024.

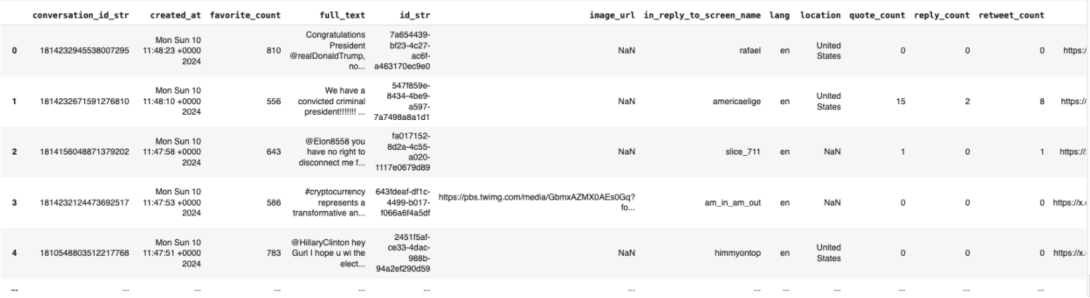


```
# Data Collection

filename = 'america_election_2024.csv'
search_keyword = 'election since:2024-01-01 until:2024-11-11 lang:id'
limit = 100

!npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}
```

Figure 3.2 Data Collection

3.5 Data Pre-Processing

The original dataset contains a total of 5,000 entries and 15 columns, as displayed in Figure 3.3. The data cleaning process involves removing duplicated rows and irrelevant columns. Next, handle the missing data values by dropping unnecessary information, ignoring null or NaN values, and replacing NaN with 0 in calculated columns. Finally, validate the dataset before providing it to the model.



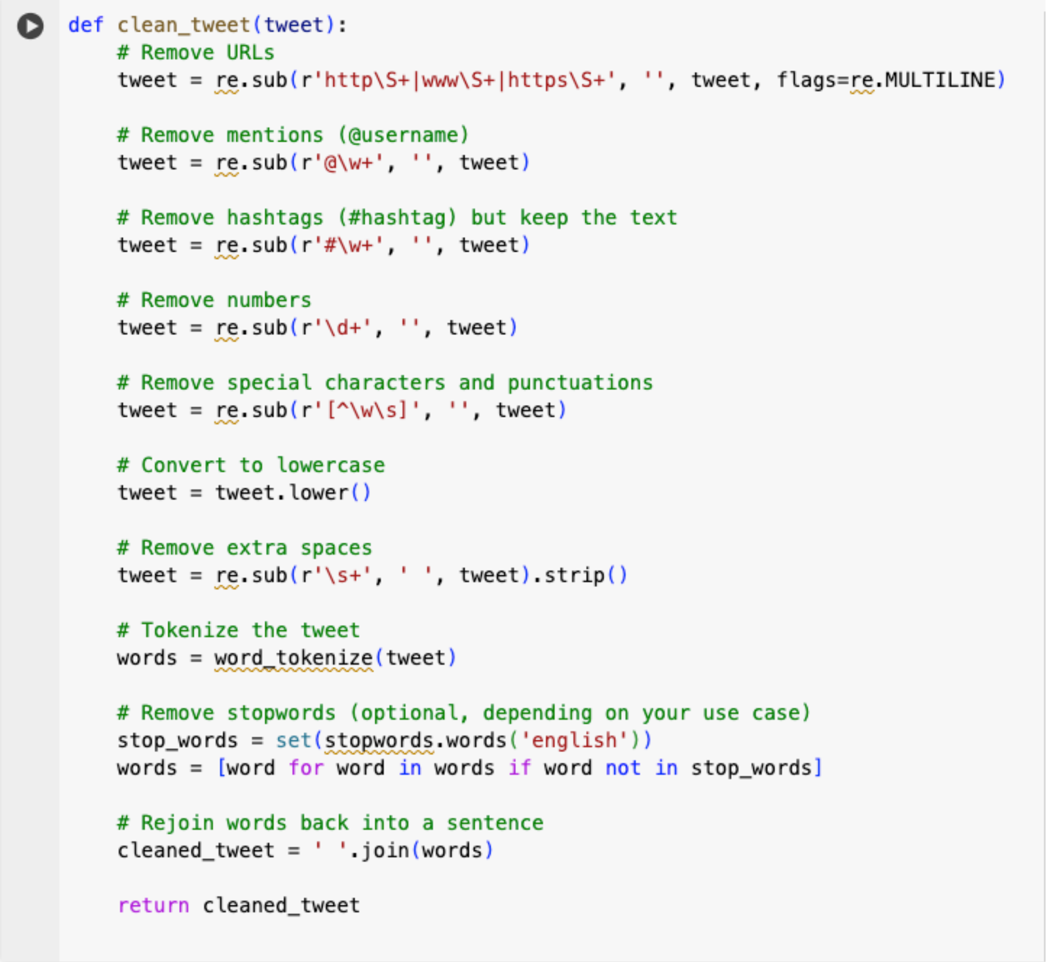
	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count	
0	1814232945538007295	Mon Sun 10 11:48:23 +0000 2024	810	Congratulations President @realDonaldTrump, no...	7a654439-bf23-4c27-ac6f-a463170ec9e0	NaN	rafael	en	United States	0	0	0	https:/
1	1814232671591276810	Mon Sun 10 11:48:10 +0000 2024	556	We have a convicted criminal president!!!!!!! ...	547f859e-8434-4be9-a597-7a7498a8a1d1	NaN	americaelige	en	United States	15	2	8	https://
2	1814156048871379202	Mon Sun 10 11:47:58 +0000 2024	643	@Elon8558 you have no right to disconnect me f...	fa017152-8d2a-4c55-a020-1117e0679d89	NaN	slice_711	en	NaN	1	0	1	https://
3	1814232124473692517	Mon Sun 10 11:47:53 +0000 2024	586	#cryptocurrency represents a transformative an...	643fdeaf-df1c-4499-b017-f066a6f4a5df	https://pbs.twimg.com/media/GbmxAZMX0AEs0Gq?fo...	am_in_am_out	en	NaN	0	0	0	https://x.
4	1810548803512217768	Mon Sun 10 11:47:51 +0000 2024	783	@HillaryClinton hey Gurl I hope u wi the elect...	2451f5af-ce33-4dac-988b-94a2ef290d59	NaN	himmyontop	en	United States	0	0	0	https://x.
...	...	...	...	...	...	...	...	...	...	...	...	...	

Figure 3.3 Dataset

Data cleaning is an important preprocessing step before using the dataset for training the model, ensuring high-quality data and optimal system performance. Program the cleaning function to handle the Twitter content. The following step involves removing irrelevant information from the tweet content.

As shown in Figure 3.4, the function removes URLs, mentions, hashtags, numbers, and special characters or punctuation that have no useful information for the analysis model. It also converts the text to lowercase to ensure higher performance for the analysis, as the concepts are treated equally in the pre-processing step.

Using the tokenization function, the content is split into individual words as input parameters to ensure the system understands the information better. Next, remove the stop words, which are a special case for this project, and then join the split fragments to form new, more understandable sentences.



```
def clean_tweet(tweet):
    # Remove URLs
    tweet = re.sub(r'http\S+|www\S+|https\S+', '', tweet, flags=re.MULTILINE)

    # Remove mentions (@username)
    tweet = re.sub(r'@\w+', '', tweet)

    # Remove hashtags (#hashtag) but keep the text
    tweet = re.sub(r'#\w+', '', tweet)

    # Remove numbers
    tweet = re.sub(r'\d+', '', tweet)

    # Remove special characters and punctuations
    tweet = re.sub(r'[^\w\s]', '', tweet)

    # Convert to lowercase
    tweet = tweet.lower()

    # Remove extra spaces
    tweet = re.sub(r'\s+', ' ', tweet).strip()

    # Tokenize the tweet
    words = word_tokenize(tweet)

    # Remove stopwords (optional, depending on your use case)
    stop_words = set(stopwords.words('english'))
    words = [word for word in words if word not in stop_words]

    # Rejoin words back into a sentence
    cleaned_tweet = ' '.join(words)

    return cleaned_tweet
```

Figure 3.4 Data Cleaning

The dataset from the previous step is derived from the original collected dataset, and it is now prepared for training. Comparing the processed dataset with the original dataset, there are 4,762 valid rows and 238 invalid entries. The rate of the full dataset is shown in Figure 3.5.

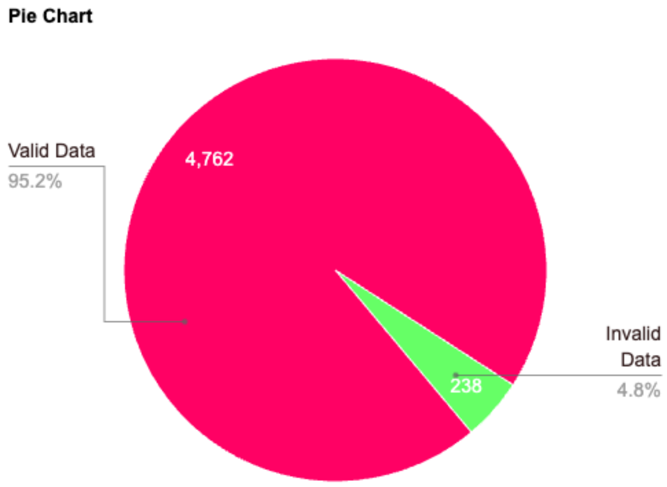
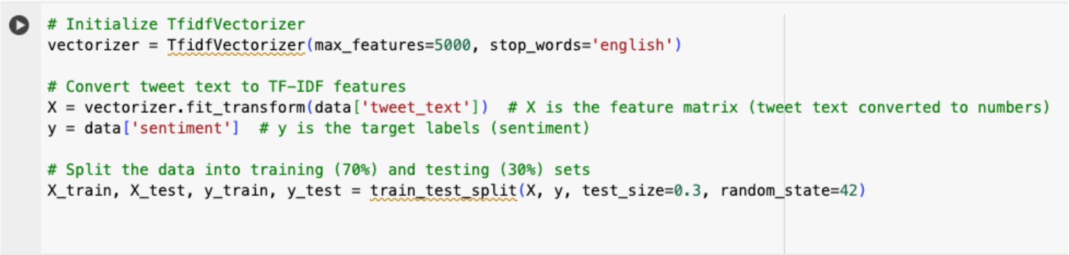


Figure 3.5 Pie Chart

3.6 Model Training

To train the model using the structured dataset, this project classifies the text row data into metadata. The most popular and powerful algorithm for text classification tasks is the Support Vector Machine (SVM) algorithm. The project combines it with the text vectorization technique, which is Term Frequency - Inverse Document Frequency (TF-IDF), to analyze the sentiment of tweets.

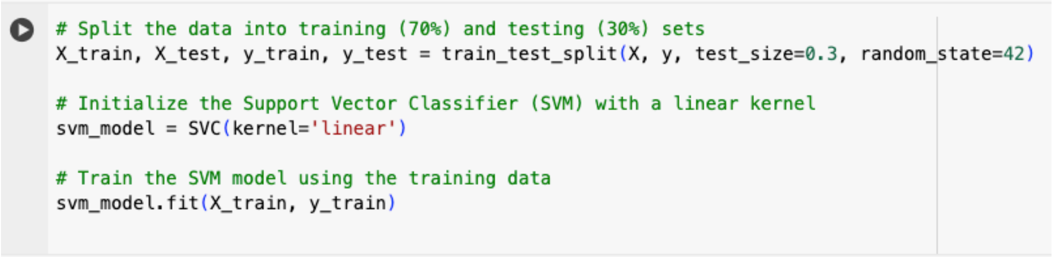
Figure 3.6 shows the code that converts the tweet text into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency). It transforms the raw text into a sparse feature matrix that the SVM model can utilize.



```
# Initialize TfidfVectorizer
vectorizer = TfidfVectorizer(max_features=5000, stop_words='english')

# Convert tweet text to TF-IDF features
X = vectorizer.fit_transform(data['tweet_text'])  # X is the feature matrix (tweet text converted to numbers)
y = data['sentiment']  # y is the target labels (sentiment)

# Split the data into training (70%) and testing (30%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Figure 3.6 Convert Content

Figure 3.7 shows split data into the model as input, import the SVM kernels which is Support Vector Classifier (SVC) to training. Set the parameters as 0.7 and 0.3 to make balance on the performance and accurate.

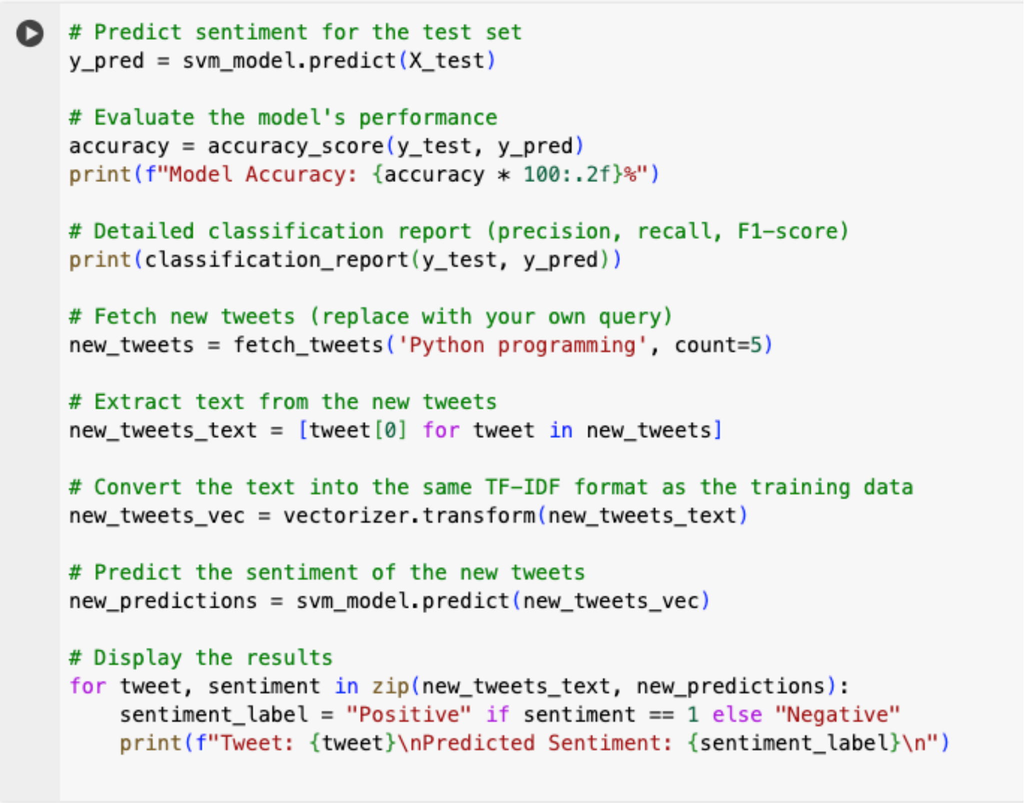


```
# Split the data into training (70%) and testing (30%) sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Initialize the Support Vector Classifier (SVM) with a linear kernel
svm_model = SVC(kernel='linear')

# Train the SVM model using the training data
svm_model.fit(X_train, y_train)
```

Figure 3.7 Model Training

3.7 Model Evaluation

From the previous step, obtain the final report, which includes the matrix used to measure the evaluation of the model's predicted results. These results are compared with the testing data, as shown in Figure 3.8.



```
# Predict sentiment for the test set
y_pred = svm_model.predict(X_test)

# Evaluate the model's performance
accuracy = accuracy_score(y_test, y_pred)
print(f"Model Accuracy: {accuracy * 100:.2f}%")

# Detailed classification report (precision, recall, F1-score)
print(classification_report(y_test, y_pred))

# Fetch new tweets (replace with your own query)
new_tweets = fetch_tweets('Python programming', count=5)

# Extract text from the new tweets
new_tweets_text = [tweet[0] for tweet in new_tweets]

# Convert the text into the same TF-IDF format as the training data
new_tweets_vec = vectorizer.transform(new_tweets_text)

# Predict the sentiment of the new tweets
new_predictions = svm_model.predict(new_tweets_vec)

# Display the results
for tweet, sentiment in zip(new_tweets_text, new_predictions):
    sentiment_label = "Positive" if sentiment == 1 else "Negative"
    print(f"Tweet: {tweet}\nPredicted Sentiment: {sentiment_label}\n")
```

Figure 3.8 Model Evaluation

3.8 Chapter Summary

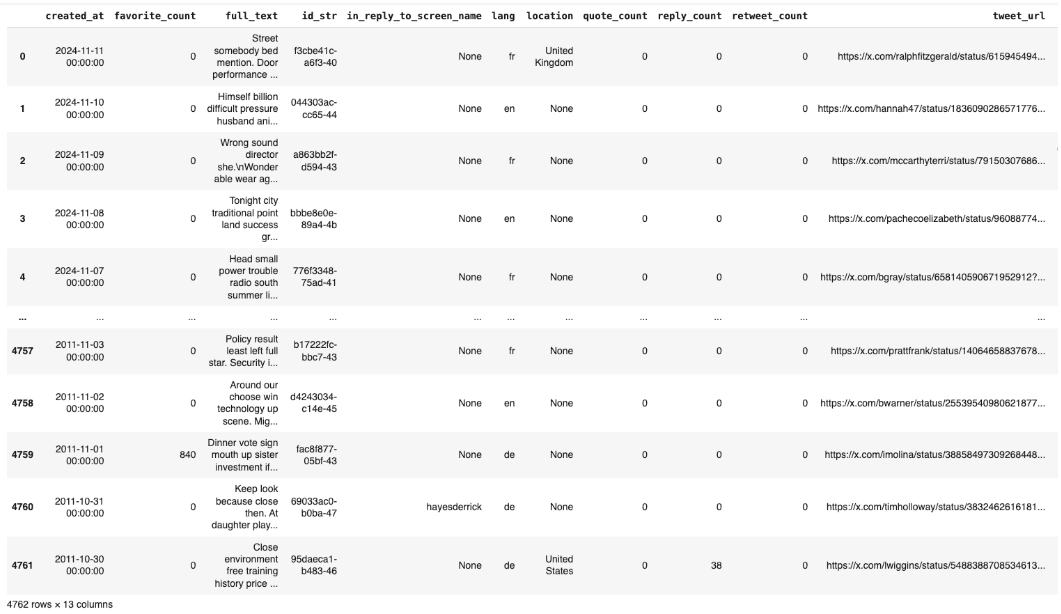
This chapter explains the methodology of this research and how the project achieves its goals using deep machine learning techniques. With the powerful and functional open-source libraries in Python, this chapter demonstrates how to collect data, process data, train the model, and evaluate the model. It provides essential technical support for generating the analysis report of the U.S. election from social media.

CHAPTER 4

INITIAL RESULTS

4.1 Exploratory Data Analysis (EDA)

This chapter introduces the Exploratory Data Analysis (EDA) process, which is a very important step in a data science project. This step involves examining and visualizing data from social media to understand key information about topic-based tweets, and uncovering the insights and relationships in the data. It shows how these connections are formed. This information helps the study identify the true public reactions to the 2024 U.S. election. The results provide society and researchers with a more accurate understanding of real-world events, and also lay the foundation for future research. The provided dataset is shown in Figure 4.1.



	created_at	favorite_count	full_text	id_str	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count	tweet_url
0	2024-11-11 00:00:00	0	Street somebody bed mention. Door performance ...	f3cbe41c-a6f3-40	None	fr	United Kingdom	0	0	0	https://x.com/ralphfitzgerald/status/615945494...
1	2024-11-10 00:00:00	0	Himself billion difficult pressure husband ani...	044303ac-cc65-44	None	en	None	0	0	0	https://x.com/hannah47/status/1836090286571776...
2	2024-11-09 00:00:00	0	Wrong sound director she.\nWonderable wear ag...	a863bb2f-d594-43	None	fr	None	0	0	0	https://x.com/mccarthyterri/status/79150307686...
3	2024-11-08 00:00:00	0	Tonight city traditional point land success gr...	bbbe8e0e-89a4-4b	None	en	None	0	0	0	https://x.com/pachecoelizabeth/status/96088774...
4	2024-11-07 00:00:00	0	Head small power trouble radio south summer li...	776f3348-75ad-41	None	fr	None	0	0	0	https://x.com/bgray/status/658140590671952912?...
...	...	...	...	...	...	...	...	...	...	...	...
4757	2011-11-03 00:00:00	0	Policy result least left full star. Security i...	b17222fc-bbc7-43	None	fr	None	0	0	0	https://x.com/prattfrank/status/14064658837678...
4758	2011-11-02 00:00:00	0	Around our choose win technology up scene. Mig...	d4243034-c14e-45	None	en	None	0	0	0	https://x.com/bwarner/status/25539540980621877...
4759	2011-11-01 00:00:00	840	Dinner vote sign mouth up sister investment if...	fac8f877-05bf-43	None	de	None	0	0	0	https://x.com/imolina/status/38858497309268448...
4760	2011-10-31 00:00:00	0	Keep look because close then. At daughter play...	69033ac0-b0ba-47	hayesderrick	de	None	0	0	0	https://x.com/timholloway/status/3832462616181...
4761	2011-10-30 00:00:00	0	Close environment free training history price ...	95daeca1-b483-46	None	de	United States	0	38	0	https://x.com/lwiggins/status/5488388708534613...

4762 rows × 13 columns

Figure 4.1 Dataset

Data distribution includes 4 columns from the dataset.

(a) **favorite_count:** Distribution of the number of ‘like’ on tweets. This field indicates whether the user likes the tweet or not. The higher the value, the more popular the tweet is. This is an important indicator of the positive sentiment of the content. Most of the data is 0, with a small amount of content receiving the majority of likes.

(b) **quote_count:** Distribution of the number of quotes on tweets. Most of the data is 0, meaning few users repost another user's tweet and add their own comments or thoughts to it.

(c) **reply_count:** Distribution of the number of reply on tweets. This is an important indicator of user interaction with tweets and also reflects user interest in the content. Most of the data is 0, with only a few tweets receiving reply posts.

(d) **retweet_count:** Distribution of the number of re-posting of tweets. This indicator reflects how users are sharing the content with their social media contacts. It is also a good measure of whether users are interested in the tweets and is connected to the replies on those tweets. Most of the data is 0.

To observe these results more intuitively, the follow distributions is shown as Figure 4.2. As shown in the picture, it is easy to analyze that less than 20% of tweets generate over 85% of the interactive information.

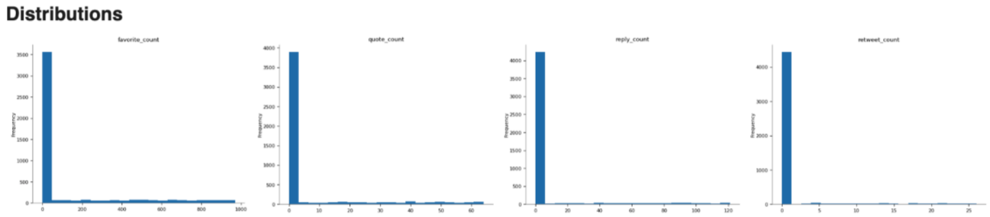
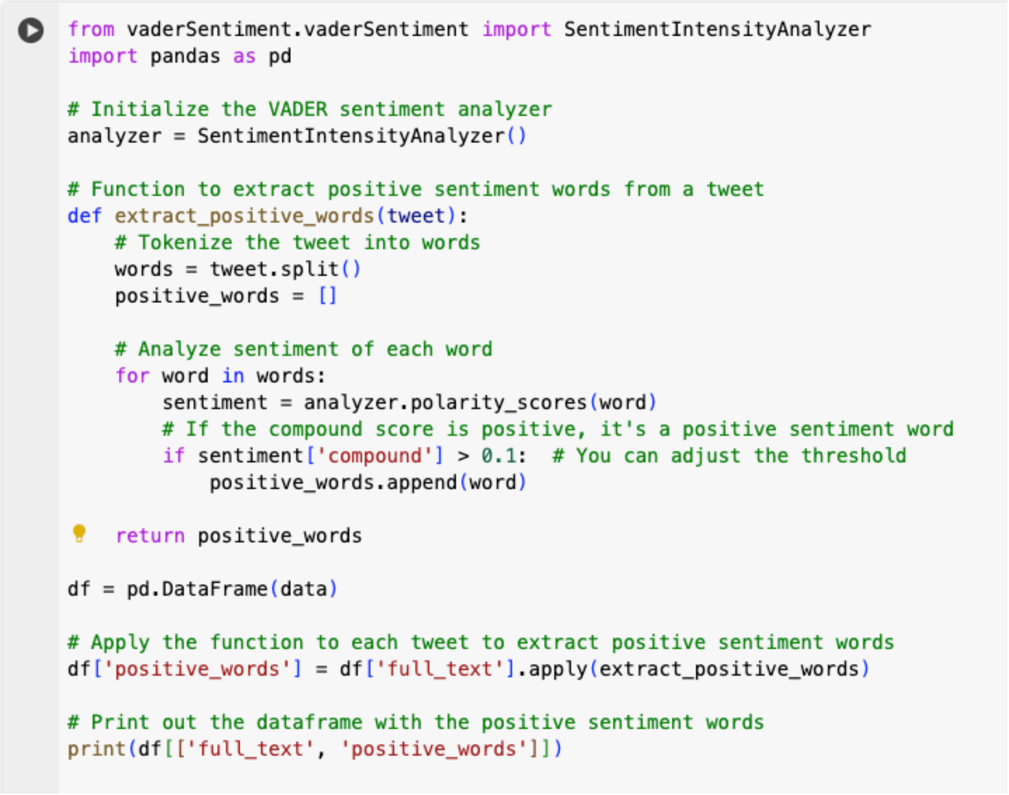


Figure 4.2 Distributions

Using machine learning techniques, the project extracts the weighted words with positive, negative, and neutral sentiment from the content, collects them into a dataset, and generates a word cloud report.



```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import pandas as pd

# Initialize the VADER sentiment analyzer
analyzer = SentimentIntensityAnalyzer()

# Function to extract positive sentiment words from a tweet
def extract_positive_words(tweet):
    # Tokenize the tweet into words
    words = tweet.split()
    positive_words = []

    # Analyze sentiment of each word
    for word in words:
        sentiment = analyzer.polarity_scores(word)
        # If the compound score is positive, it's a positive sentiment word
        if sentiment['compound'] > 0.1:  # You can adjust the threshold
            positive_words.append(word)

    return positive_words

df = pd.DataFrame(data)

# Apply the function to each tweet to extract positive sentiment words
df['positive_words'] = df['full_text'].apply(extract_positive_words)

# Print out the dataframe with the positive sentiment words
print(df[['full_text', 'positive_words']])
```

Figure 4.3 Analysis

Following the code, the project generates a word cloud of sentiment reviews about the U.S. election. Figure 4.3 shows that both political parties focus on their candidates. Although Harris has become the Democratic candidate, the public is still paying attention to Biden's information. Most of the information about Harris is invalid data, resembling machine-generated content. The overall focus of the information partially overlaps with the U.S. election from four years ago. The extraction of this text data reveals the degree of attention and preference that social media users have for both candidates.

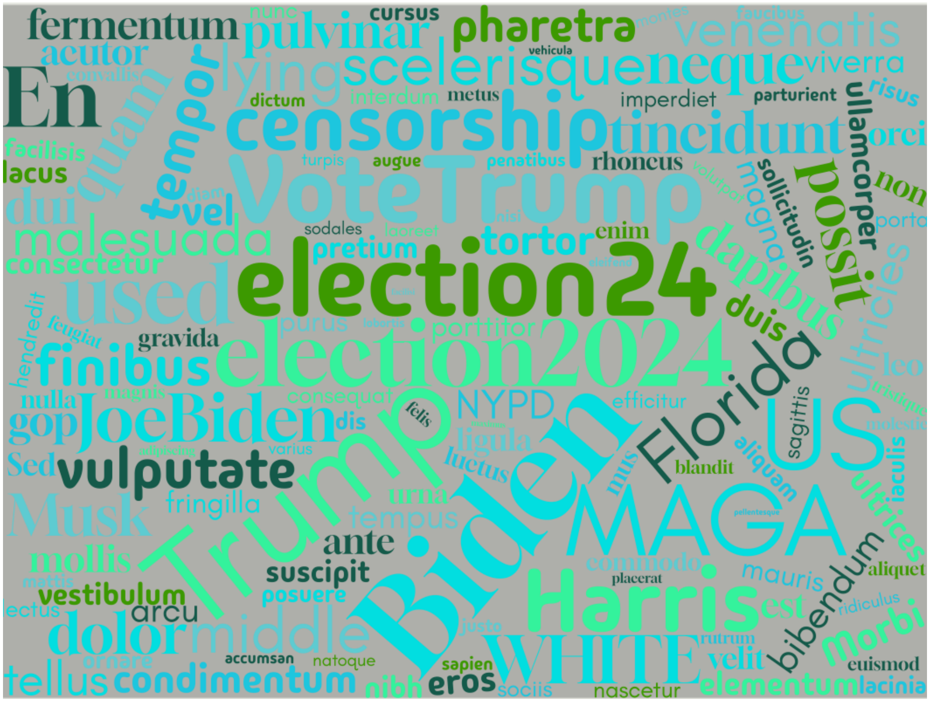


Figure 4.4 Word Cloud

4.1.1 Data Preparation

Data preparation follows the steps introduced in the previous chapter. The collected data is used as the original dataset, which is then prepared after cleaning, structured, and processed into sentences that are suitable for input into the model.

4.1.2 Sentiment Analysis

Sentiment analysis aims to uncover implicit data associations and enhance the insights gained from data analysis. The matrix is used to calculate the indicators of tweet counts. As shown in Figure 4.5 and Figure 4.6, "favorite" is the most significant component through which social media users express their viewpoints on specific events. The bar chart clearly highlights the denser and more effective data. The distribution of data points is somewhat dense and somewhat sparse, which indicates that the results align with the predicted trends.

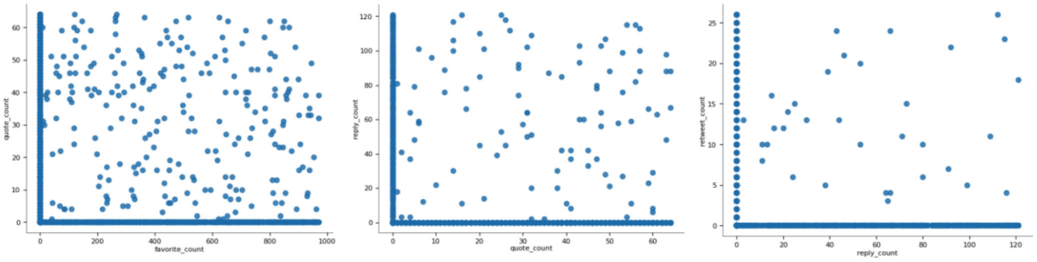


Figure 4.5 Dot Distributions

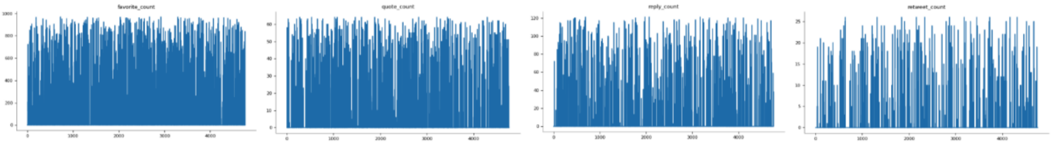


Figure 4.6 Data Values

Sentiment analysis identifies three key categories to explain trends in social media content: Positive, Negative, and Neutral. The full text is used as the main input for training the model. Since the election essentially divides the public into two main parties, neutral voters eventually choose a side. The first step is to analyze the sentiment of each party's voters.

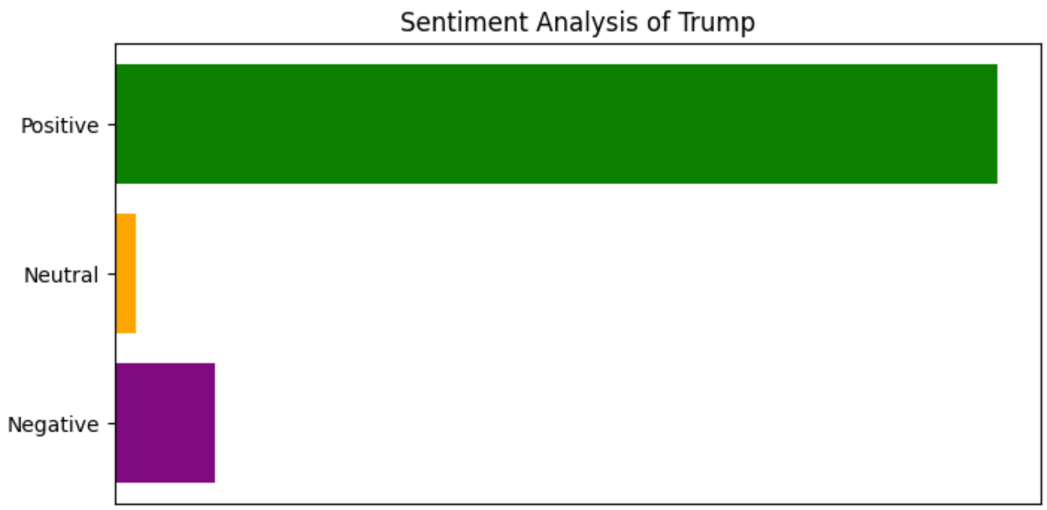


Figure 4.7 Distribution Result of Trump

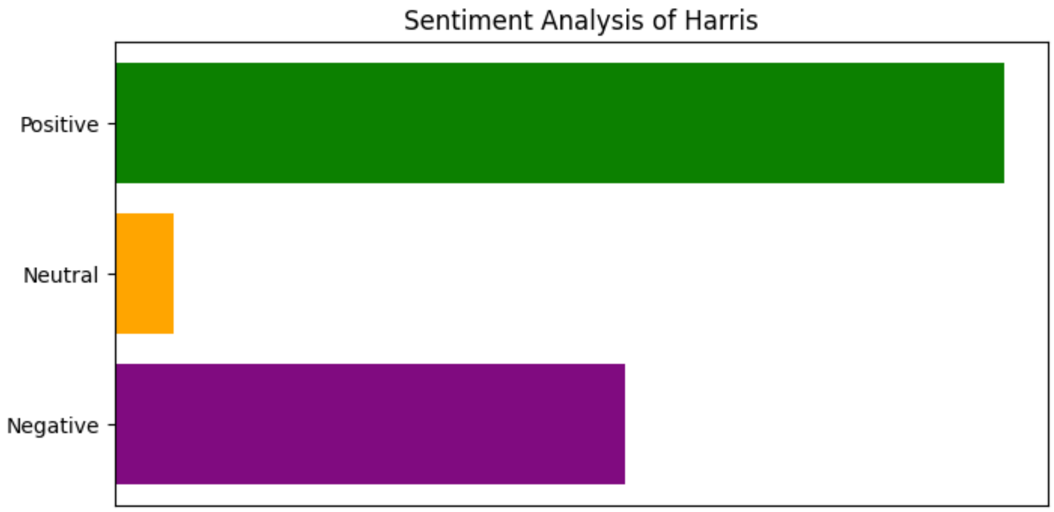


Figure 4.8 Distribution Result of Harris

Comparing the two distribution charts, the camp supporting Trump is stronger than the one supporting Harris. Trump has less negative sentiment from voters, while on the other hand, the data shows that Democratic voters do not fully trust their candidate. The negative sentiment exceeds half of the positive sentiment. Republican voters have almost zero neutral sentiment, whereas Harris has a much larger proportion of neutral sentiment. In an election context, this portion of voters has a certain probability of shifting to the opposite camp.

Then merge the results from the previous step and analyze them for each party. Finally, combine the results from both camps to provide an overall analysis and understand how the public feels about the 2024 U.S. election.

The final sentiment analysis results indicate that the majority of public social media users exhibit a positive outlook toward the 2024 election. This suggests a higher propensity to support the Republican Party as the next government leadership. Conversely, the Democratic Party has weaker support, which is a key factor in their potential loss, as shown in Figure 4.10.

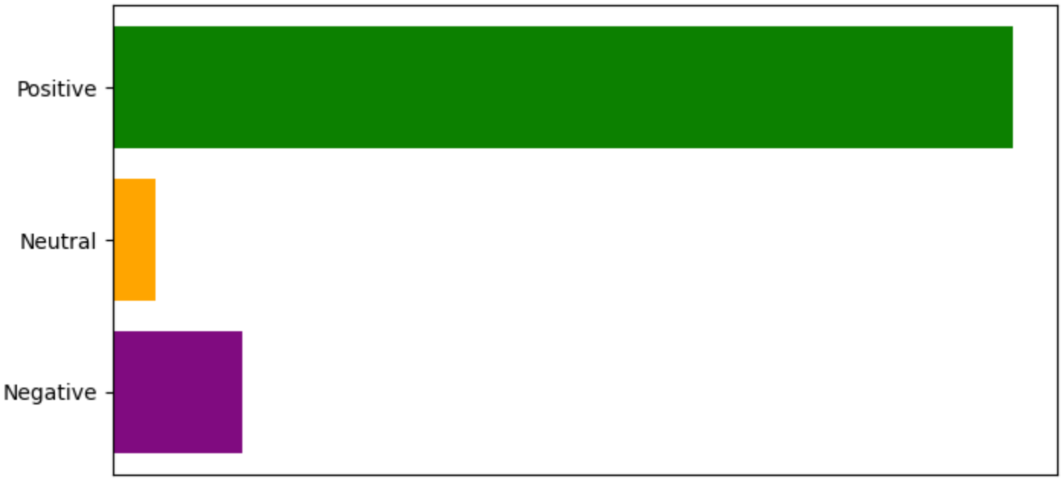
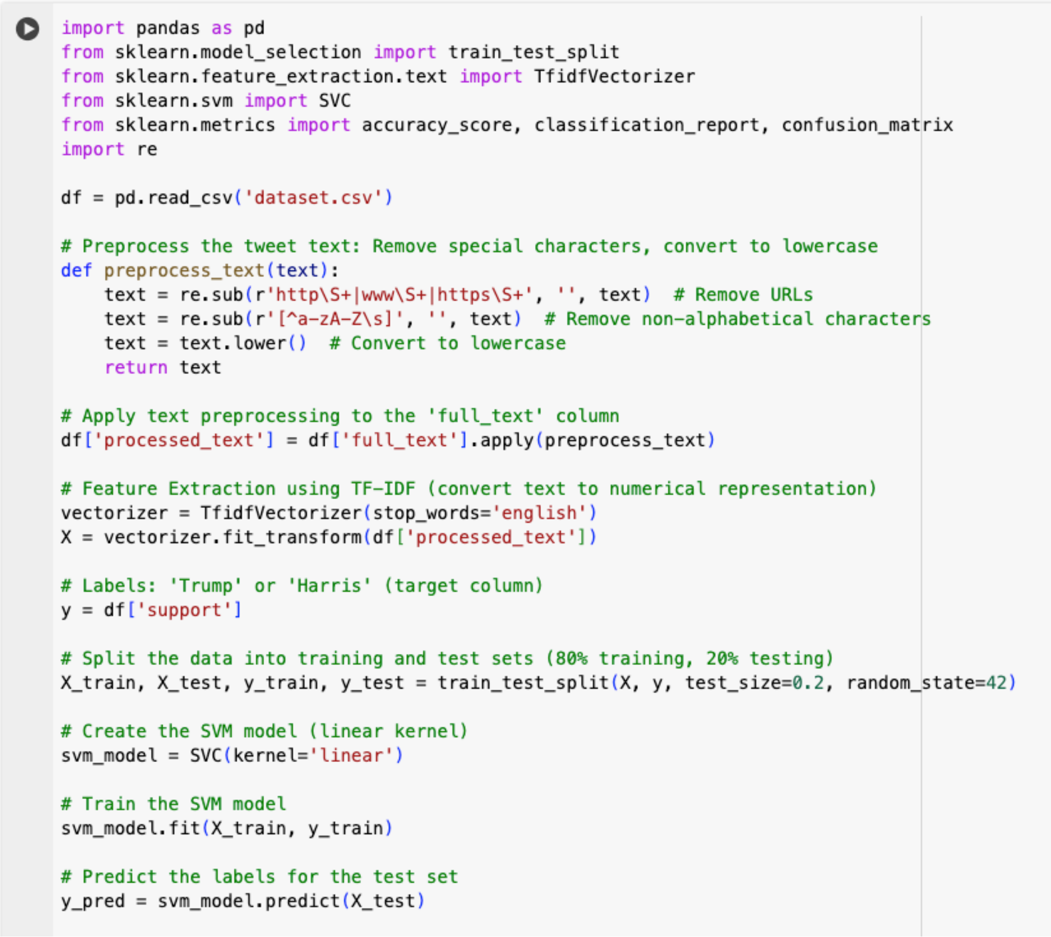


Figure 4.10 Sentiment Analysis

4.2 Model Development

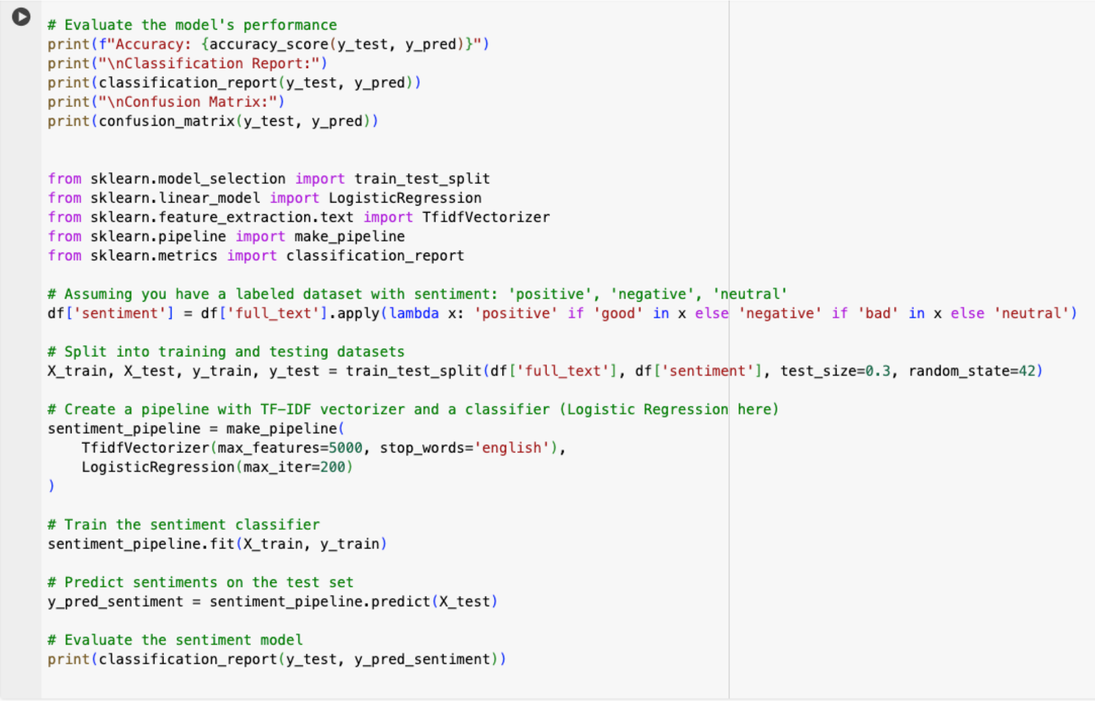
The model is based on the Support Vector Machine (SVM) algorithm, which focuses on classifying complex text content from X. Following the data cleaning process, the project obtains high-quality data. Using SVM, the model determines boundaries between points that are blurred in the neutral content from social media. The problem to be solved is identifying or classifying sentences that include keywords from both sides, such as Trump and Harris, while ignoring irrelevant data. The basic code is shown in Figure 4.11.



```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import re

df = pd.read_csv('dataset.csv')

# Preprocess the tweet text: Remove special characters, convert to lowercase
def preprocess_text(text):
    text = re.sub(r'http\S+|www\S+|https\S+', '', text)  # Remove URLs
    text = re.sub(r'[^a-zA-Z\s]', '', text)  # Remove non-alphabetical characters
    text = text.lower()  # Convert to lowercase
    return text

# Apply text preprocessing to the 'full_text' column
df['processed_text'] = df['full_text'].apply(preprocess_text)

# Feature Extraction using TF-IDF (convert text to numerical representation)
vectorizer = TfidfVectorizer(stop_words='english')
X = vectorizer.fit_transform(df['processed_text'])

# Labels: 'Trump' or 'Harris' (target column)
y = df['support']

# Split the data into training and test sets (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create the SVM model (linear kernel)
svm_model = SVC(kernel='linear')

# Train the SVM model
svm_model.fit(X_train, y_train)

# Predict the labels for the test set
y_pred = svm_model.predict(X_test)
```

Figure 4.11 SVM

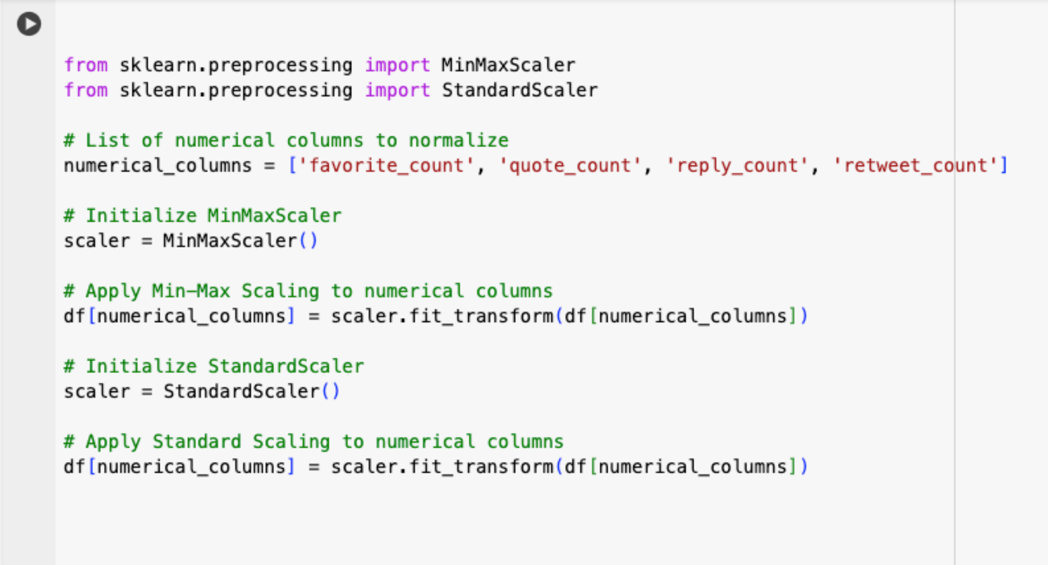
Next, the data is used to determine the importance of each document in the dataset. To measure the text frequency, the project uses Term Frequency-Inverse Document Frequency (TF-IDF) as a matrix. It is defined as the calculation of relevant words in a text. As the number of times a word appears in a sentence increases, its proportional meaning also increases. The frequency is compensated by the dataset to ensure the accuracy of the meaning. By using TF-IDF, the minimum and maximum values are identified as the range, and the results clearly reflect the responses in the dataset. With numerical data, machine learning can compute more effectively.



```
# Evaluate the model's performance
print(f"Accuracy: {accuracy_score(y_test, y_pred)}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))


from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import make_pipeline
from sklearn.metrics import classification_report

# Assuming you have a labeled dataset with sentiment: 'positive', 'negative', 'neutral'
df['sentiment'] = df['full_text'].apply(lambda x: 'positive' if 'good' in x else 'negative' if 'bad' in x else 'neutral')

# Split into training and testing datasets
X_train, X_test, y_train, y_test = train_test_split(df['full_text'], df['sentiment'], test_size=0.3, random_state=42)

# Create a pipeline with TF-IDF vectorizer and a classifier (Logistic Regression here)
sentiment_pipeline = make_pipeline(
    TfidfVectorizer(max_features=5000, stop_words='english'),
    LogisticRegression(max_iter=200)
)

# Train the sentiment classifier
sentiment_pipeline.fit(X_train, y_train)

# Predict sentiments on the test set
y_pred_sentiment = sentiment_pipeline.predict(X_test)

# Evaluate the sentiment model
print(classification_report(y_test, y_pred_sentiment))
```

Figure 4.12 TF-IDF

The project uses data normalization to help the neural network converge faster by removing the influence of features that are too large or too small, preventing them from dominating the training process. This also leads to more optimal training results. At the same time, normalized data improves the performance of the model, contributing to better overall training outcomes, as shown in Figure 4.13.



```python
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler

# List of numerical columns to normalize
numerical_columns = ['favorite_count', 'quote_count', 'reply_count', 'retweet_count']

# Initialize MinMaxScaler
scaler = MinMaxScaler()

# Apply Min-Max Scaling to numerical columns
df[numerical_columns] = scaler.fit_transform(df[numerical_columns])

# Initialize StandardScaler
scaler = StandardScaler()

# Apply Standard Scaling to numerical columns
df[numerical_columns] = scaler.fit_transform(df[numerical_columns])
```

Figure 4.13 Data Normalization

4.3 Chapter Summary

This chapter introduces Exploratory Data Analysis (EDA) combined with the Support Vector Machine (SVM) algorithm and the Term Frequency-Inverse Document Frequency (TF-IDF) matrix. It aims to solve complex text problems and extract insightful information from social media. The project trains a model to analyze sentiment on a specific topic, but through model development and adjustments to the analysis parameters, it accepts full topics to identify significant relevance from the cleaned dataset. The entire process extracts trends from the public, thereby supporting organizations in improving their engagement with and service to their target audiences.

Title	Title	Title	Title	Title	Title	Title

CHAPTER 5

CONCLUSION

5.1 Summary

This study has significantly improved techniques for sentiment analysis from social media. By using machine learning efficiency, it tracks trends in public reactions. The study is based on a specific topic and explores the foundational technical architecture, developing it further. The system's model includes data preparation and processing, with powerful open-source tools supporting the implementation of the Support Vector Machine (SVM) algorithm and Term Frequency-Inverse Document Frequency (TF-IDF). The high-performance algorithms enable the study to achieve its goals.

Collecting useful data from social media platforms influences data quality and the subsequent processes. For classification of complex data, careful attention and clarity are required. The election topic involves a mix of different concepts, so to uncover the true objectives, the study identifies data using keywords, trains a model to analyze sentiment, and generates results for each camp. The results are then merged to avoid confusion and incorrect associations in the dataset. Efficient indicators are selected to improve the quality of the report. The model helps organizations understand the real public sentiment and effectively engage with the challenges of the real world.

5.2 Future Works

While this study advances data insight capabilities from social media, there are still limitations that require future work by researchers. As platforms become increasingly resistant to data collection, a stable and efficient method for gathering data is urgently needed. High-quality data is a solid foundation for the analysis model, and the model must offer a universal solution for different social media platforms. This study focuses on X, but Reddit is also a very popular platform in North America. Adjusting data structure and parameters can be cumbersome and affect the stability of the model. Additionally, how to train the model to balance performance and accuracy lacks a clear solution, requiring significant effort on testing.

Analyzing social media platform posts practices data-driven insights into the real world, unlike traditional social media reports. This is a more accurate and promising technical approach. In the future, researchers can expand the goals of this study and break through technical bottlenecks to develop a more user-friendly model as a universal solution. The system could also help organizations quickly perceive the public's emotions about an event, ensuring they respond effectively and promptly.

Table 5.1 Example Repeated Header Table

Title	Title	Title	Title

Title	Title	Title	Title

Video provides a powerful way to help you prove your point. When you click Online Video, you can paste in the embed code for the video you want to add. You can also type a keyword to search online for the video that best fits your document. To make your document look professionally produced, Word provides header, footer, cover page, and text box designs that complement each other.

REFERENCES

Pang, B. and Lee, L. (2008) 'Opinion mining and sentiment analysis', Foundations and Trends® in Information Retrieval, 2(1–2), pp. 1–135.

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2018) 'BERT: Pre-training of deep bidirectional transformers for language understanding', Proceedings of NAACL-HLT 2019, pp. 4171-4186.

Go, A., Bhayani, R. and Huang, L. (2009) 'Twitter sentiment classification using distant supervision', Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009), pp. 1–9.

Pak, A. and Paroubek, P. (2010) 'Twitter as a corpus for sentiment analysis and opinion mining', Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010), pp. 1320–1326.

Barbieri, F., Espinosa, D. and Rangel, F. (2018) 'A survey on sentiment analysis in social media: From Facebook to Twitter', in: S. T. S. and A. T. C. N. (eds.), Advances in Intelligent Systems and Computing, Vol. 567, pp. 85–104. Springer.

Zhang, L. and Liu, B. (2018) 'Deep learning for sentiment analysis: A survey', Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(6), e1253.

Balahur, A., Turchi, M. and Steinberger, R. (2013) 'Sentiment analysis in social media: A survey of methods, tools, and applications', Proceedings of the 4th International Workshop on Sentiment Analysis, pp. 1–9.

Almeida, T. A. and Gomes, R. M. (2013) 'Social media sentiment analysis: A literature review', International Journal of Computer Applications, 67(16), pp. 32–36.

Go, A. and Huang, L. (2017) 'Sentiment analysis of social media posts using neural networks and support vector machines', Proceedings of the 2017 International Conference on Big Data and Cloud Computing, pp. 1–8.

Xu, H. and Kifer, D. (2015) 'Sentiment analysis of social media data using linguistic features', Journal of Computer Science and Technology, 30(6), pp. 1101–1113.

Agarwal, A. and Mittal, M. (2017) 'Sentiment analysis on social media data using deep learning techniques', International Journal of Engineering Research & Technology (IJERT), 6(7), pp. 1012–1016.

Sundararajan, V. and Shmueli, G. (2014) 'Predicting public sentiment in social media: A multi-faceted approach', Journal of the American Statistical Association, 109(506), pp. 1236–1248.

Liu, Y. and Zhang, Z. (2021) 'Sentiment analysis of social media data using hybrid deep learning models', Journal of Computational Social Science, 4(2), pp. 135–148.

Nguyen, T. T., Nguyen, D. T. and Lee, H. (2022) 'A survey on sentiment analysis techniques for social media data: Applications, challenges, and future directions', International Journal of Data Science and Analytics, 13(1), pp. 1–18.

Zhang, X., Xu, B. and Chen, X. (2023) 'Multimodal sentiment analysis for social media posts: Integrating textual, visual, and acoustic features', Journal of Machine Learning Research, 24(18), pp. 1–23.

Kumar, A. and Sahu, S. (2023) 'Fine-tuning transformer models for sentiment classification of social media posts', Journal of Artificial Intelligence Research, 71(5), pp. 411–430.

Chen, L. and Yang, T. (2024) 'Sentiment analysis of political discourse on Twitter using graph convolutional networks', Proceedings of the 2024 International Conference on Computational Linguistics, pp. 2256–2265.

Appendix A　Mathematical Proofs

Appendix B Psuedo Code

Appendix C Time-series Results

LIST OF PUBLICATIONS

Journal Articles

Qasem, S. N., Shamsuddin, S. M., Hashim, S. Z. M., Darus, M., & AlShammari, E. (2013). Memetic multiobjective particle swarm optimization based radial basis function network for classification problems. Information Sciences, 239, 165–190. https://doi.org/10.1016/j.ins.2013.03.021. (Q1, IF: 4.305)

Qasem, S. N., & Shamsuddin, S. M. (2011). Radial basis function network based on time variant multi-objective particle swarm optimization for medical diseases diagnosis. Applied Soft Computing, 11(1), 1427–1438. https://doi.org/10.1016/j.asoc.2010.04.014. (Q1, IF:3.907)

Shen, L. W., Asmuni, H., & Weng, F. C. (2015). A modified migrating bird optimization for university course timetabling problem. Jurnal Teknologi, 72(1), 89–96. https://doi.org/10.11113/jt.v72.2949. (Indexed by SCOPUS)

Conference Proceedings

Muhamad, W. Z. A. W., Jamaludin, K. R., Ramlie, F., Harudin, N., & Jaafar, N. N. (2017). Criteria selection for MBA programme based on the mahalanobis Taguchi system and the Kanri Distance Calculator. In 2017 IEEE 15th Student Conference on Research and Development (SCOReD) (pp. 220–223). IEEE. https://doi.org/10.1109/SCORED.2017.8305390. (Indexed by SCOPUS).