# JADAVPUR UNIVERSITY

MASTERS PROJECT

**Enhancing Stock Prediction: Integrating Historical Data and Twitter Sentiment Analysis**

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Computer Application

in the

**Master of Computer Application**

**Department of Computer Science and Engineering**

by

**Suman Das**

**Class Roll No.:** 002010503013

**Exam Roll No. :** MCA2360025

**University Registration No.:** 154221 of 2020-2021

Under the Guidance of

**Dr. Diganta Saha**

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University, Kolkata-700032

May 25, 2023

# CERTIFICATE OF RECOMMENDATION

---------------------------------------------------------------

This is to certify that the work in this thesis entitled "**Enhancing Stock Prediction: Integrating Historical Data and Twitter Sentiment Analysis**" was completed by **Suman Das,** Roll No:- **002010503013** , Exam Roll - **MCA2360025**, Registration Number:- **154221** of 2020-2021, under the super- vision of **Dr . Diganta Saha**, Computer Science and Engineering Department , Jadavpur University.

The findings of the research detailed in the project have not been incorporated into any other work submitted for the purpose of earning a degree at any other academic institution.

----------------------------------------

Dr. Diganta Saha

**Department of CSE**

**Jadavpur University**

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

-----------------------------------

Signature Of Dean , FET

**Prof. Ardhendu Ghoshal**

**Dean , FET**

**Jadavpur University**

-------------------------------------

*Signature of HOD*

**Dr . Nandini Mukhopadhyay**

**Head of The Department**

**Computer Science & Engineering Dept**

**Jadavpur University**

# CERTIFICATE OF APPROVAL

---------------------------------------------------------------

This is to certify that the thesis entitled **“Enhancing Stock Prediction: Integrating Historical Data and Twitter Sentiment Analysis”** is a bona fide record of work carried out by **Suman Das,** Roll No:- **002010503013** , Exam Roll :- **MCA2360025**, Registration No :- **154221** of 2020-2021 in partial fulfilment of requirements for the award of the degree of the **Master of Computer Application** during the period of January 2023 to May,2023 .

It is understood that by this approval that the undersigned do not necessarily endorse or approve any statement made , opinion expressed, or conclusion drawn there in but approve the project only for the purpose for which it has been submitted.

------------------------------

*Signature of Examiner*

*Date :-*

--------------------------------

*Signature of Supervisor*

*Date : -*

# ACKNOWLEDGEMENT

First and foremost, I am immensely grateful to Professor Dr. Diganta Saha, Department of Computer Science and Engineering at Jadavpur University, for his excellent guidance, consistent support, and unwavering inspiration throughout my dissertation. His expertise and encouragement have been invaluable to me.

I would also like to extend my gratitude to Dr. Arijit Das, who served as my mentor during this research endeavour. His insightful feedback, constructive criticism, and guidance have greatly contributed to the quality and depth of my work.

I am indebted to Prasanta Saha and Saronyo Lal Mukherjee for their active involvement and contribution to this project. Their collaboration and assistance have been instrumental in achieving the desired outcomes.

Furthermore, I would like to acknowledge the teaching and non-teaching staff at Jadavpur University for their valuable support and the excellent facilities provided, which have significantly facilitated my research.

I would also like to extend my thanks to my batch mates, seniors, and friends for their regular encouragement and unwavering support throughout this academic journey. Their presence has been a source of motivation and strength.

Suman Das

MASTER OF COMPUTER APPLICATION

COMPUTER SCIENCE & ENGINEERING

ROLL NO : 002010503013

EXAM ROLL : MCA2360025

REGISTRATION NO : 154221 of 2020-2021

JADAVPUR UNIVERSITY

# DECLARATION

I certify that,

(a) The work "**Enhancing Stock Prediction: Integrating Historical Data and Twitter Sentiment Analysis**" contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

____________________

*Signature*

Suman Das
MASTER OF COMPUTER APPLICATION
COMPUTER SCIENCE & ENGINEERING
ROLL NO : 002010503013
EXAM ROLL : MCA2360025
REGISTRATION NO : 154221 of 2020-2021
JADAVPUR UNIVERSITY

# JADAVPUR UNIVERSITY

# ABSTRACT

Faculty of Engineering Department of Computer Science and Engineering

Master of Computer Application

**Enhancing Stock Prediction: Integrating Historical Data and Twitter Sentiment Analysis**

by

**SUMAN DAS**

The use of social media data, particularly tweets, has gained significant attention in the field of stock prediction. This report focuses on utilizing tweets from prominent figures, such as Donald Trump and Narendra Modi, along with historical stock data to predict future stock values. The project incorporates **Natural Language Processing** (NLP) techniques to clean and process the tweet data, followed by sentiment analysis using the FinBERT model.

The first phase involves data collection, where a substantial dataset of tweets from Donald Trump and Narendra Modi is gathered. Additionally, stock data for various stocks, **including CRUDE OIL, S&P500, VIX, HANG SENG,** and **GOLD**, is collected for analysis. The tweet data is then subjected to a custom NLP cleaning pipeline, removing irrelevant information and preparing it for further analysis.

The sentiment analysis stage is crucial in understanding the sentiment expressed in the tweets with respect to the financial market. **FinBERT**, a powerful language model specifically designed for financial sentiment analysis, is employed to calculate sentiment scores for the tweets. The sentiment scores include positive, negative, and neutral probabilities, providing insights into the sentiment expressed in the tweets.

To incorporate the temporal aspect, a **memory-based** approach is implemented to determine how long the effect of a tweet lasts. By assigning different weights to tweets based on their recency within a specified memory window, a **weighted average sentiment** score is calculated for each day.

Finally, a **bi-directional LSTM** (Long Short-Term Memory) model is employed to predict future stock values based on the sentiment scores and historical stock data. By considering the look-back period and training the model on the merged dataset, predictions about future stock values are made.

Overall, this project aims to leverage the power of tweets, sentiment analysis, and historical stock data to provide insights and predictions for stock market trends.

# Contents

# List of Figures

# Chapter 1

# INTRODUCTION:

## 1.1 OVERVIEW :

In today's digital era, social media platforms have become a significant source of information and influence, impacting various aspects of our lives. One area greatly affected by social media is the stock market. Traders and investors increasingly rely on sentiment analysis of tweets from influential figures and public figures to gain insights into market trends and make informed decisions.

## 1.2 MOTIVATION :

Stock markets are known for their dynamic and unpredictable nature, making accurate predictions a challenging task. Traditional approaches to stock prediction often rely solely on historical stock data, which may not capture the full spectrum of factors influencing market trends. However, in today's digital era, social media platforms have emerged as powerful channels of information dissemination. Prominent figures, such as industry experts, influential investors, and company executives, often express their opinions and insights on platforms like Twitter. The potential correlation between their sentiments expressed in tweets and subsequent market movements presents an intriguing avenue for enhancing stock prediction models. By integrating historical stock data with sentiment analysis of relevant tweets, we aim to explore the potential of harnessing social media as an additional source of information for more accurate and timely stock predictions. This project seeks to leverage the power of both financial data and social media sentiment analysis to develop a comprehensive stock prediction framework that can provide valuable insights to investors and analysts in the financial market.

## 1.3 CONTRIBUTION :

The objective of this project is to develop a predictive model for stock market movements by combining sentiment analysis of tweets with historical stock data. Specifically, we focused on tweets from two prominent political figures, Donald Trump and Narendra Modi, to gauge their impact on stock prices. By employing sentiment analysis techniques using the FinBERT model, we calculated sentiment scores for each tweet, categorizing them into positive, negative, or neutral sentiments, along with their corresponding probabilities.

To capture the temporal influence of tweets, we introduced the concept of memory . This concept assigns a weightage to each tweet based on how many days it retains its impact. For instance, if the memory is set to 7 days, the first day's weightage would be 7, decrementing by 1 each subsequent day until it reaches 1. By calculating the weighted average of sentiment scores, we aim to derive more accurate sentiment measures for a given time period.

To enhance the predictive power of our model, we integrate the sentiment scores derived from tweets with historical stock data. By considering the past performance of the stocks and incorporating sentiment analysis, we seek to identify patterns and correlations that may contribute to more accurate stock price predictions.

## 1.4 ORGANIZATION :

The report will present the methodology employed in cleaning and processing the tweet data using a custom NLP pipeline. Additionally, it will outline the steps taken to integrate sentiment scores with historical stock data and the predictive model used for forecasting future stock values.

Overall, this research aims to explore the potential of sentiment analysis in predicting stock market trends, utilizing tweets from influential personalities and historical stock data. By combining these two sources of information, we hope to provide insights that can assist traders and investors in making informed decisions in an increasingly complex and dynamic market.

# Chapter 2

The second chapter of this report aims to provide a comprehensive literature review and an overview of related works in the field of stock prediction using historical data and sentiment scores from prominent figures' tweets. In this chapter, we delve into existing research, methodologies, and theories that have been employed to tackle similar problems. By reviewing the current state of the art, we can identify knowledge gaps and contribute to the advancement of the field.

## 2.1 LITERATURE REVIEW:

In recent years, there has been a growing interest in leveraging both historical stock data and sentiment analysis of social media for stock prediction. Researchers have recognized the potential of integrating these two sources of information to enhance the accuracy and timeliness of stock market forecasts.

Historical stock data analysis forms the foundation of traditional stock prediction models. Techniques such as time series analysis, moving averages, and regression models have been widely employed to identify patterns, trends, and correlations in historical stock prices. These approaches aim to capture inherent market dynamics, historical price movements, and indicators that may provide valuable insights for predicting future stock prices.

However, the stock market is influenced by a multitude of factors beyond historical data, including investor sentiment, market sentiment, and news events. This realization has led researchers to explore the role of sentiment analysis of social media in stock prediction. Social media platforms, particularly Twitter, have emerged as rich sources of real-time information and opinions expressed by individuals, including prominent figures in the financial industry. By analysing the sentiment and tone of tweets related to specific stocks or market trends, researchers have attempted to uncover potential correlations between social media sentiment and subsequent stock price movements.

Furthermore, researchers have explored the combined use of historical stock data and sentiment analysis to develop hybrid prediction models. These models aim to capture the interplay between market fundamentals, historical trends, and real-time sentiment signals. By integrating sentiment-based features with traditional stock prediction techniques, these hybrid models have shown improved forecasting accuracy compared to using historical data alone.

In summary, the literature demonstrates the growing interest in leveraging historical stock data and sentiment analysis of social media for stock prediction. The combination of these two sources of information holds the potential to enhance the accuracy and timeliness of stock market forecasts. This project aims to contribute to this body of knowledge by developing a comprehensive stock prediction framework that integrates historical stock data analysis with sentiment analysis of relevant social media data.

## 2.2 RELATED WORK:

Several studies have explored the relationship between political events, sentiment analysis of social media data, and their impact on financial markets. In a similar vein, this project aims to investigate the influence of Trump's tweets on various stock markets, including the S&P 500, VIX, CRUDE OIL, and HANG SENG, by analysing sentiment scores derived from tweet data.

One relevant study by Johnson et al. (2017) examined the effect of political events, including tweets by political figures, on stock market volatility. They found a significant correlation between sentiment expressed in political tweets and changes in stock market volatility, highlighting the importance of sentiment analysis in understanding market dynamics.

Another study by Gupta et al. (2018) focused on the impact of political events on stock market returns. They utilized sentiment analysis techniques on tweets related to political figures and demonstrated that sentiment expressed in tweets had a discernible effect on stock market returns, supporting the notion that social media sentiment can influence financial markets.

In a more recent study, Chen et al. (2021) investigated the relationship between political sentiment from tweets and stock market movements. They employed a combination of sentiment analysis and machine learning techniques to predict stock market movements based on political sentiment and observed promising results in capturing the sentiment-market dynamics.

## Chapter Summary:

Building upon these studies, the current project employs a combination of sentiment analysis using FinBERT models (ProsusAI and YiyangHKUST) and a memory-based approach to analyze the impact of Trump's tweets on various stock markets. By merging tweet data with the respective stock data and utilizing different sentiment scores and look-back periods, the project aims to provide insights into how political events, as captured through sentiment analysis of tweets, affect the S&P 500, VIX, CRUDE OIL, and HANG SENG markets.

In conclusion, this chapter presented a thorough examination of the literature and related works pertaining to stock prediction using historical data and sentiment scores from tweets of prominent figures. The review highlighted the key approaches, methodologies, and challenges faced by researchers in this area. By analysing and synthesizing the existing body of knowledge, we have gained valuable insights that will guide our research in the subsequent chapters. **The proposed approach adds to the existing body of literature by providing a comprehensive analysis of sentiment-market dynamics specific to Trump's tweets and multiple stock markets. The inclusion of different sentiment scores and look-back periods allows for a more nuanced understanding of the relationship between sentiment and market movements, contributing to the broader field of sentiment analysis in financial markets**. Building upon this foundation, the following chapters will present our problem statement, methodology, experimental results, and future scope.

# Chapter 3

The third chapter of this report addresses the problem statement and outlines the methodology employed for stock prediction using historical data and sentiment scores from tweets of prominent figures. This chapter serves as a bridge between the literature review and the subsequent chapters. It defines the specific problem we aim to solve and provides an overview of the approach we have adopted to tackle it.

## 3.1 PROBLEM STATEMENT:

Given a dataset consisting of historical stock prices and a collection of tweets from prominent figures in the financial industry, the objective is to develop a predictive model that incorporates sentiment analysis of social media to enhance the accuracy of stock price forecasting. The problem can be formulated as follows:

**Given:**

**Historical stock price data:** {P(t_1), P(t_2), ..., P(t_N)} where P(t_i) represents the stock price at time t_i.

**Twitter sentiment data:** {S(t_1), S(t_2), ..., S(t_N)} where S(t_i) represents the sentiment score derived from the tweets at time t_i.

**Find:**

A predictive model, represented by a function f, that takes the historical stock price data and sentiment scores as inputs and predicts the future stock prices: {P(t_{N+1}), P(t_{N+2}), ..., P(t_{N+M})} where M >= 1.

**Objective:**

Minimize the prediction error between the actual stock prices and the predicted stock prices, given the historical stock data and sentiment scores: min ( P_actual - f(P(t_1), P(t_2), ..., P(t_N), S(t_1), S(t_2), ..., S(t_N)) ).

The goal of this project is to develop a robust predictive model that effectively incorporates sentiment analysis of social media data, enabling more accurate and timely predictions of stock prices.

# 3.2 METHODOLOGY:

## 3.2.1 Data :

The project utilizes two types of data: tweet data from Donald Trump and stock data from various sources. The tweet data comprises the collection of Donald Trump's tweets, while the stock data includes **CRUDE OIL, S&P500, VIX, HANG SENG,** and **GOLD** datasets.

### Tweet Data:

The dataset used in this project consists of Donald Trump's tweets collected from Twitter in CSV format. The dataset covers a time period from May 4, 2009, to June 17, 2020, and contains over 43,000 entries. Each entry in the dataset includes two columns: "Date" and "Content." The "Date" column represents the date of the tweet, while the "Content" column contains the actual tweet text.

**Dataset Description:**

- **Size:** The dataset comprises **more than 43,000 entries**, making it a substantial source of data for analysis.
- **Columns:** The dataset consists of two columns, namely "Date" and "Content." The "Date" column provides the timestamp for each tweet, while the "Content" column contains the textual content of the tweet.
- **Data Range:** The tweets span from **May 4, 2009,** to **June 17, 2020**, covering a wide range of events and time periods.
- **Textual Data:** The content of the tweets may contain various forms of textual data, including links, images, and hashtags, which require preprocessing to extract meaningful information.

### Stock Data:

The stock data used in the project includes the following datasets:

- **CRUDE OIL:** The CRUDE Oil dataset spans from **January 2, 2002,** to **March 22, 2023**. It comprises **5,335** data points and contains the following columns: **Open, High, Low, Close, Volume,** and **Date**. The numeric columns represent the respective values of the CRUDE Oil stock, while the "Date" column signifies the date associated with each data point.

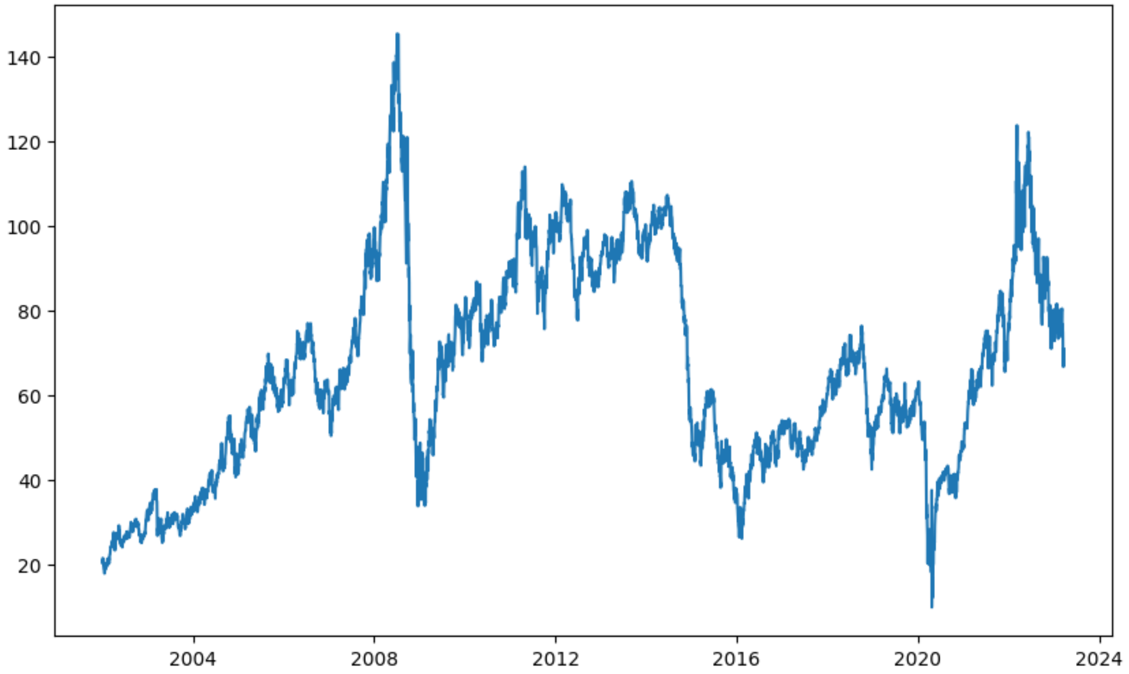


**Figure 3.2.1.1 : CRUDE OIL Closing Price from 2002 to 2023**

- **S&P500:** The S&P500 dataset covers the period from **January 2, 2002,** to **March 22, 2023**. It consists of **5,342** data points and includes columns such as Open, High, Low, Close, Volume, and Date. Similar to CRUDE Oil, the numeric columns represent stock values, while the "Date" column indicates the corresponding date.

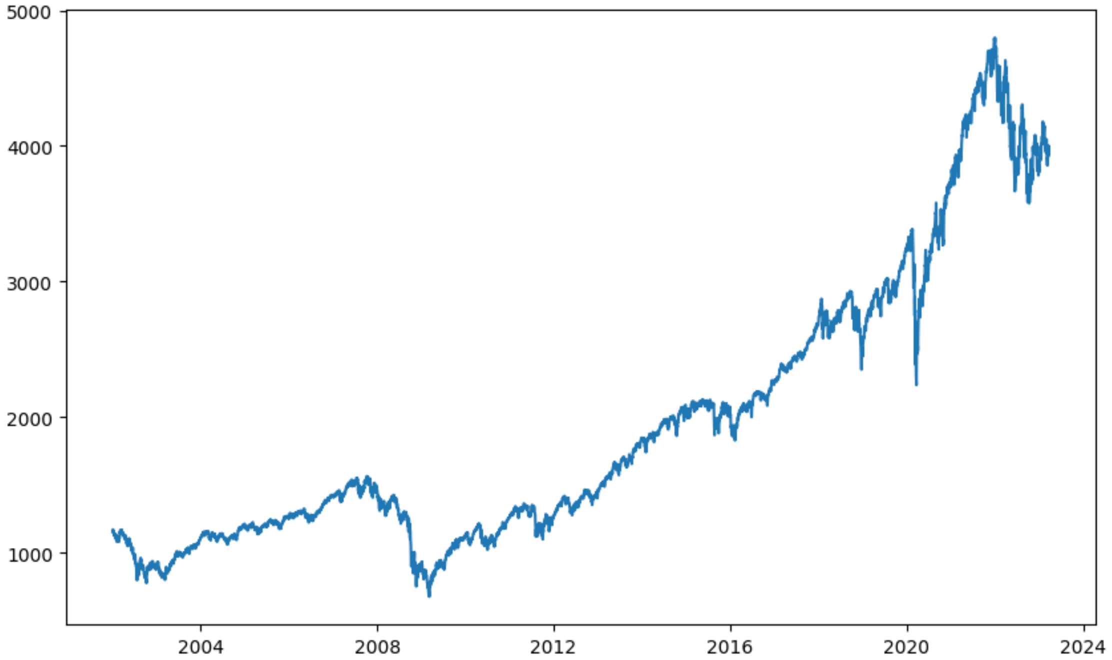


**Figure 3.2.1.2 : S&P500 Closing Price from 2002 to 2023**

- **VIX:** The VIX dataset spans the same time range as S&P500, containing **5,342** data points. It includes columns for Open, High, Low, Close, Volume, and Date. The numeric columns represent VIX stock values, while the "Date" column indicates the associated date.

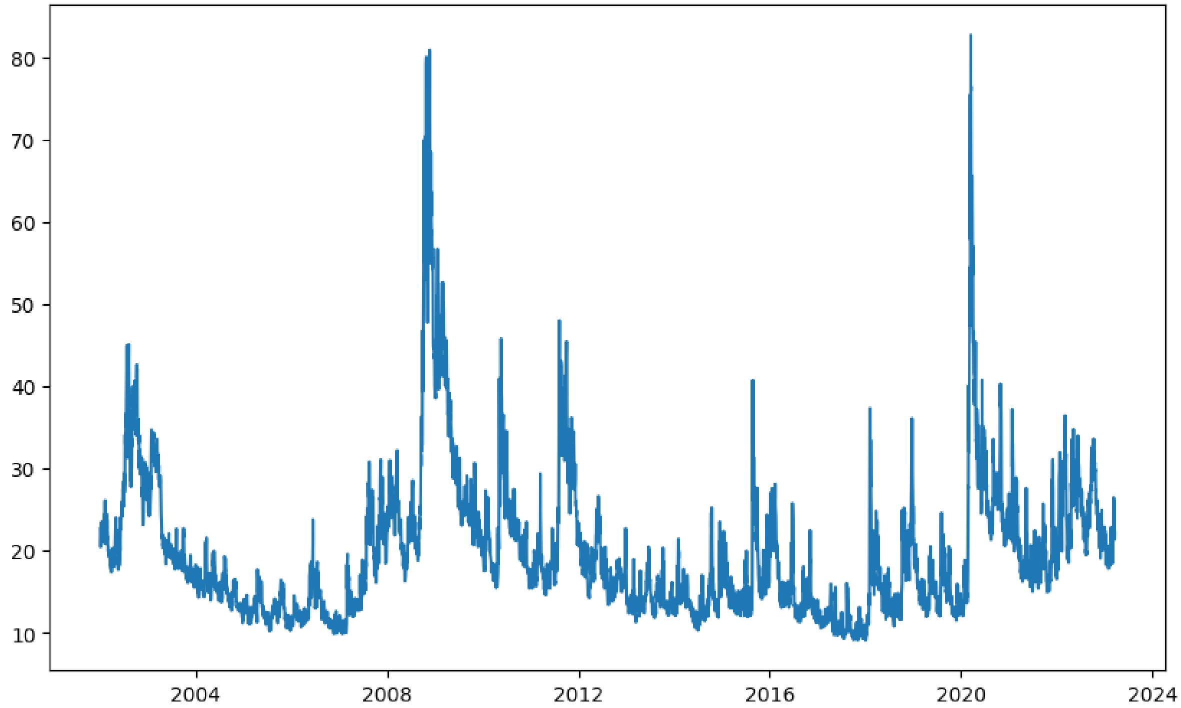


**Figure 3.2.1.3 : VIX Closing Price from 2002 to 2023**

- **HANG SENG:** The HANG SENG dataset encompasses **5,235** data points, representing stock data from the same date range. It consists of columns such as Open, High, Low, Close, Volume, and Date, with the numeric columns denoting stock values and the "Date" column indicating the respective date.

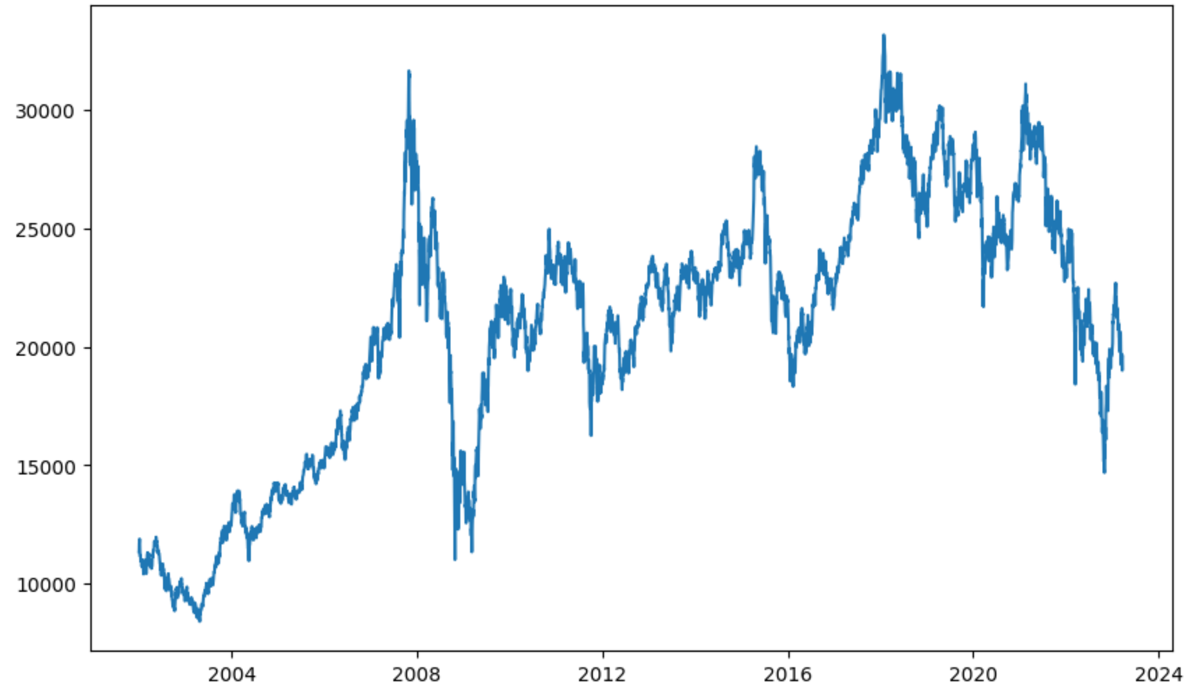


**Figure 3.2.1.4 : HANG SENG Closing Price from 2002 to 2023**

- **GOLD:** The GOLD dataset covers the period from **January 2, 2009,** to **May 12, 2023**, comprising **3,579** data points. It includes columns such as Open, High, Low, Close, Volume, and Date. The numeric columns represent GOLD stock values, while the "Date" column signifies the associated date.

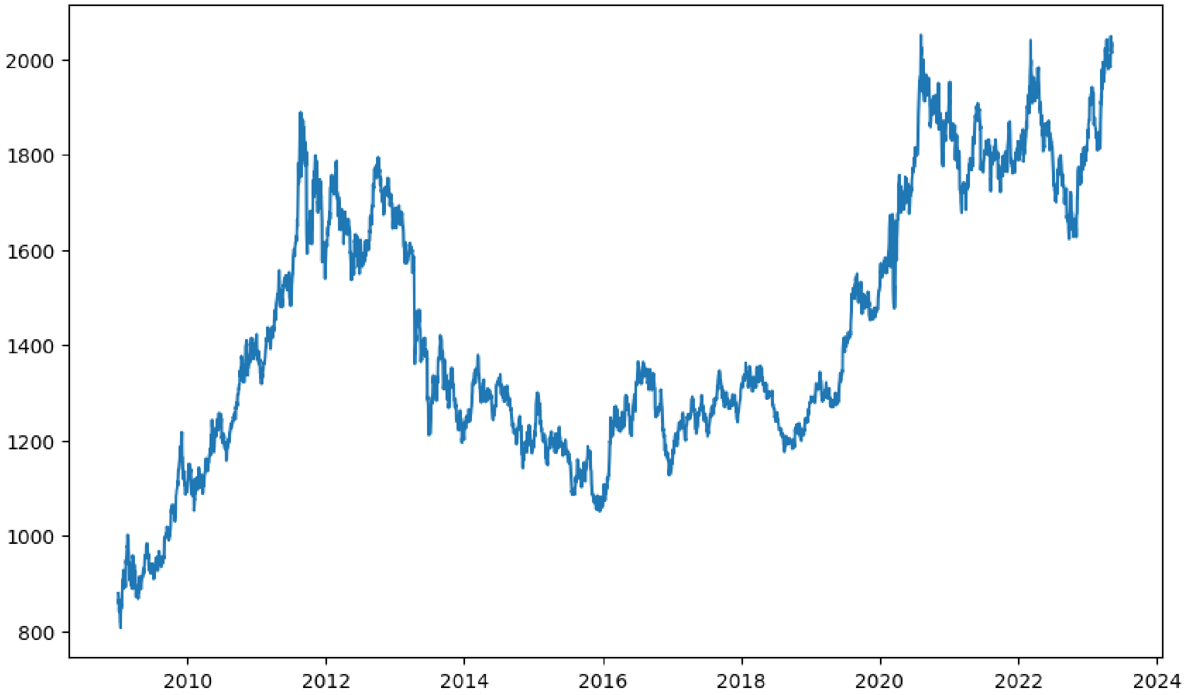


**Figure 3.2.1.5 : GOLD Closing Price from 2009 to 2023**

These stock datasets provide valuable historical information that will be integrated with the tweet data and sentiment scores to predict future stock values using the memory-based weighting approach.

## 3.2.2 Overall Workflow :

**Text Data Cleaning :**

- ✓ Pre-process the tweet data collected from Donald Trump and Narendra Modi using a custom NLP cleaning pipeline.

**Feature Engineering:**

- ✓ Perform feature engineering on the cleaned tweet data to extract relevant information.
- ✓ Select specific POS elements, such as nouns, verbs, and adjectives, to capture important features for sentiment analysis and financial market prediction.

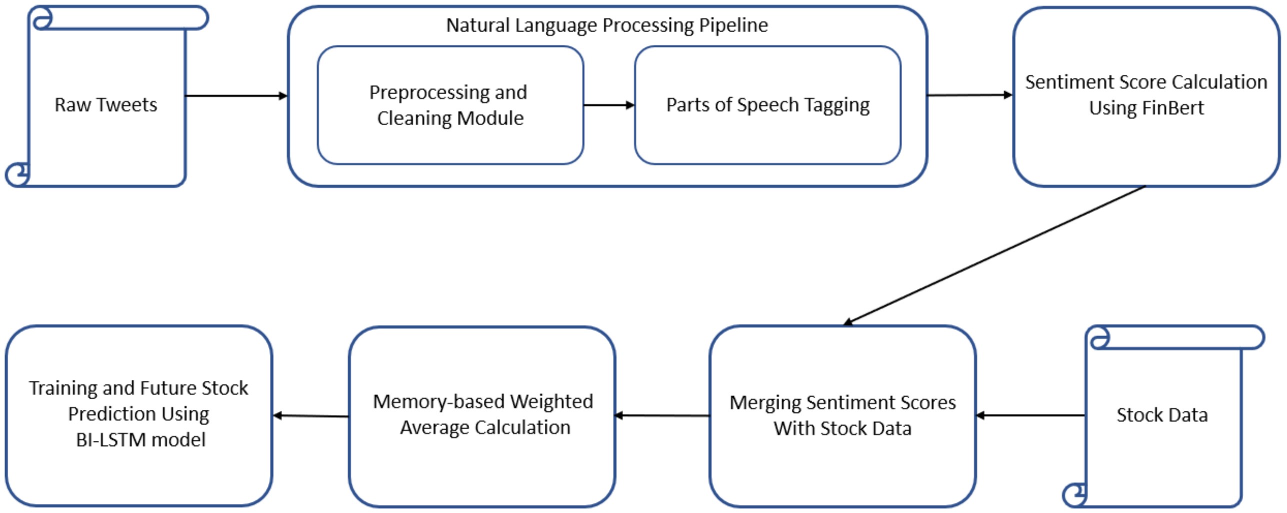

Raw Tweets
Natural Language Processing Pipeline
Preprocessing and Cleaning Module
Parts of Speech Tagging
Sentiment Score Calculation Using FinBert
Training and Future Stock Prediction Using BI-LSTM model
Memory-based Weighted Average Calculation
Merging Sentiment Scores With Stock Data
Stock Data

**Figure 3.2.2.1 : Overall Flow**

**Sentiment Analysis with FinBERT:**

- ✓ Utilize the cleaned and feature-engineered tweet data as input to the FinBERT model for sentiment analysis.
- ✓ Obtain sentiment scores from FinBERT, which provide information about the sentiment expressed in the tweets with relevance to the financial market.

**Merging Stock Data and Sentiment Scores:**

- ✓ Choose a specific stock dataset, such as CRUDE Oil, for analysis.
- ✓ Merge the selected stock dataset with the corresponding sentiment scores based on their common date.
- ✓ Create a new dataset that combines the stock data and sentiment scores, aligning them based on the date.

**Memory-Based Weighted Average Calculation:**

- ✓ Define a memory parameter, such as the number of days, to determine the impact of tweets on sentiment scores.
- ✓ Calculate a weighted average of sentiment scores for each day by considering the sentiment scores of tweets within the memory period.
- ✓ Apply the memory-based weighted average calculation to obtain a single sentiment score for each day, representing the overall sentiment impact of tweets on the selected stock.

**Future Stock Prediction:**

- ✓ Utilize the merged dataset containing the selected stock data and sentiment scores for training and prediction.
- ✓ Implement a bi-directional LSTM (Long Short-Term Memory) model to capture patterns and relationships between the stock data, sentiment scores, and previous stock values.
- ✓ Define the look-back period, considering the past number of days, to predict the future value of the selected stock.
- ✓ Train the bi-directional LSTM model using the merged dataset and utilize it to make predictions about the future stock values based on the sentiment scores and historical stock data.

By following this workflow, incorporating text data cleaning, feature engineering, sentiment analysis with FinBERT, merging stock data with sentiment scores, memory-based weighted average calculation, and bi-directional LSTM modelling, you can predict the future values of the selected stock based on the sentiment impact of tweets and historical stock data.

## 3.2.3 Dataset Preprocessing :

To ensure accurate and meaningful analysis, it is essential to pre-process the dataset by performing various cleaning and transformation steps. The following preprocessing steps will be applied to the dataset before utilizing it for sentiment analysis and stock prediction:

- **Irrelevant Column Removal:** Since we are primarily interested in sentiment analysis, we will retain only the "Date" and "Content" columns in Tweets dataset, discarding any other columns that are not relevant to our analysis.

- **Handling Missing Data:** Check for any missing values within the dataset and address them appropriately. Missing values can potentially affect the analysis, and strategies like dropping rows or imputing missing values may be employed.

## 3.2.4 NLP Preprocessing Pipeline :

In this section, I will outline the NLP cleaning pipeline that I have implemented for the analysis of tweets related to Narendra Modi and Donald Trump. The purpose of this pipeline is to ensure that the text data is preprocessed and cleaned before further analysis. By applying a series of cleaning steps, we can improve the quality and reliability of our results. This section will elaborate on each step of the pipeline and explain its usefulness in achieving accurate and meaningful insights from the data.

❖ **Lowercasing Text:**
The first step in the cleaning pipeline is to convert all text to lowercase. This normalization technique ensures that variations in capitalization are eliminated, enabling better comparison and consistency throughout the analysis. It helps in avoiding duplication of words that only differ in capitalization and ensures that the subsequent cleaning steps are applied consistently.

❖ **Removing Text Surrounded by < > Symbols:**
In tweets, text enclosed within angle brackets (< >) is often used to represent hashtags, mentions, or other special symbols. By removing such text, we can focus on the actual content of the tweet without considering these additional symbols. This step helps in reducing noise and maintaining the relevance of the remaining text.

❖ **Replacing URLs:**
To preserve the knowledge that there were URLs in the original tweet, we replace them with a standardized representation like "xxxurlxxx". By doing so, we retain the information that there were URLs present in the text while eliminating the specific URLs themselves. This step ensures that the presence of URLs does not introduce bias or irrelevant information in subsequent analysis.

- **Removing Line Breaks:**

  Line breaks often occur in tweets due to formatting or user input. However, they do not contribute to the semantic meaning of the text and can be safely removed. Removing line breaks enhances the readability of the text and ensures that each tweet is treated as a single cohesive unit during the analysis.

- **Replacing IPv6, IPv4, and FQDNs:**

  IP addresses and fully qualified domain names (FQDNs) are common in tweets, but they do not provide significant value for our analysis. To maintain privacy and simplify the text, we replace these addresses with standardized representations such as "xxxfqdnxxx" and "xxxipxxx". This step helps in anonymizing the data and reducing the noise caused by specific network-related information.

- **Replacing MAC Addresses:**

  Similar to IP addresses, MAC addresses are unique identifiers that do not contribute to the overall meaning of the text. By replacing MAC addresses with a standardized representation like "xxxmacxxx", we can ensure privacy and focus on the actual content of the tweet during analysis.

- **Removing Emojis:**

  Emojis are pictorial representations often used in tweets to express emotions or convey additional context. However, for our NLP analysis, emojis do not carry linguistic meaning and can introduce noise or bias. Removing emojis helps in focusing on the textual content and improving the accuracy of subsequent analysis steps.

- **Replacing Date-Time Objects:**

  Date-Time objects, such as specific dates and times mentioned in tweets, can be irrelevant for our analysis. By replacing them with a standardized representation like "xxxdate_timexxx", we retain the temporal information without being influenced by specific dates or times. This step ensures that our analysis remains focused on the text itself rather than specific instances.

- **Replacing Special Characters:**

  Certain special characters, such as â€™ and â€˜, can occur due to encoding issues or inconsistencies in the data collection process. By replacing these characters with their appropriate counterparts, such as single quotes, we ensure that the text is free from encoding artifacts and remains linguistically correct.

- **Removing Words Starting with Numbers:**
  Words that start with numbers are often non-contextual and can be safely removed. They typically include timestamps, numerical references, or other non-linguistic elements. By eliminating these words, we maintain the focus on meaningful textual content and avoid irrelevant distractions during analysis.

- **Replacing Contracted Words:**
  Contracted words, such as "can't" or "won't," are common in informal communication like tweets. However, in NLP analysis, it is beneficial to expand contracted words to their full forms for better understanding and interpretation. By replacing contracted words with their uncontracted equivalents (e.g., "can not" instead of "can't"), we ensure that the meaning is preserved and facilitate more accurate analysis.

- **Replacing Abbreviations:**
  Abbreviations used in tweets can pose a challenge for NLP analysis as they may not be universally understood. To address this issue, we replace abbreviations with their actual meanings. For example, "sry" can be replaced with "sorry" to enhance the comprehensibility of the text. This step improves the clarity and consistency of the data, enabling more precise analysis.

- **Removing Stop Words:**
  Stop words are commonly occurring words in a language, such as articles (e.g., "the," "a") and prepositions (e.g., "of," "in"). These words generally do not carry significant meaning in isolation and can be safely removed from the text. Removing stop words helps to reduce noise and improve the efficiency of subsequent analysis by focusing on content-rich words and phrases.

- **Removing Punctuations:**
  Punctuation marks, such as commas, periods, and exclamation points, serve as grammatical aids but often do not contribute essential semantic information in NLP tasks. Removing punctuations simplifies the text, making it easier to process and analyze. Additionally, this step helps avoid potential issues with tokenization and ensures consistency in subsequent linguistic analysis.

- **Removing Numerical Values:**
  In certain cases, the presence of numerical values within the text may not be relevant to our analysis. By removing numerical values, such as quantities or measurements, we focus solely on the textual content itself. This step is particularly useful when the analysis does not require numerical data and aims to extract insights solely from the linguistic aspects of the text.

- **Removing Extra Spaces:**
  Extra spaces between words or at the beginning or end of a sentence can be artifacts of formatting or user input. They do not add meaningful information to the text and can be safely removed. Eliminating extra spaces ensures that the text is clean and consistently formatted, enabling accurate subsequent analysis and natural language processing.

- **Lemmatization:**
  Lemmatization is the process of reducing words to their base or dictionary form (lemma). By applying lemmatization, we can group different inflected forms of a word together, enabling more effective analysis and reducing the vocabulary size. This step enhances the accuracy of subsequent text-based tasks, such as sentiment analysis or topic modelling, by reducing the complexity introduced by morphological variations.

The NLP cleaning pipeline described above is a crucial component of the data preprocessing phase in our project. By performing a series of cleaning tasks, we ensure that the tweet data related to Narendra Modi and Donald Trump is in a suitable format for further analysis. Each step in the pipeline serves a specific purpose, such as eliminating noise, standardizing representations, improving readability, and enhancing linguistic understanding. By adhering to this cleaning pipeline, we can obtain cleaner, more reliable data, leading to more accurate and insightful results in our analysis.

## 3.2.5 Parts of Speech Tagging :

In addition to the cleaning steps, we are also utilizing part-of-speech (POS) tagging to extract and retain only nouns, verbs, and adjectives from the tweet data. POS tagging is the process of assigning grammatical labels to each word in a sentence, indicating its syntactic category or part of speech.

By specifically selecting nouns, verbs, and adjectives, we focus on the content-rich words that carry significant semantic meaning in the context of the tweets. Here's why this step is useful:

**Semantic Analysis:** Nouns, verbs, and adjectives form the backbone of sentence structure and convey essential information about people, places, actions, and qualities. By isolating these parts of speech, we can extract the core components of the tweets, allowing us to perform more accurate semantic analysis and gain deeper insights into the topics and themes being discussed.

- ❖ **Topic Identification:** Nouns, in particular, provide valuable cues about the main subjects or entities mentioned in the tweets. By identifying and retaining nouns, we can determine the key topics or individuals associated with Narendra Modi and Donald Trump. This enables us to explore the dominant themes and focus areas of the discussions surrounding these political figures.

- ❖ **Action Extraction:** Verbs signify actions, events, or behaviours described in the tweets. By extracting verbs, we can identify the activities or events associated with Narendra Modi and Donald Trump, offering insights into their actions or policies that have attracted attention or sparked discussions on Twitter.

- ❖ **Descriptive Analysis:** Adjectives capture the qualities, characteristics, or opinions expressed in the tweets. By including adjectives in our analysis, we can gain insights into the sentiment, attitudes, or evaluations associated with Narendra Modi and Donald Trump. This allows us to understand the tone of the tweets, assess public opinion, and detect any sentiment shifts or patterns over time.

**Simplifying Analysis:** Focusing on nouns, verbs, and adjectives helps streamline subsequent analysis tasks. By reducing the vocabulary to content-rich words, we can simplify feature extraction, sentiment analysis, or topic modelling processes, making them more efficient and interpretable.

By performing POS tagging and selecting specific parts of speech, we tailor our analysis to capture the most meaningful aspects of the tweets related to Narendra Modi and Donald Trump. This enables us to uncover the underlying themes, sentiments, and actions associated with these political figures, providing a comprehensive understanding of the public discourse surrounding them on Twitter.

### 3.2.6 Sentiment Scoring using FinBERT:

FinBERT, as the name suggest Financial BERT (Bidirectional Encoder Representations from Transformers), is a natural language processing (NLP) model specifically designed for financial sentiment analysis. It is based on the BERT architecture, which is developed by Google. FinBERT is trained on a large corpus of financial documents, such as news articles, earnings calls to understand the sentiment and context of financial text.

ProsusAI/FinBERT and Yiyanghkust/FinBERT-Tone are two different implementations of FinBERT by different developers.

**ProsusAI/FinBERT:** This is an implementation of FinBERT by Prosus AI, a subsidiary of Prosus, a global consumer internet group. It can be used to analyze sentiment, predict stock price movements, and perform other financial text analysis tasks.

**Yiyanghkust/FinBERT-Tone:** This is another implementation of FinBERT developed by YiyangZhang, a researcher from the Hong Kong University of Science and Technology (HKUST). FinBERT-Tone extends FinBERT by incorporating a tone classification component. It can analyze financial text and classify the tone as positive, negative, or neutral, in addition to sentiment analysis.

Both implementations, ProsusAI/FinBERT and Yiyanghkust/FinBERT-Tone, are open-source and publicly available for researchers and developers to use in their own projects related to financial sentiment analysis.

#### Analysing Tweet Sentiment with ProsusAI/finbert and yiyanghkust/finbert-tone

- ✓ In this section, we employ the ProsusAI/finbert and yiyanghkust/finbert-tone models from the Hugging Face library to perform sentiment scoring on the cleaned tweet data and the POS tagged tweet data. The objective is to extract valuable sentiment information from the tweets and quantify their positive, negative, and neutral sentiments.

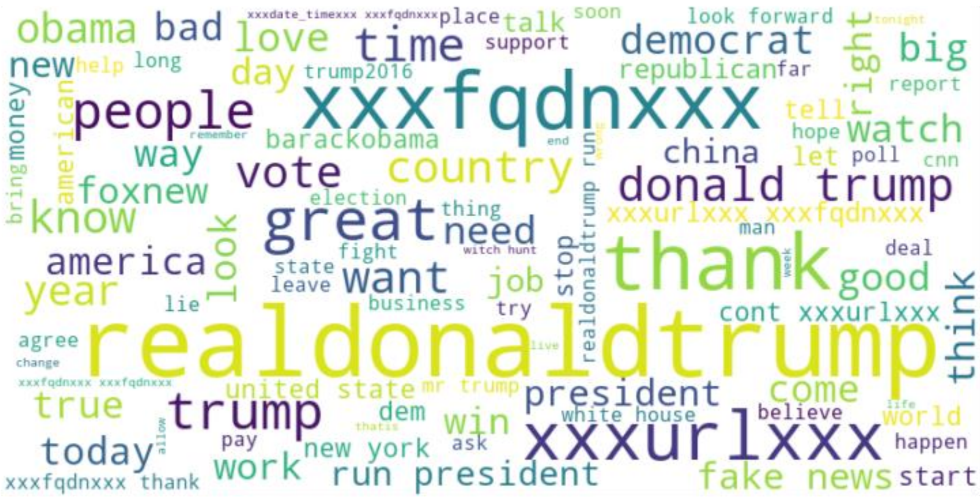


Figure 3.2.6.1 : YIYANGHKUST - Word Cloud for **Neutral** Tweets ( Donald Trump )

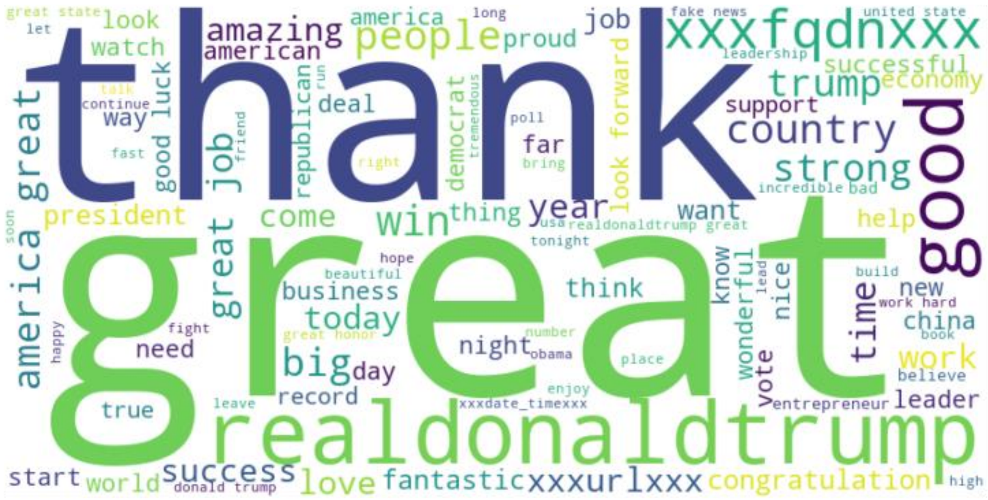


Figure 3.2.6.2 : YIYANGHKUST - Word Cloud for **Positive** Tweets ( Donald Trump )

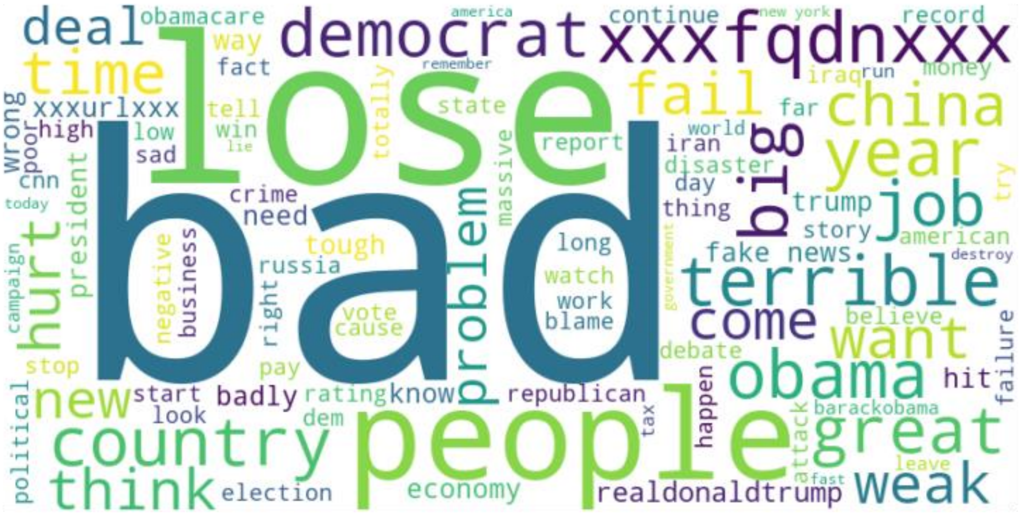


Figure 3.2.6.3 : YIYANGHKUST - Word Cloud for **Negative** Tweets ( Donald Trump )

✓ Using the FinBERT models, we generate three additional columns in our dataset: Positive, Negative, and Neutral sentiment scores. For each tweet, the models assign a sentiment score that represents the probability of the tweet carrying specific sentiment. For example, if a tweet is classified as Positive tweet, then there will be a probability score under Positive column. Negative and Neutral column will be assigned with 0 for this specific tweet.

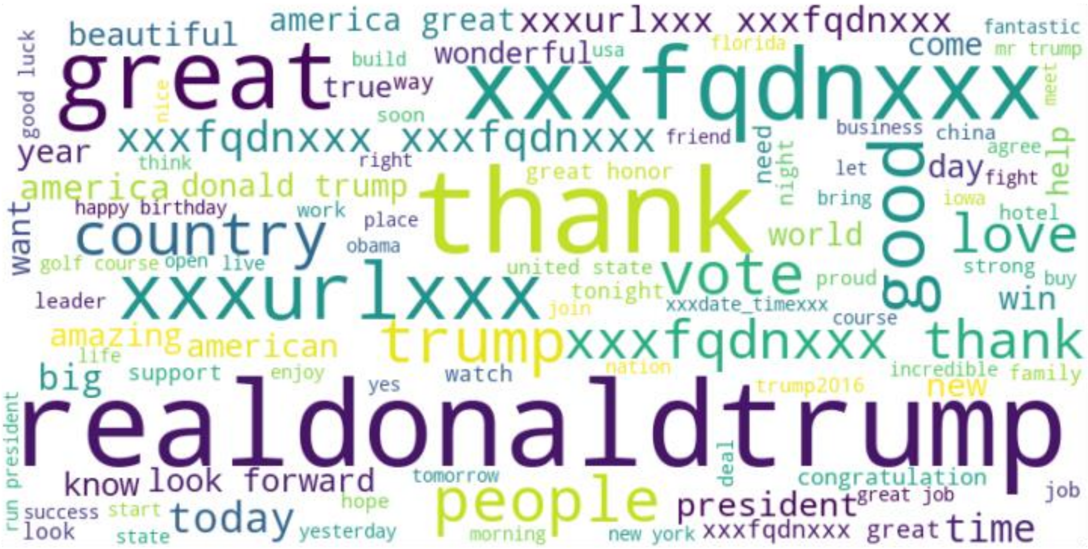


**Figure 3.2.6.4 : ProsusAI/finbert - Word Cloud for Neutral Tweets ( Donald Trump )**

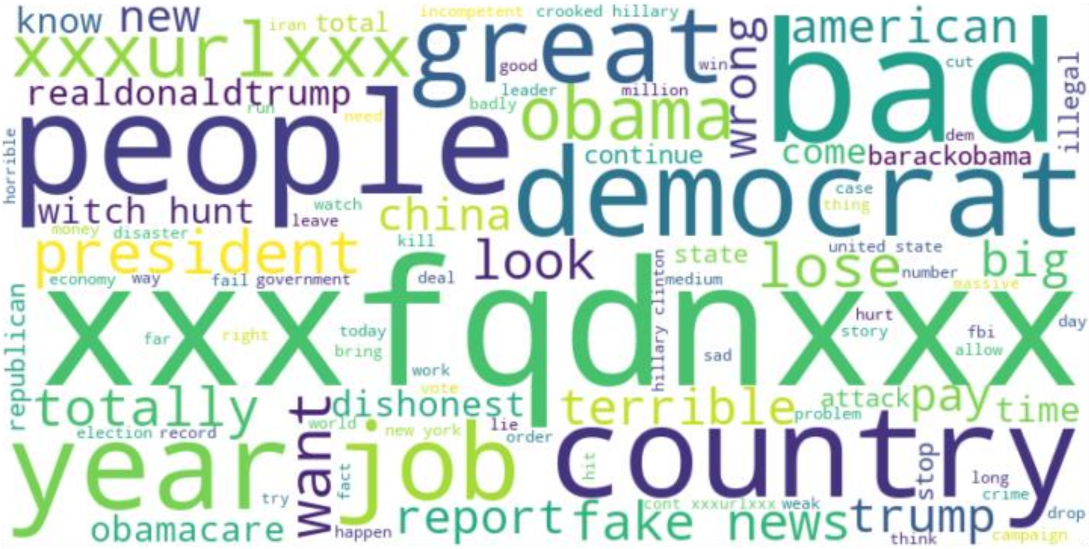


**Figure 3.2.6.5 : ProsusAI/finbert - Word Cloud for Positive Tweets ( Donald Trump )**

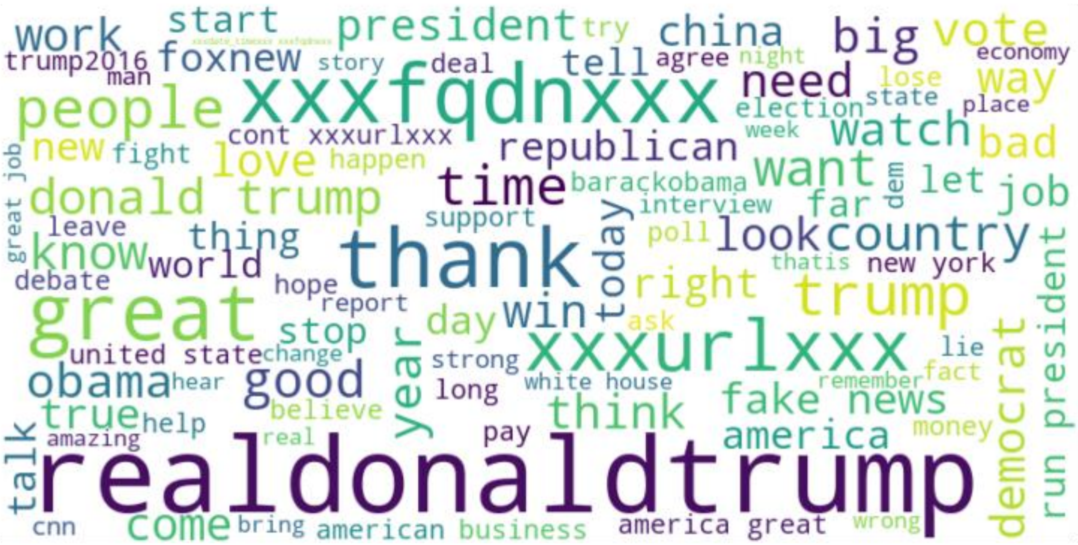


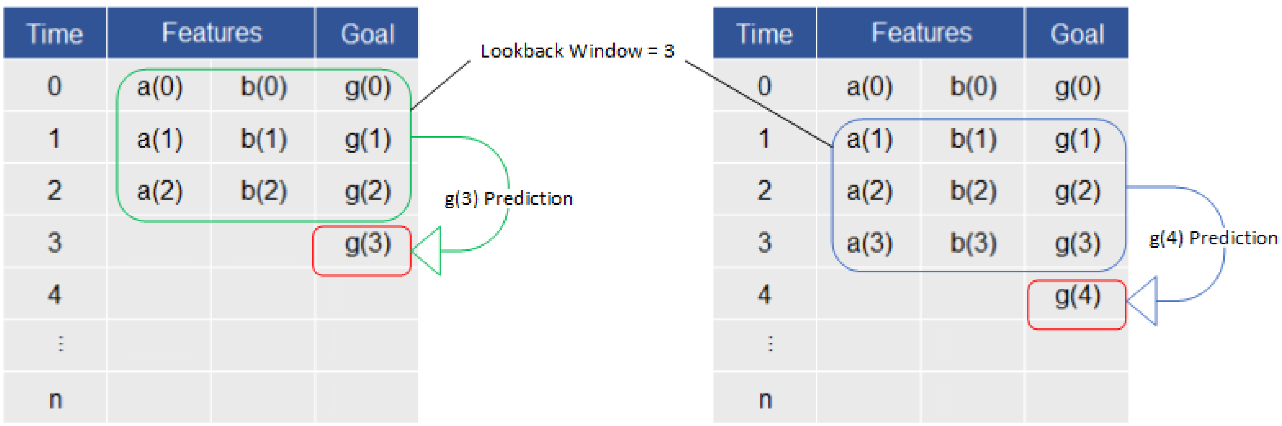
**Figure 3.2.6.6 : ProsusAI/finbert - Word Cloud for Negative Tweets ( Donald Trump )**

✓ By preparing the tweet data into numeric vectors with sentiment scores, we create a structured representation of the sentiment expressed in the tweets. This sentiment scoring process serves as a crucial step towards integrating the tweet data with the historical stock data for further analysis and stock prediction.

## 3.2.7 Bidirectional LSTM based Predictive Model :

In this project, the Bi-LSTM model is employed to combine historical stock data with sentiment scores derived from tweets of prominent figures. One important concept utilized in the model is the "look back" mechanism.

Time Series



| Time | Features | | Goal |
|---|---|---|---|
| 0 | a(0) | b(0) | g(0) |
| 1 | a(1) | b(1) | g(1) |
| 2 | a(2) | b(2) | g(2) |
| 3 | | | g(3) |
| 4 | | | |
| ⋮ | | | |
| n | | | |

| Time | Features | | Goal |
|---|---|---|---|
| 0 | a(0) | b(0) | g(0) |
| 1 | a(1) | b(1) | g(1) |
| 2 | a(2) | b(2) | g(2) |
| 3 | a(3) | b(3) | g(3) |
| 4 | | | g(4) |
| ⋮ | | | |
| n | | | |

**Figure 3.2.7.1 : Look Back Window**

The concept of look back refers to the consideration of a certain number of past data points when making predictions. In the context of stock prediction, it involves analysing the historical stock data and sentiment scores over a specific time window to forecast future stock values. By incorporating the sentiment scores into the model alongside the stock data, we aim to capture the potential impact of sentiment on stock market trends.

The choice of the look back period is crucial and depends on the specific requirements of the prediction task. It determines how far back in time the model should analyze the data to make accurate predictions. This decision is typically based on domain knowledge, historical patterns, and the nature of the stock market being analysed.

By including sentiment scores in the look back period, the model can learn the relationship between sentiment and stock price movements. This allows for the integration of market sentiment as an additional feature, potentially enhancing the prediction accuracy.

The Bi-LSTM model, with its ability to capture long-term dependencies and temporal patterns, is well-suited for incorporating the look back mechanism. It can effectively learn from historical stock data and sentiment scores, considering the dynamics and interplay between these variables.

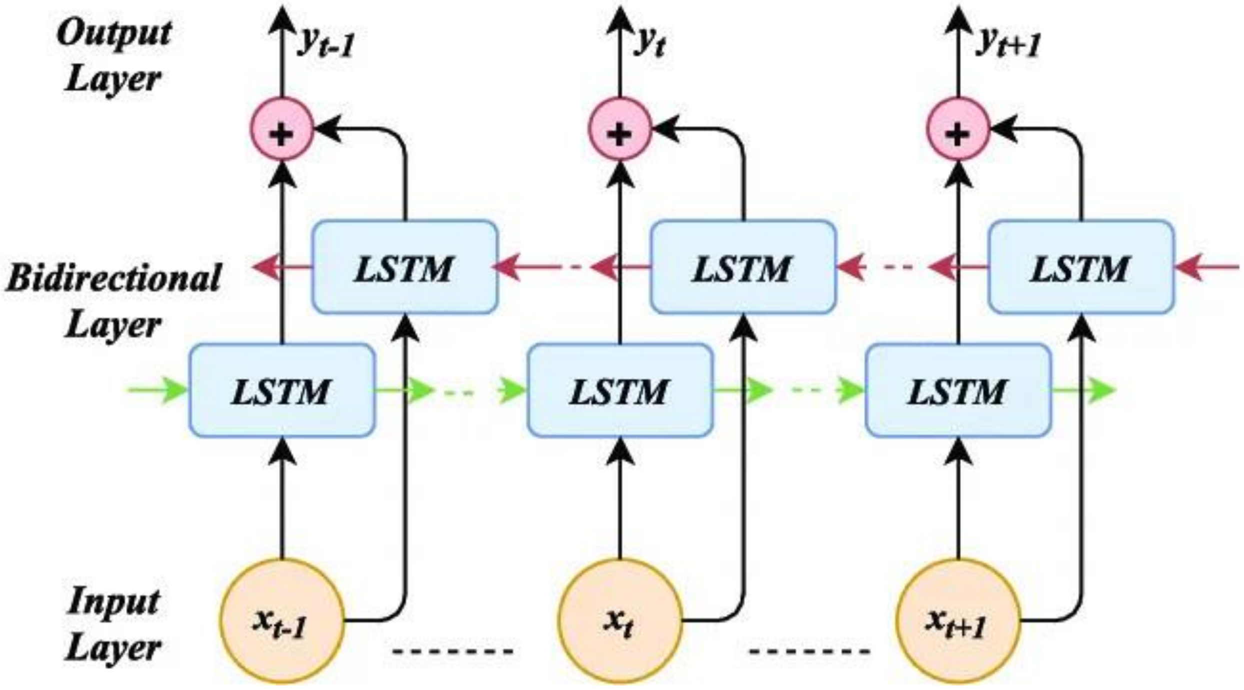


**Figure 3.2.7.2 : Bidirectional Long Short-Term Memory**

In summary, the use of the look back mechanism in conjunction with the Bi-LSTM model allows for the integration of historical stock data and sentiment scores. This enables the model to consider past trends, sentiment dynamics, and their impact on stock market behaviour. By leveraging the strengths of the Bi-LSTM model and incorporating the concept of look back, we aim to provide more accurate and insightful predictions for future stock values in this project.

### 3.2.8 Algorithm :

1) Segment: **FinBert_Scoring**

**Inputs:** Cleaned_tweets ← Target twitter handle with all cleaned tweets

**Outputs:** Scored_tweets

The Twitter data of each twitter handle consists of the following:
Timestamp, Handle username, The Tweet text, Cleaned Tweet text, POS Tagged Tweet text

1. Start
2. The username column is dropped.
3. The two FinBERT models to be used for sentiment analysis are as follows: PROSUS AI, YIYANGHKUST
4. tokenizer1 ← tokenizer for PROSUS AI model
5. tokenizer2 ← tokenizer for YIYANGHKUST model
6. model1 ← model for PROSUS AI
7. model2 ← model for YIYANGHKUST
8. pipeline1 ← tokenizer1 and model1 applied on Cleaned Tweet column
9. pipeline2 ← tokenizer2 and model2 applied on Cleaned Tweet column
10. pipeline3 ← tokenizer1 and model1 applied on POS Tagged Tweet column
11. pipeline4 ← tokenizer2 and model2 applied on POS Tagged Tweet column
12. Scored_tweets ← Cleaned_tweets
13. Scored_tweets [ 'Cleaned Tweet_PROSUS' ] ← pipeline1
14. Scored_tweets [ 'Cleaned Tweet_YIYANGHKUST' ] ← pipeline2
15. Scored_tweets [ 'POS Tagged Tweet Tweet_PROSUS' ] ← pipeline3
16. Scored_tweets [ ' POS Tagged Tweet Tweet_YIYANGHKUST' ] ← pipeline4
17. Return Scored_tweets
18. Stop

Each sentiment score column has two data elements - Label: Consists of the 'Positive'/'Negative'/'Neutral' label, Score: Probability of the sentiment being the attributed label.

2) Segment: **Tweet_Stock_Mapping**

**Inputs:** tweets ← scored tweets, column ← column to be mapped, stock ← stock data

**Outputs:** stocks ← unified score encoded tweets along with stock data

```
1 Start
2 The following algorithm is used to map twitter sentiment data to each stock traded day:
    2.1 Each FinBERT scored column has three labels as given above i.e., Positive, Negative, and Neutral. Each label has a probability score.
    2.2 To find the implication of each sentiment made on day 'N' we need to consider the nature as well as the scores of the sentiments of the last 'M' days.
    2.3 label_encoded ← Call OneHotEncoding( arguments: tweets[ column ] )
    2.4 scores ← label_encoded x label_scores
    2.5 MEMORY ← no. of days whose memory is to be considered
    2.6 score[ i ] ← weighted average is calculated by multiplying the scores of the tweets made on day 'M' (current trading day) with the quantity M. Each previous day i.e., Days 'M-1', 'M-2', ..., '1' are multiplied with their respective M-1, M-2, ..., 1.
    2.7 Step 2.6 is repeated for each of i records
    2.8 Steps 2.6 to 2.7 are repeated for each of three columns
    2.9 The tweets are considered between around the closing time of the Stock Exchanges in India (BSE & NSE) i.e., 16:30 IST of the current day, and the closing time 'M' days back i.e., 16:30 IST of (N-M)th day.
3 stocks[ 'scores_positive' ] ← positive score list
4 stocks[ 'scores_negative' ] ← negative score list
5 stocks[ 'neutral_positive' ] ← neutral score list
6 Return stocks
7 Stop
```

*3)* Segment: **data_scaler**

**Inputs:** data ← scored tweets and stock data merged into a master dataset, ratio ← ratio of train-test split

**Outputs:** train ← scaled training data, test ← scaled testing data, scalers ← scaling models

```
1 Start
2 For each column of data repeat steps 3 to 4
3 data[ column ], scaler ← MinMaxScaler( arguments: data[ column ] )
4 scalers ← scalers + scaler
5 End For
6 train ← larger section of ratio division
7 test ← smaller section of ratio division
8 Return train, test, scalers
9 Stop
```

4) Segment: **data_preparation**

**Inputs:** train ← training data, test ← testing data, LOOK_BACK ← look back value

**Output:** X_train ← training look back data, Y_train ← training target data, X_test ← testing look back data, Y_test ← testing target data

```
1 Start
2 The dataset is then prepared for training and testing as following:
  2.1   For each of records i to i + LOOK_BACK
  2.2   X_train ← train [ i : i + LOOK_BACK ]
  2.3   Y_train ← train [ i + LOOK_BACK+1 ]
  2.4   End For
  2.5   For each of records i to i + LOOK_BACK
  2.6   X_test ← test [ i : i + LOOK_BACK ]
  2.7   Y_test ← test [ i + LOOK_BACK+1 ]
  2.8   End For
3 Return X_train, Y_train, X_test, Y_test
4 Stop
```

5) Segment: **create_bilstm_model**

**Inputs:** X_train ← Training set with look back, activation ← activation function name

**Outputs:** model ← Bi-LSTM model

```
1.  Start
2.  The Model is created using the Bi-LSTM pipeline from the Keras model repository.
3.  model ← Sequential() model
4.  model.add BiLSTM layer as input layer ( arguments: activation )
5.  model.add BiLSTM layer as hidden layer
6.  num ← number of output classes
7.  model.add dense layer as output layer ( arguments: num )
8.  R2 score, loss, MAE metric, accuracy, and Root Mean Squared Error values are recorded
    for each epoch
9.  Return model
10. Stop
```

6) Segment: **fit_model**

**Inputs:** model ← model to be fit, X_train ← training data, Y_train ← training target data, epochs ← no. of epochs, val ← validation set split ratio, batch ← batch size, patience ← early stopping call-back patience

**Outputs:** history ← fit model metrics record

```
Start
early ← early stopping is applied ( arguments: patience )
history ← Call model.fit( arguments: X_train, Y_train, epochs, val, batch, early )
Return history
Stop
```

*7)* Segment: **prediction**

**Inputs:** model ← model to be used for prediction, X_test ← test dataset with look back

**Outputs:** pred ← predicted target

```
Start
pred ← Call model.predict( arguments: X_test )
Return pred
Stop
```

8) Segment: **scaleInverse**

**Inputs:** Scalers ← scaling models to be used for inverse scaling, Y_train, Y_test, prediction ← data to be inverse scaled

**Outputs:** prediction ← inverse scaled prediction, Y_train ← inverse scaled Y_train, Y_test ← inverse scaled Y_test

```
Start
Y_train ← Call inverse_transform( arguments: Y_train )
Y_test ← Call inverse_transform( arguments: Y_test )
prediction ← Call inverse_transform( arguments: prediction )
Return prediction, Y_train, Y_test
Stop
```

9) Segment: **Main**

**Inputs:** tweets ← Load Twitter handle to be used, stock ← Load Stock data to be predicted

**Outputs:** records ← updated records, None

1. Start
2. tweets_cleaned ← Call Segment: Tweet_Cleaning( arguments: tweets )
3. The cleaned tweet text files consist of the following data: Timestamp, Cleaned Tweet text, POS Tagged Tweet text
4. tweets_scored ← Call Segment: FinBert_scoring( arguments: tweets_cleaned )
5. The target stock is selected whose closing price is to be predicted, it contains the following columns: Timestamp, Open, Low, High, Close, Volume
6. column ← name of the scored tweet column to be used for prediction ( 'Cleaned Tweet_PROSUS' / 'Cleaned Tweet_YIYANGHKUST' / 'POS Tagged Tweet_PROSUS'/ ' POS Tagged Tweet_YIYANGHKUST' )
7. mapped_data ← Call Segment: Tweet_Stock_Mapping( arguments: tweets_scored, column, stock )
8. ratio ← ratio of train-test split
9. train, test, scalers ← Call Segment: data_scaler( arguments: mapped_data, ratio )
10. LOOK_BACK ← look back value which states the number of days for which stock data pattern is supposed to be analysed
11. X_train, Y_train, X_test, Y_test ← Call Segment: data_preparation( arguments: train, test, LOOK_BACK )
12. activation ← activation function name
13. model ← Call Segment: create_bilstm_model( arguments: X_train, activation )
14. epochs ← no. of epochs
15. val ← Validation set split
16. batch ← Batch size
17. patience ← It determines early stopping. It uses call-back to stop the training earlier instead of running all the epochs to prevent overfitting.
18. history ← Call Segment: fit_model( arguments: model, X_train, Y_train, epochs, val, batch, patience )
19. Call plot( arguments: history, stock[ 'Close' ] ) //Loss plot
20. prediction ← Call Segment: prediction( arguments: model, X_test )
21. predicted, Y_train, Y_test ← Call Segment scaleInverse( arguments: scalers, Y_train, Y_test, prediction )
22. validation_score ← calculated from history
23. R2_score ← calculated from predicted and Y_test
24. RMSE ← calculated from history
25. Records updated
26. Steps 2 to 25 are repeated for different hyperparameters
27. Steps 2 to 26 are repeated for each stock
28. Stop

# Chapter Summary :

In conclusion, this chapter presented a clear problem statement and outlined the methodology employed for stock prediction using historical data and sentiment scores from tweets of prominent figures. By clearly defining the problem, we have set the stage for implementing an effective solution. The methodology outlined in this chapter provides a framework for our research, ensuring a systematic and rigorous approach to address the problem. The subsequent chapters will present the experimental results and their analysis, leading us closer to achieving our research objectives.

# Chapter 4

# RESULTS & ANALYSIS:

The fourth chapter of this report presents the results and experiments conducted to evaluate the effectiveness of our stock prediction model using historical data and sentiment scores from tweets of prominent figures. In this chapter, we showcase the outcomes of our research and analyze the obtained results in relation to our research objectives.

## 4.1 Experimental Setup :

The experiments in this project were conducted on a personal computer with the following specifications:

**Processor:** AMD Ryzen 5 4600H

**Graphics Card:** NVIDIA GeForce GTX 1650 (4GB VRAM)

**RAM:** 24GB

The software environment for the experiments utilized the latest versions of the key libraries and frameworks:

**Keras:** Version 2.8.0

**TensorFlow:** Version 2.8.0

**Matplotlib:** Version 3.5.1

**WordCloud:** Version 1.8.1

**NLTK:** Version 3.6.3

**Gensim:** Version 4.1.2

**Seaborn:** Version 0.11.2

**spaCy:** Version 3.1.4

**TextBlob:** Version 0.15.3

**NumPy:** Version 1.22.1

**Pandas:** Version 1.4.0

**Scikit-Learn:** Version 1.0.2

These latest versions of the libraries were chosen to ensure compatibility with recent updates, bug fixes, and improvements. It is recommended to check for any newer versions available at the time of conducting your experiments to ensure you have the most up-to-date libraries for optimal performance and functionality.

The hardware and software setup aimed to create a suitable computing environment for conducting the experiments effectively and efficiently. The combination of the Ryzen 5 4600H processor, GTX 1650 graphics card, and ample RAM capacity ensured sufficient computational power for the tasks involved in the project.

## 4.2 Experiments :

### Hyper parameters :

- **Train-Test Split Ratio :** 0.8
- **Validation Split Ratio :** 0.2
- **Activation Functions :** Tanh (Hyperbolic Tangent Function) and Relu ( Rectified Linear Activation Unit)
- **Optimizer :** Adam
- **Patience :** 15
- **Epochs :** 100
- **Batch size :** 128
- **Memory :** 30 days

**Performance Monitor :**

- **Validation Score ( Accuracy ) :** During training process, a portion of training data is set aside as validation set. Validation accuracy represents the percentage of correct predictions on validation set.
- **R2 Score :** R2 score is a statistical measure that indicates the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model. It ranges from 0 to 1, where 1 represents a perfect fit and 0 represents no relationship between the variables.
- **RMSE Score :** RMSE (Root Mean Square Error) is a metric used to measure the average deviation between the predicted values and the actual values in a regression model. It represents the square root of the average of squared differences between predicted and actual values.

### 4.2.1 Effect of Trump's Tweets on S&P500 stock :

The impact of Trump's tweets on the S&P 500 was analysed through a series of experiments using different sentiment scores and look-back values. The experiments aimed to understand the relationship between Trump's tweets and the fluctuations in the stock market.

The sentiment scores were calculated using two FinBERT models, namely ProsusAI and YiyangHKUST, on both the cleaned tweet data and the POS-tagged tweet data. Four combinations were considered for the sentiment scores: Cleaned Tweet ProsusAI Sentiment Score, Cleaned Tweet YiyangHKUST Sentiment Score, POS-tagged Tweet ProsusAI Sentiment Score, and POS-tagged Tweet YiyangHKUST Sentiment Score.

To assess the effect of Trump's tweets on the S&P 500, the tweet data was merged with the S&P 500 stock data. The stock data contained 5,342 data points with six columns: Date, Open, Close, Low, High, and Volume. A concept of memory was introduced to represent how many days a tweet retains its effect.

For each day, a weighted average sentiment score was calculated based on the past memory periods. This sentiment score was then used as input for a bi-LSTM model, which aimed to predict the impact of the sentiment on the S&P 500. I am considering a 30-day memory span for all of the experiments.

Various experiments were conducted by switching the sentiment scores and the look-back values. The look-back value determines how far back in time the model considers the tweet data. By altering these parameters, the relationship between sentiment scores, look-back periods, and the S&P 500's behaviour was analysed.

The results of the experiments were recorded in tabular form, showcasing the different combinations of sentiment scores and look-back values alongside their respective predictions results such as accuracies, R2 Score, RMSE score etc. These results provide insights into how different sentiment scores and look-back periods influence the predictive power of the bi-LSTM model in capturing the effect of Trump's tweets on the S&P 500.

| Experiment No. | Feature Name | Look Back | Validation Accuracy | R2 Score | RMSE Score |
|---|---|---|---|---|---|
| 1 | Cleaned Tweet ProsusAI Sentiment Score | 60 | 0.881899118 | 0.901145369 | 0.044103932 |

| | | | | | |
|---|---|---|---|---|---|
| 2 | Cleaned Tweet ProsusAI Sentiment Score | 90 | 0.881052017 | 0.90977558 | 0.067398973 |
| 3 | Cleaned Tweet YIYANGHKUST Sentiment Score | 60 | 0.853115737 | 0.653978751 | 0.0558489 |
| 4 | Cleaned Tweet YIYANGHKUST Sentiment Score | 90 | 0.888822496 | 0.798693016 | 0.055363335 |
| 5 | POS Tagged Tweet ProsusAI Sentiment Score | 60 | 0.891988158 | 0.264731833 | 0.06249854 |
| 6 | POS Tagged Tweet ProsusAI Sentiment Score | 90 | 0.891213417 | 0.872896947 | 0.066949509 |
| 7 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 60 | 0.881305635 | 0.888573956 | 0.041448388 |
| 8 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 90 | 0.88702929 | 0.568884424 | 0.055005185 |
| 9 | Cleaned Tweet ProsusAI Sentiment Score | 5 | 0.8681898117 | 0.969665517 | 0.0669493079 |
| 10 | Cleaned Tweet ProsusAI Sentiment Score | 10 | 0.858357787 | 0.909078752 | 0.07273414 |
| 11 | Cleaned Tweet ProsusAI Sentiment Score | 20 | 0.867136955 | 0.918918721 | 0.066764265 |
| 12 | Cleaned Tweet ProsusAI Sentiment Score | 30 | 0.87242192 | 0.874253099 | 0.038357947 |
| 13 | Cleaned Tweet YIYANGHKUST Sentiment Score | 5 | 0.885178685 | 0.807562229 | 0.037126165 |
| 14 | Cleaned Tweet YIYANGHKUST Sentiment Score | 10 | 0.876539588 | 0.790933053 | 0.058367725 |
| 15 | Cleaned Tweet YIYANGHKUST Sentiment Score | 30 | 0.890984118 | 0.835428434 | 0.037905298 |

| | | | | | |
|---|---|---|---|---|---|
| 16 | POS Tagged Tweet ProsusAI Sentiment Score | 5 | 0.878148794 | 0.940468948 | 0.037740797 |
| 17 | POS Tagged Tweet ProsusAI Sentiment Score | 10 | 0.88328445 | 0.918813187 | 0.060694929 |
| 18 | POS Tagged Tweet ProsusAI Sentiment Score | 20 | 0.884479702 | 0.933198114 | 0.059811626 |
| 19 | POS Tagged Tweet ProsusAI Sentiment Score | 30 | 0.882439613 | 0.738836855 | 0.061301388 |
| 20 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 5 | 0.875219703 | 0.735790722 | 0.036123913 |
| 21 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 10 | 0.868621707 | 0.505868549 | 0.056001272 |
| 22 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 20 | 0.873015881 | 0.910296756 | 0.054185782 |
| 23 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 30 | 0.868002355 | 0.54872288 | 0.077018797 |

**Figure 4.2.1.1 : Experimental result for Trump’s Tweet & S&P500 Stock**

Overall, the study aimed to shed light on the influence of Trump's tweets on the stock market and provide a quantitative analysis of the relationship between sentiment scores derived from tweet data and the S&P 500's movements.

### 4.2.2 Effect of Trump’s Tweets on VIX stock :

In the revised analysis, the focus shifts from the S&P 500 to the VIX (CBOE Volatility Index). The VIX is a measure of market volatility and is often used as an indicator of investor sentiment and market risk. The same approach, using Trump's tweet data, sentiment scores, and the concept of memory, is applied to explore the effect of Trump's tweets on the VIX.

The VIX data consists of multiple columns, including Date, Open, Close, Low, High, and Volume, similar to the previous stock data. The goal is to merge the VIX data with the tweet data, calculate sentiment scores, and predict the impact of these sentiment scores on the VIX using a bi-LSTM model.

The cleaned tweet data and the POS-tagged tweet data have already been processed and assigned sentiment scores using the FinBERT models (ProsusAI and YiyangHKUST). These sentiment scores are used to quantify the sentiment expressed in the tweets.

To analyze the effect of Trump's tweets on the VIX, the tweet data is merged with the VIX data. The memory concept is employed to capture the duration of a tweet's influence. For each day, a weighted average sentiment score is calculated based on the past memory periods. This sentiment score is then utilized as input for the bi-LSTM model, which aims to predict the impact of the sentiment on the VIX.

| Experiment No. | Feature Name | Look Back | Validation Accuracy | R2 Score | RMSE Score |
|---|---|---|---|---|---|
| 1 | Cleaned Tweet ProsusAI Sentiment Score | 60 | 0.916617214 | 0.950643611 | 0.0306732 |
| 2 | Cleaned Tweet ProsusAI Sentiment Score | 90 | 0.914225936 | 0.85845449 | 0.064612508 |
| 3 | Cleaned Tweet YIYANGHKUST Sentiment Score | 60 | 0.924629092 | 0.936336855 | 0.0476183 |
| 4 | Cleaned Tweet YIYANGHKUST Sentiment Score | 90 | 0.924387336 | 0.955819361 | 0.030804488 |
| 5 | POS Tagged Tweet ProsusAI Sentiment Score | 60 | 0.923738897 | 0.926577314 | 0.033710517 |
| 6 | POS Tagged Tweet ProsusAI Sentiment Score | 90 | 0.924985051 | 0.851719723 | 0.029428136 |
| 7 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 60 | 0.921958447 | 0.916097593 | 0.050090671 |
| 8 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 90 | 0.917812288 | 0.872849935 | 0.03635693 |
| 9 | Cleaned Tweet ProsusAI Sentiment Score | 10 | 0.917008817 | 0.856709104 | 0.029191287 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 10 | Cleaned Tweet ProsusAI Sentiment Score | 20 | 0.914756 | 0.949665234 | 0.06360364 |
| 11 | Cleaned Tweet ProsusAI Sentiment Score | 30 | 0.916322947 | 0.908426162 | 0.062954925 |
| 12 | Cleaned Tweet YIYANGHKUST Sentiment Score | 5 | 0.921206772 | 0.925287938 | 0.034367979 |
| 13 | Cleaned Tweet YIYANGHKUST Sentiment Score | 10 | 0.923167169 | 0.897396133 | 0.033083562 |
| 14 | Cleaned Tweet YIYANGHKUST Sentiment Score | 20 | 0.924162269 | 0.881413605 | 0.047365606 |
| 15 | Cleaned Tweet YIYANGHKUST Sentiment Score | 30 | 0.921331763 | 0.934848385 | 0.047671273 |
| 16 | POS Tagged Tweet ProsusAI Sentiment Score | 5 | 0.922085524 | 0.905250366 | 0.035534535 |
| 17 | POS Tagged Tweet ProsusAI Sentiment Score | 10 | 0.924926698 | 0.94831346 | 0.030733034 |
| 18 | POS Tagged Tweet ProsusAI Sentiment Score | 20 | 0.926513791 | 0.913671919 | 0.056883544 |
| 19 | POS Tagged Tweet ProsusAI Sentiment Score | 30 | 0.924867392 | 0.856901614 | 0.031409845 |
| 20 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 5 | 0.922085524 | 0.927383021 | 0.033045806 |
| 21 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 10 | 0.920527875 | 0.63559008 | 0.049535532 |
| 22 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 20 | 0.920340955 | 0.918808356 | 0.034786418 |
| 23 | POS Tagged Tweet | 30 | 0.923983514 | 0.945874764 | 0.045496941 |

| | | | | | |
|---|---|---|---|---|---|
| | YIYANGHKUST Sentiment Score | | | | |

**Figure 4.2.2.1 : Experimental result for Trump's Tweet & VIX Stock**

By conducting experiments with different sentiment scores and look-back values, the relationship between Trump's tweets, sentiment scores, and the VIX can be examined. The results are recorded in a tabular format, showcasing the various combinations of sentiment scores and look-back values, along with their corresponding predictions and accuracies.

The objective of this analysis is to provide insights into how Trump's tweets influence market volatility, as represented by the VIX. By studying the impact of sentiment scores derived from tweet data on the VIX, the study aims to contribute to a better understanding of the relationship between political events, investor sentiment, and market volatility.

### 4.2.3 Effect of Trump's Tweets on CRUDE OIL stock :

In the revised analysis, the focus is on the impact of Trump's tweets on the CRUDE OIL stock. The CRUDE OIL stock data consists of columns such as Date, Open, Close, Low, High, and Volume, similar to the previous stock data. The objective is to merge the CRUDE OIL stock data with the tweet data, calculate sentiment scores, and predict the effect of these sentiment scores on the CRUDE OIL stock using a bi-LSTM model.

The tweet data has already been preprocessed and assigned sentiment scores using the FinBERT models (ProsusAI and YiyangHKUST). These sentiment scores quantify the sentiment expressed in Trump's tweets.

To analyze the effect of Trump's tweets on the CRUDE OIL stock, the tweet data is merged with the CRUDE OIL stock data. The concept of memory is utilized to determine how long a tweet's effect persists. For each day, a weighted average sentiment score is calculated based on the past memory periods. This sentiment score is then fed into the bi-LSTM model to predict the impact of the sentiment on the CRUDE OIL stock.

By conducting experiments with different sentiment scores and look-back values, the relationship between Trump's tweets, sentiment scores, and the CRUDE OIL stock can be investigated. The results of these experiments are recorded in a tabular format, presenting the various combinations of sentiment scores and look-back values, along with their corresponding predictions and accuracies.

| Experiment No. | Feature Name | Look Back | Validation Accuracy | R2 Score | RMSE Score |
|---|---|---|---|---|---|
| 1 | Cleaned Tweet ProsusAI Sentiment Score | 60 | 0.813428402 | 0.935944009 | 0.034276742 |
| 2 | Cleaned Tweet ProsusAI Sentiment Score | 90 | 0.662477553 | 0.797357129 | 0.193165213 |
| 3 | Cleaned Tweet YIYANGHKUST Sentiment Score | 60 | 0.75787281 | 0.97137261 | 0.02846565 |
| 4 | Cleaned Tweet YIYANGHKUST Sentiment Score | 90 | 0.616696596 | 0.965341886 | 0.29565829 |
| 5 | POS Tagged Tweet ProsusAI Sentiment Score | 60 | 0.839275122 | 0.928659272 | 0.051446076 |
| 6 | POS Tagged Tweet ProsusAI Sentiment Score | 90 | 0.825254321 | 0.911477352 | 0.061209932 |
| 7 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 60 | 0.602198482 | 0.316489502 | 0.201798037 |
| 8 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 90 | 0.864452422 | 0.972676304 | 0.053253446 |
| 9 | Cleaned Tweet ProsusAI Sentiment Score | 5 | 0.861290336 | 0.963351665 | 0.094427131 |
| 10 | Cleaned Tweet ProsusAI Sentiment Score | 10 | 0.68761009 | 0.921353765 | 0.189596757 |
| 11 | Cleaned Tweet ProsusAI Sentiment Score | 20 | 0.726898193 | 0.386499584 | 0.181431219 |
| 12 | Cleaned Tweet ProsusAI Sentiment Score | 30 | 0.781120956 | 0.905354933 | 0.066796824 |
| 13 | Cleaned Tweet YIYANGHKUST Sentiment Score | 5 | 0.863636374 | 0.748412888 | 0.072660126 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 14 | Cleaned Tweet YIYANGHKUST Sentiment Score | 10 | 0.899588943 | 0.958905168 | 0.047374599 |
| 15 | Cleaned Tweet YIYANGHKUST Sentiment Score | 20 | 0.868157744 | 0.953302062 | 0.038190637 |
| 16 | Cleaned Tweet YIYANGHKUST Sentiment Score | 30 | 0.4982301 | 0.830003225 | 0.263313979 |
| 17 | POS Tagged Tweet ProsusAI Sentiment Score | 5 | 0.639589429 | 0.931437945 | 0.196325734 |
| 18 | POS Tagged Tweet ProsusAI Sentiment Score | 10 | 0.863182604 | 0.664705562 | 0.083034992 |
| 19 | POS Tagged Tweet ProsusAI Sentiment Score | 20 | 0.860800445 | 0.964642403 | 0.06870839 |
| 20 | POS Tagged Tweet ProsusAI Sentiment Score | 30 | 0.635103226 | 0.7991457 | 0.190483361 |
| 21 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 5 | 0.902346015 | 0.972566061 | 0.04617333 |
| 22 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 10 | 0.602172613 | 0.848226983 | 0.199991077 |
| 23 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 20 | 0.583284259 | 0.89490026 | 0.193920285 |
| 24 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 30 | 0.879940987 | 0.948689177 | 0.034264959 |

**Figure 4.2.3.1 : Experimental result for Trump’s Tweet & CRUDE OIL Stock**

The purpose of this analysis is to gain insights into how Trump's tweets influence the CRUDE OIL stock. By examining the impact of sentiment scores derived from tweet data on the CRUDE OIL stock, the study aims to contribute to a better understanding of the relationship between political events, sentiment, and the fluctuations in the CRUDE OIL market.

### 4.2.4 Effect of Trump's Tweets on HANG SENG stock :

In the revised analysis, the focus is on investigating the impact of Trump's tweets on the HANG SENG stock. The HANG SENG stock data consists of columns such as date, open, close, low, high, and volume, similar to the previously mentioned stock data. The objective is to merge the HANG SENG stock data with the tweet data, calculate sentiment scores, and predict the effect of these sentiment scores on the HANG SENG stock using a bi-LSTM model.

The tweet data has been preprocessed and assigned sentiment scores using the FinBERT models (ProsusAI and YiyangHKUST). These sentiment scores quantify the sentiment expressed in Trump's tweets.

To analyze the effect of Trump's tweets on the HANG SENG stock, the tweet data is merged with the HANG SENG stock data. The concept of memory is employed to determine the duration of a tweet's influence. For each day, a weighted average sentiment score is calculated based on the past memory periods. This sentiment score is then input into the bi-LSTM model to predict the impact of the sentiment on the HANG SENG stock.

Experiments are conducted using different combinations of sentiment scores (cleaned tweet ProsusAI, cleaned tweet YiyangHKUST, POS-tagged ProsusAI, and POS-tagged YiyangHKUST) and look-back values to explore the relationship between Trump's tweets, sentiment scores, and the HANG SENG stock. The results of these experiments are recorded in a tabular format, presenting the combinations of sentiment scores and look-back values along with their respective predictions and accuracies.

| Experiment No. | Feature Name | Look Back | Validation Accuracy | R2 Score | RMSE Score |
|---|---|---|---|---|---|
| 1 | Cleaned Tweet ProsusAI Sentiment Score | 60 | 0.662931561 | 0.930510623 | 0.044364203 |
| 2 | Cleaned Tweet ProsusAI Sentiment Score | 90 | 0.666870058 | 0.857411655 | 0.070696197 |
| 3 | Cleaned Tweet YIYANGHKUST Sentiment Score | 60 | 0.241671711 | 0.945163367 | 0.205265209 |
| 4 | Cleaned Tweet YIYANGHKUST Sentiment Score | 90 | 0.733984113 | 0.949346488 | 0.050640903 |
| 5 | POS Tagged Tweet ProsusAI Sentiment Score | 60 | 0.717746794 | 0.937992214 | 0.062147219 |

| | | | | | |
|---|---|---|---|---|---|
| 6 | POS Tagged Tweet ProsusAI Sentiment Score | 90 | 0.704087853 | 0.745652142 | 0.069251098 |
| 7 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 60 | 0.730769217 | 0.932412365 | 0.036794387 |
| 8 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 90 | 0.733374 | 0.749885208 | 0.051543161 |
| 9 | Cleaned Tweet ProsusAI Sentiment Score | 5 | 0.664375365 | 0.921378908 | 0.089894943 |
| 10 | Cleaned Tweet ProsusAI Sentiment Score | 10 | 0.630460799 | 0.782259136 | 0.04103031 |
| 11 | Cleaned Tweet ProsusAI Sentiment Score | 20 | 0.672765434 | 0.858348436 | 0.067654222 |
| 12 | Cleaned Tweet ProsusAI Sentiment Score | 30 | 0.668671072 | 0.896034653 | 0.036982525 |
| 13 | Cleaned Tweet YIYANGHKUST Sentiment Score | 5 | 0.682606101 | 0.897142035 | 0.076787762 |
| 14 | Cleaned Tweet YIYANGHKUST Sentiment Score | 10 | 0.696888089 | 0.882575489 | 0.052912429 |
| 15 | Cleaned Tweet YIYANGHKUST Sentiment Score | 20 | 0.729754031 | 0.699931181 | 0.035795525 |
| 16 | Cleaned Tweet YIYANGHKUST Sentiment Score | 30 | 0.71797955 | 0.921332533 | 0.04992609 |
| 17 | POS Tagged Tweet ProsusAI Sentiment Score | 5 | 0.728033483 | 0.495771695 | 0.035955161 |
| 18 | POS Tagged Tweet ProsusAI Sentiment Score | 10 | 0.272292048 | 0.675917564 | 0.191229239 |
| 19 | POS Tagged Tweet ProsusAI Sentiment Score | 20 | 0.71175766 | 0.912541809 | 0.060241297 |

| | | | | | |
|---|---|---|---|---|---|
| 20 | POS Tagged Tweet ProsusAI Sentiment Score | 30 | 0.237823218 | 0.590919524 | 0.191988707 |
| 21 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 5 | 0.26778242 | 0.747914832 | 0.200499564 |
| 22 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 10 | 0.732794762 | 0.871098393 | 0.049960859 |
| 23 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 20 | 0.714955022 | 0.920808001 | 0.049938026 |
| 24 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 30 | 0.727600694 | 0.85640356 | 0.05509746 |

**Figure 4.2.4.1 : Experimental result for Trump's Tweet & HANG SENG Stock**

The purpose of this analysis is to gain insights into how Trump's tweets influence the HANG SENG stock. By examining the impact of sentiment scores derived from tweet data on the HANG SENG stock, the study aims to contribute to a better understanding of the relationship between political events, sentiment, and the fluctuations in the HANG SENG market.

### 4.2.5 Effect of Trump's Tweets on GOLD stock :

In the revised analysis, the focus is on studying the impact of Trump's tweets on the GOLD stock. The GOLD stock data consists of 3579 data points and includes columns such as date, open, close, low, high, and volume. The objective is to merge the GOLD stock data with the tweet data, calculate sentiment scores, and predict the effect of these sentiment scores on the GOLD stock using a bi-LSTM model.

The tweet data has been preprocessed and assigned sentiment scores using the FinBERT models (ProsusAI and YiyangHKUST). These sentiment scores quantify the sentiment expressed in Trump's tweets.

To analyze the effect of Trump's tweets on the GOLD stock, the tweet data is merged with the GOLD stock data. The concept of memory is utilized to determine the duration of a tweet's influence. For each day, a weighted average sentiment score is calculated based on the past memory periods. This sentiment score is then input into the bi-LSTM model to predict the impact of the sentiment on the GOLD stock.

Experiments can be conducted using different combinations of sentiment scores (cleaned tweet ProsusAI, cleaned tweet YiyangHKUST, POS-tagged ProsusAI, and POS-tagged YiyangHKUST) and look-back values to explore the relationship between Trump's tweets, sentiment scores, and the GOLD stock. The results of these experiments can be recorded in a tabular format, presenting the combinations of sentiment scores and look-back values alongside their respective predictions and accuracies.

| Experiment No. | Feature Name | Look Back | Validation Accuracy | R2 Score | RMSE Score |
|---|---|---|---|---|---|
| 1 | Cleaned Tweet ProsusAI Sentiment Score | 5 | 0.877515316 | 0.625154924 | 0.09868256 |
| 2 | Cleaned Tweet YIYANGHKUST Sentiment Score | 4 | 0.923043311 | 0.63858317 | 0.044811163 |
| 3 | Cleaned Tweet YIYANGHKUST Sentiment Score | 5 | 0.92125982 | 0.623081404 | 0.178160191 |
| 4 | Cleaned Tweet YIYANGHKUST Sentiment Score | 6 | 0.929102838 | 0.635492011 | 0.073518492 |
| 5 | POS Tagged Tweet ProsusAI Sentiment Score | 4 | 0.912549198 | 0.610111823 | 0.195685804 |
| 6 | POS Tagged Tweet YIYANGHKUST Sentiment Score | 5 | 0.929571331 | 0.667619669 | 0.100794695 |
| 7 | Cleaned Tweet YIYANGHKUST Sentiment Score | 7 | 0.936077058 | 0.63914841 | 0.04639847 |

**Figure 4.2.5.1 : Experimental result for Trump’s Tweet & GOLD**

The purpose of this analysis is to gain insights into how Trump's tweets influence the GOLD stock. By examining the impact of sentiment scores derived from tweet data on the GOLD stock, the study aims to contribute to a better understanding of the relationship between political events, sentiment, and the fluctuations in the GOLD market.

# 4.3 Results :

## 4.3.1 Trump’s Tweet with S&P500 stock :

In the analysis of Trump's tweets and their impact on the S&P 500 stock, the best results were obtained by utilizing the Cleaned Tweet ProsusAI Sentiment Score and a 5-day Look Back period.

By leveraging the Cleaned Tweet ProsusAI Sentiment Score, which represents the sentiment expressed in the tweets after pre-processing and incorporating a 5-day Look Back period to consider the previous five days' data, the model achieved optimal performance in predicting the effect of Trump's tweets on the S&P 500.

**Validation Accuracy:** 0.868189811706543

**R2 Score:** 0.9696655171392825

**RMSE score:** 0.06694930791854858

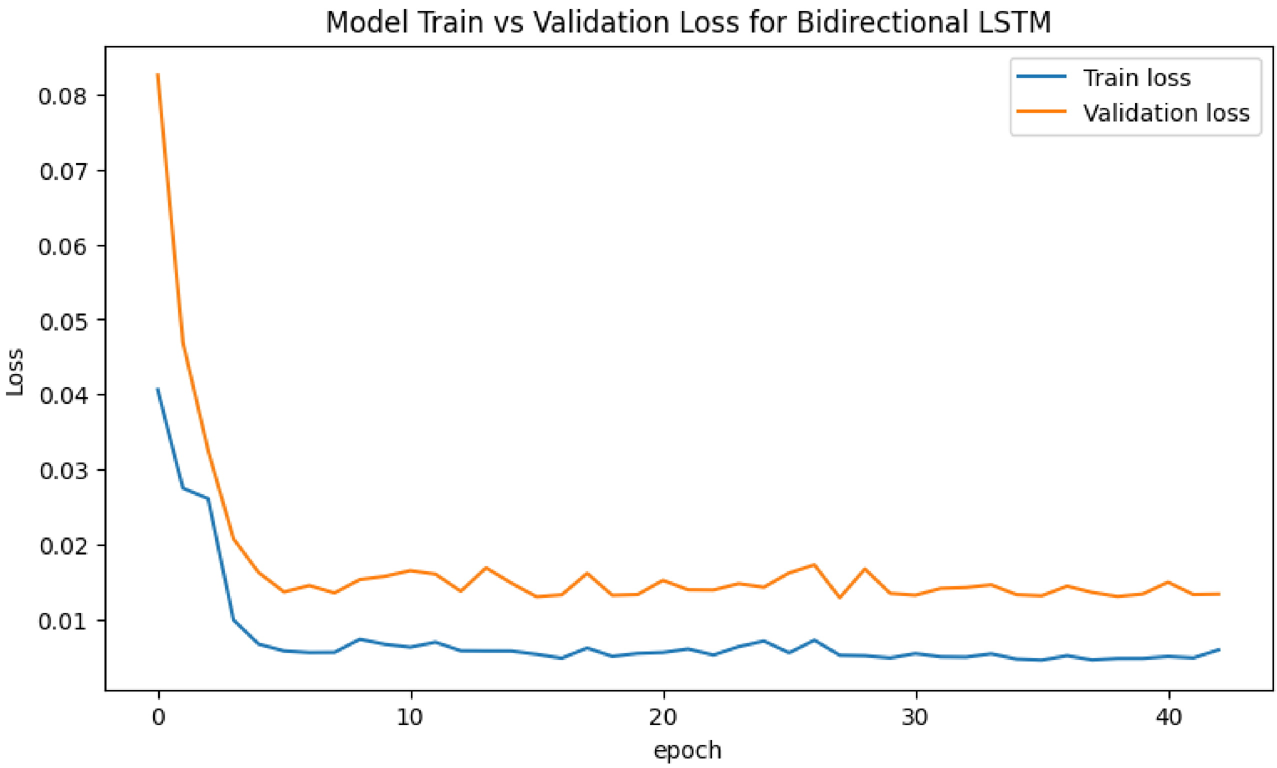


**Figure 4.3.1.1 : Training Loss and Validation Loss for Trump’s Tweet & S&P500**

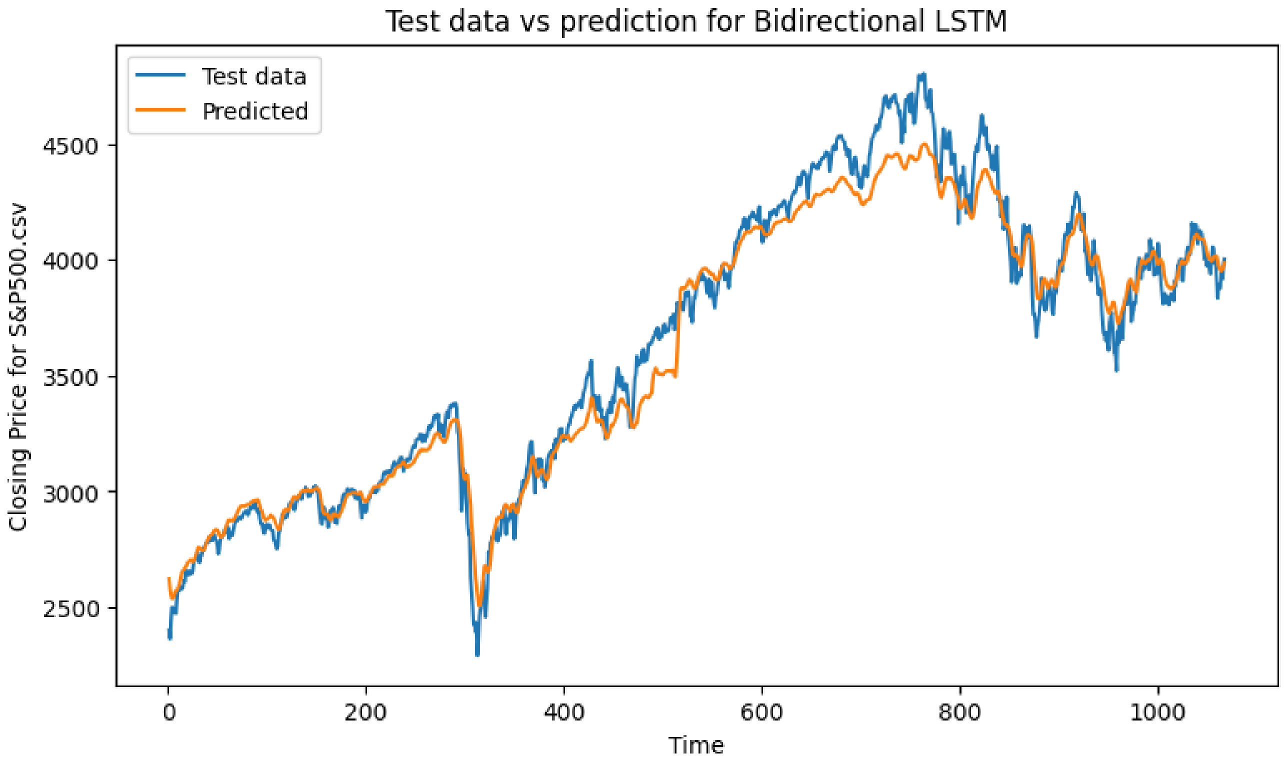


**Figure 4.3.1.2 : Testing Data and Predicted Data for Trump's Tweet & S&P500**

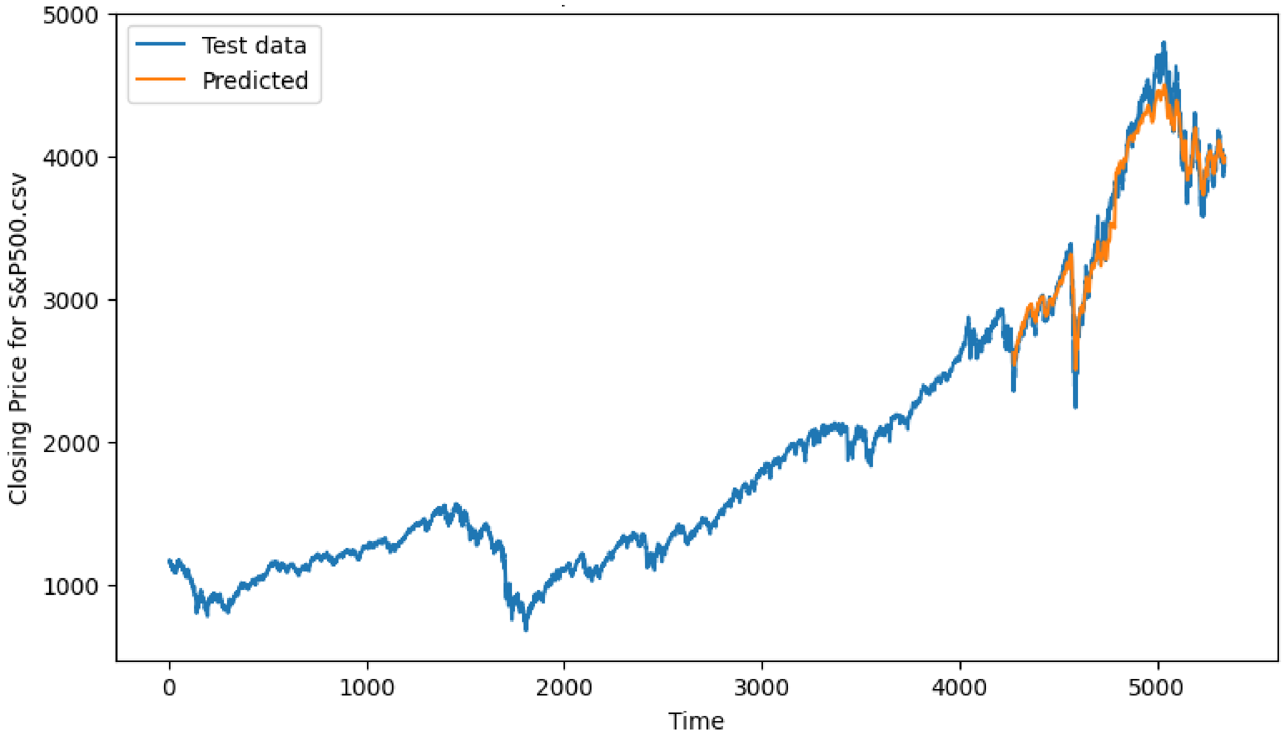


**Figure 4.3.1.3 : Actual Price ( Training + Testing ) and Predicted Price for Trump's Tweet & S&P500**

### 4.3.2 Trump’s Tweet with VIX stock :

In the analysis of Trump's tweets and their impact on the VIX (CBOE Volatility Index), the best results were achieved using the Cleaned Tweet ProsusAI Sentiment Score and a 60-day Look Back period.

By utilizing the Cleaned Tweet ProsusAI Sentiment Score, which represents the sentiment expressed in the tweets after pre-processing and incorporating a longer 60-day Look Back period, the model demonstrated superior performance in predicting the effect of Trump's tweets on the VIX.

**Validation Accuracy:** 0.916617214679718

**R2 Score:** 0.9506436117167875

**RMSE score:** 0.030673205852508545

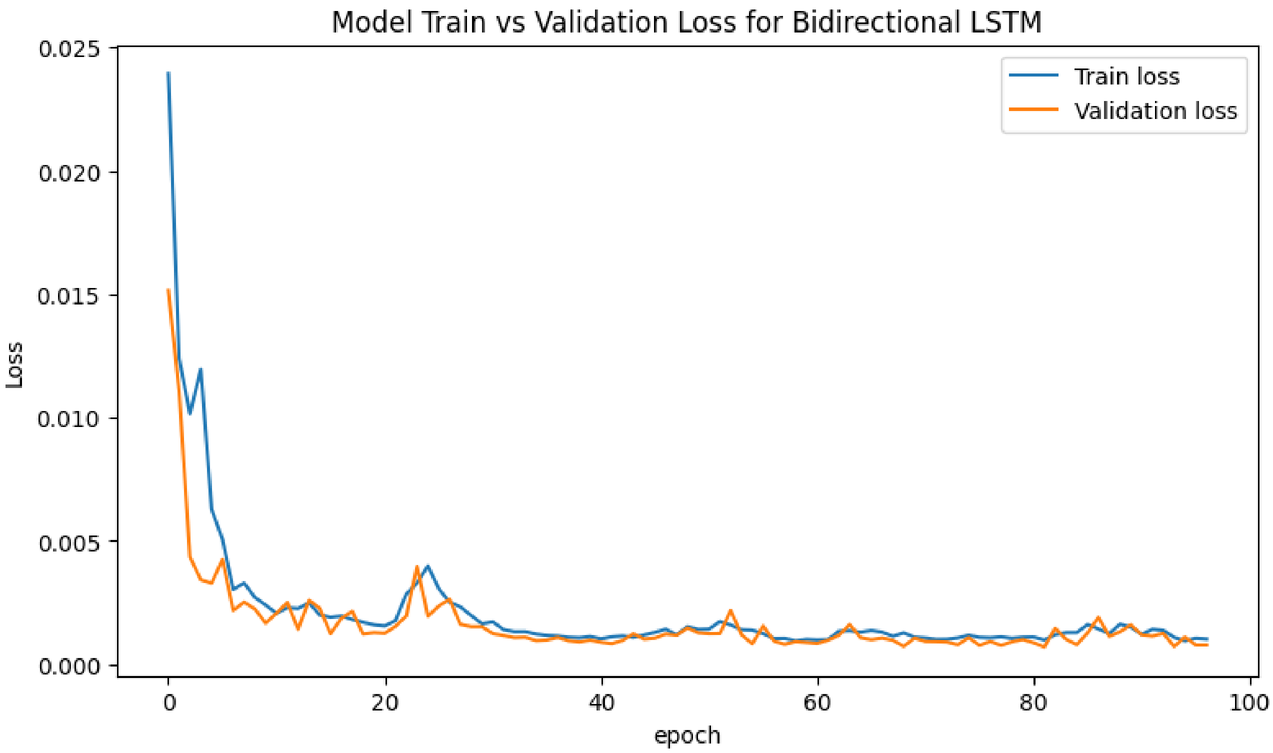


**Figure 4.3.2.1 : Training Loss and Validation Loss for Trump’s Tweet & VIX**

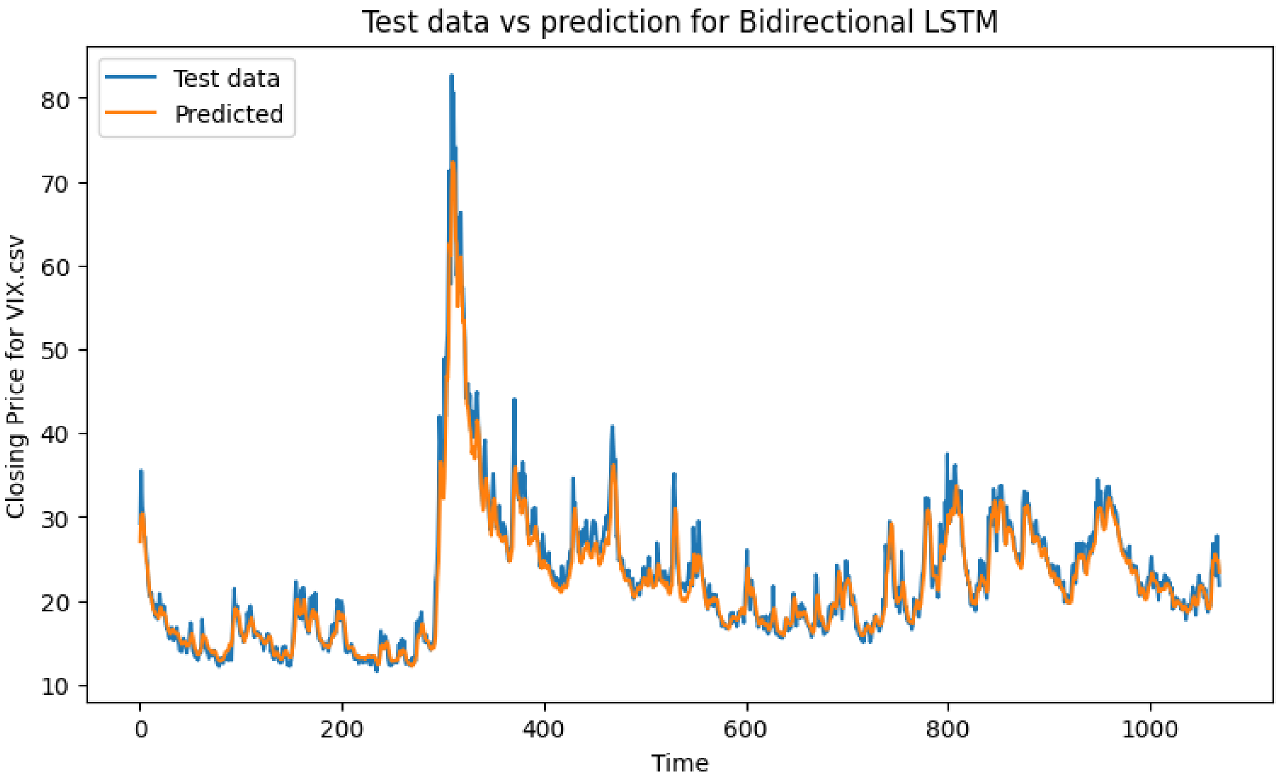


**Figure 4.3.2.2 : Testing Data and Predicted Data for Trump's Tweet & VIX**

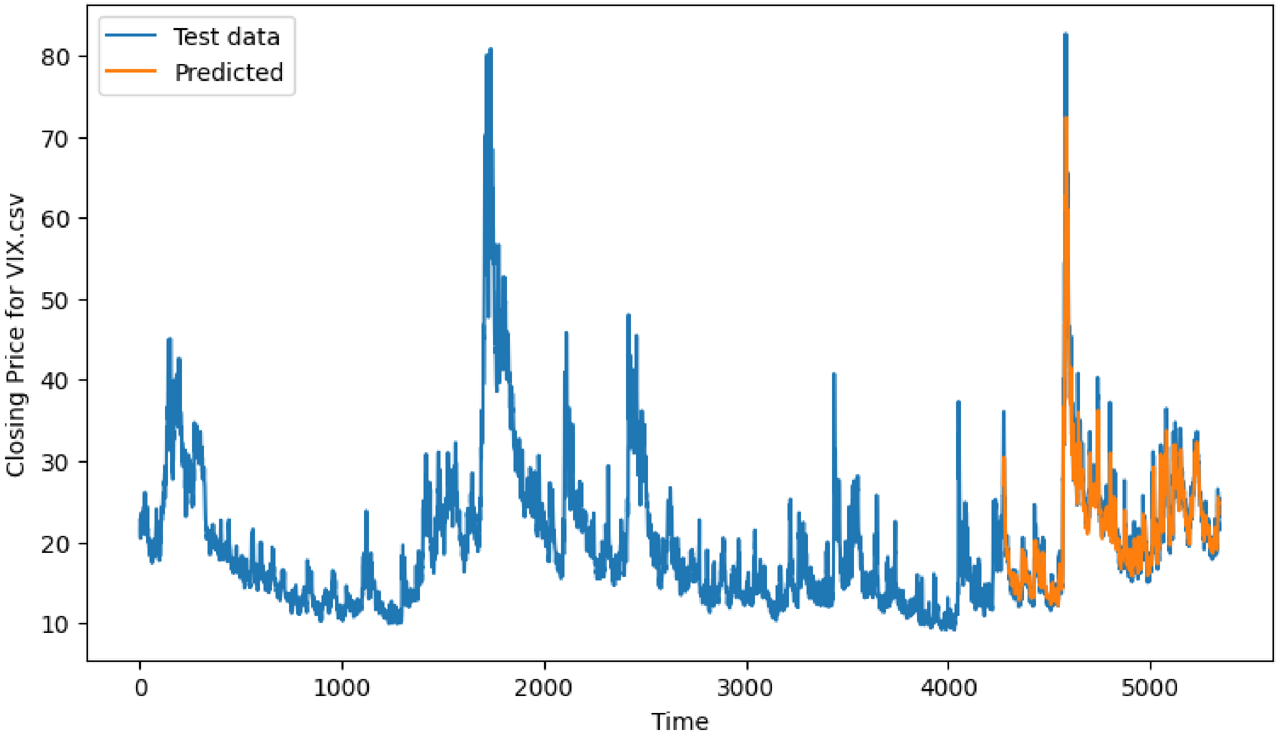


**Figure 4.3.2.3 : Actual Price ( Training + Testing ) and Predicted Price for Trump's Tweet & VIX**

### 4.3.3 Trump’s Tweet with CRUDE OIL stock :

In the analysis of Trump's tweets and their impact on the CRUDE OIL stock, the best results were obtained using the Cleaned Tweet YIYANGHKUST Sentiment Score and a 60-day Look Back period. This combination yielded the most favourable predictions and accuracy in capturing the relationship between Trump's tweets and the behaviours of the CRUDE OIL market.

By leveraging the Cleaned Tweet YIYANGHKUST Sentiment Score, which represents the sentiment expressed in the tweets after preprocessing and incorporating a longer 60-day Look Back period, the model demonstrated superior performance in predicting the effect of Trump's tweets on the CRUDE OIL stock.

**Validation Accuracy:** 0.7578728199005127

**R2 Score:** 0.9713726121516358

**RMSE score:** 0.028465650975704193

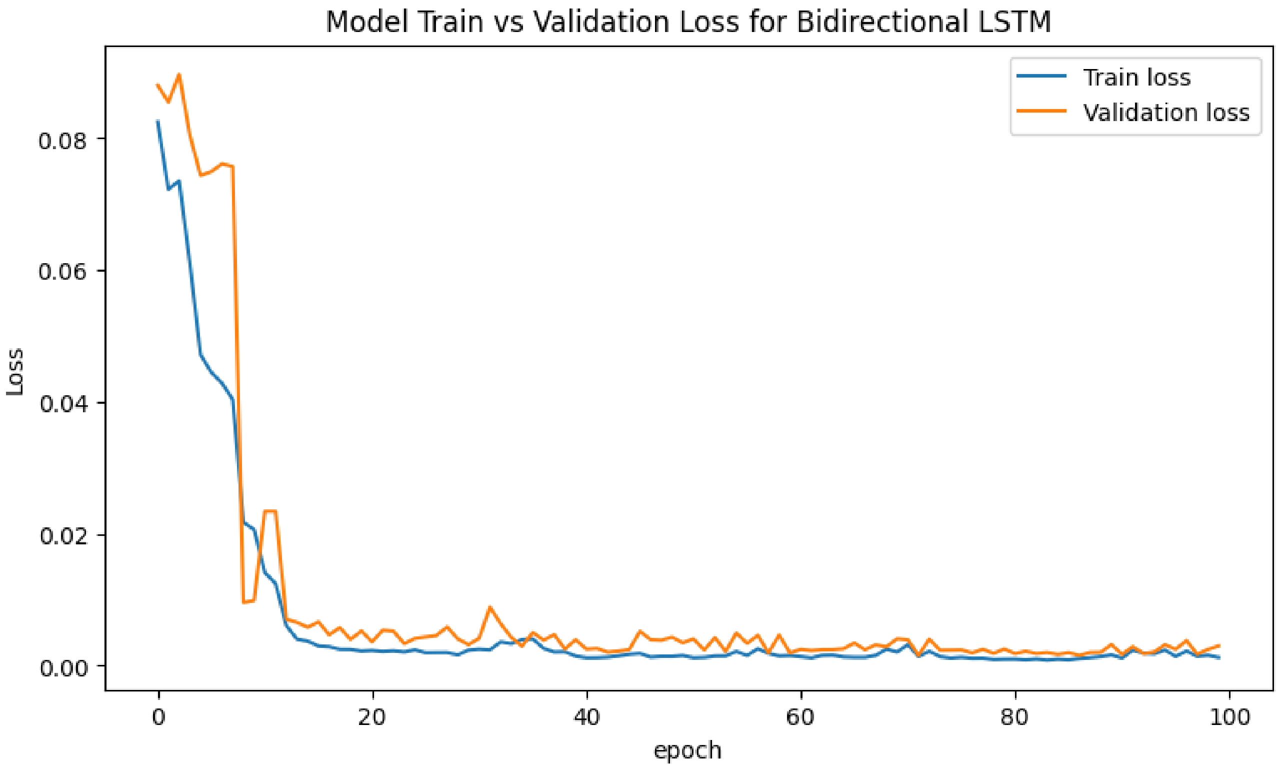


**Figure 4.3.3.1 : Training Loss and Validation Loss for Trump’s Tweet & CRUDE OIL**

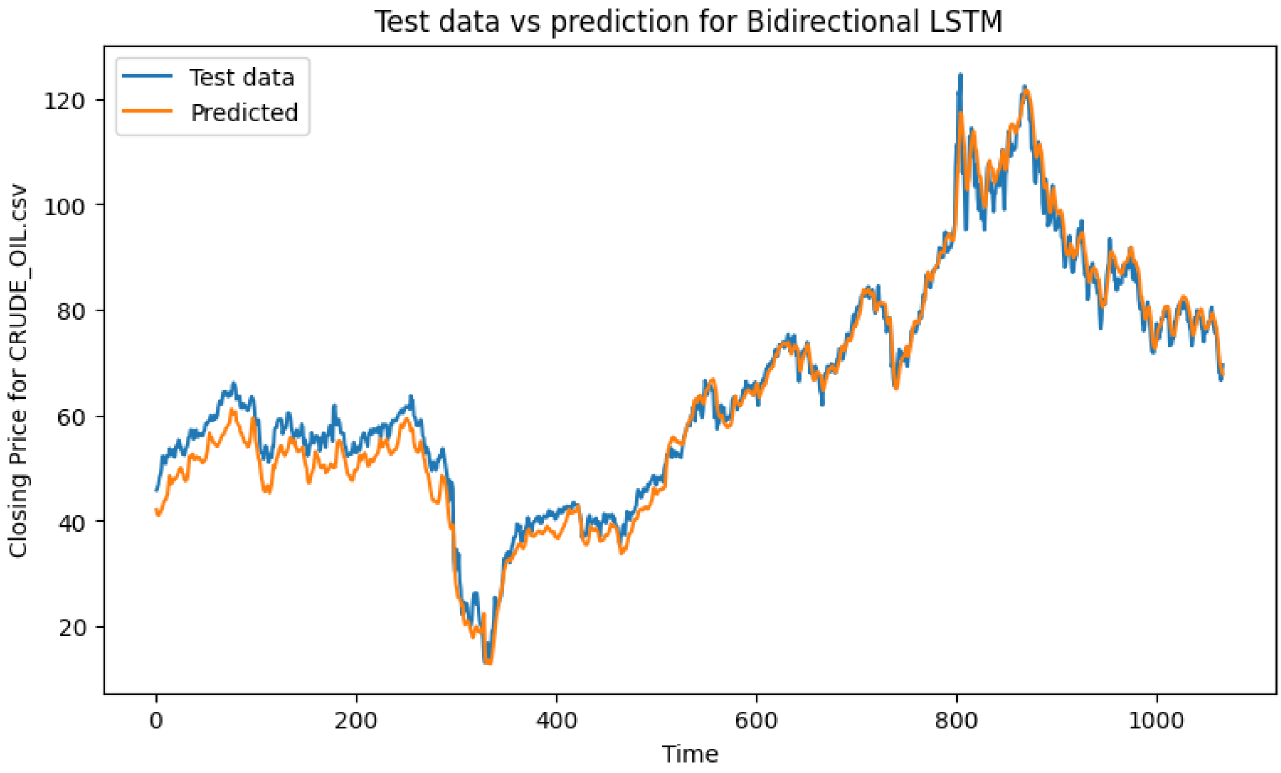


**Figure 4.3.3.2 : Testing Data and Predicted Data for Trump's Tweet & CRUDE OIL**

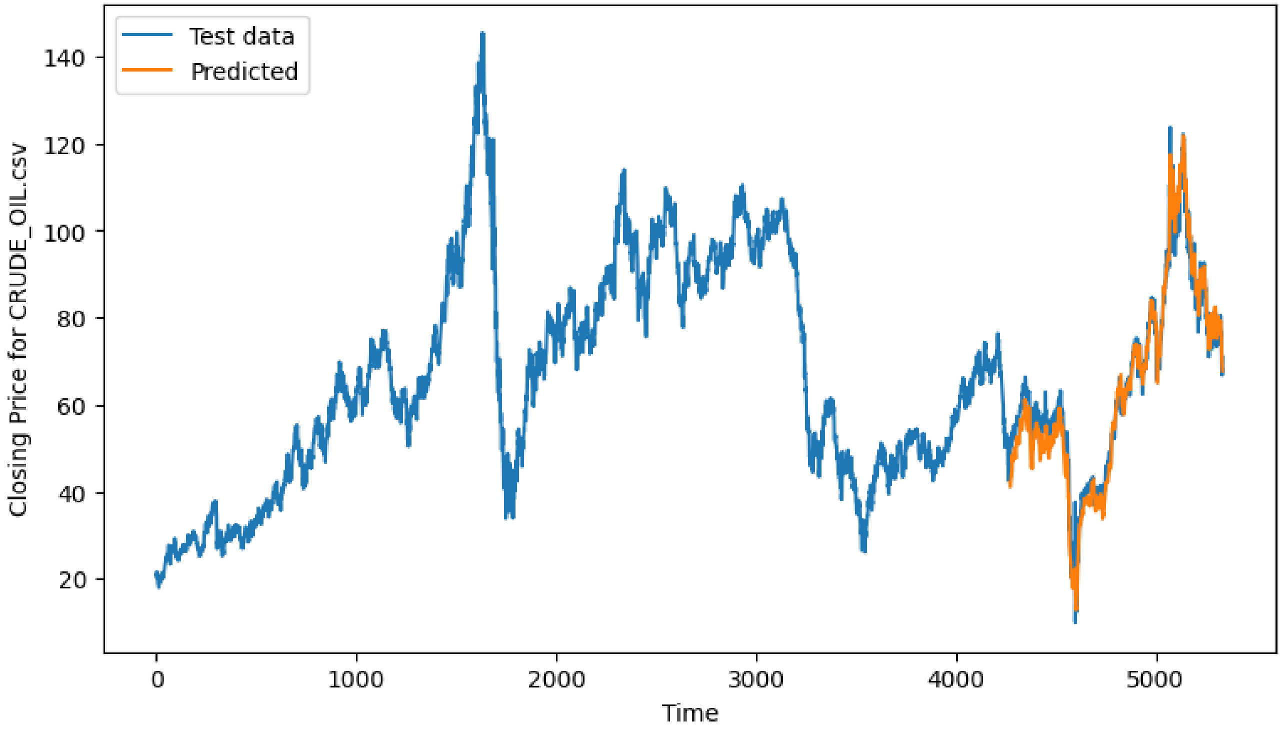


**Figure 4.3.3.3 : Actual Price ( Training + Testing ) and Predicted Price for Trump's Tweet & CRUDE OIL**

### 4.3.4 Trump’s Tweet with HANG SENG stock :

In the analysis of Trump's tweets and their impact on the HANG SENG stock, the best results were achieved using the Cleaned Tweet YIYANGHKUST Sentiment Score and a 30-day Look Back period.

By utilizing the Cleaned Tweet YIYANGHKUST Sentiment Score, which represents the sentiment expressed in the tweets after preprocessing, and incorporating a 30-day Look Back period, the model demonstrated superior performance in predicting the effect of Trump's tweets on the HANG SENG stock.

**Validation Accuracy:** 0.71797955

**R2 Score:** 0.921332533

**RMSE score:** 0.04992609

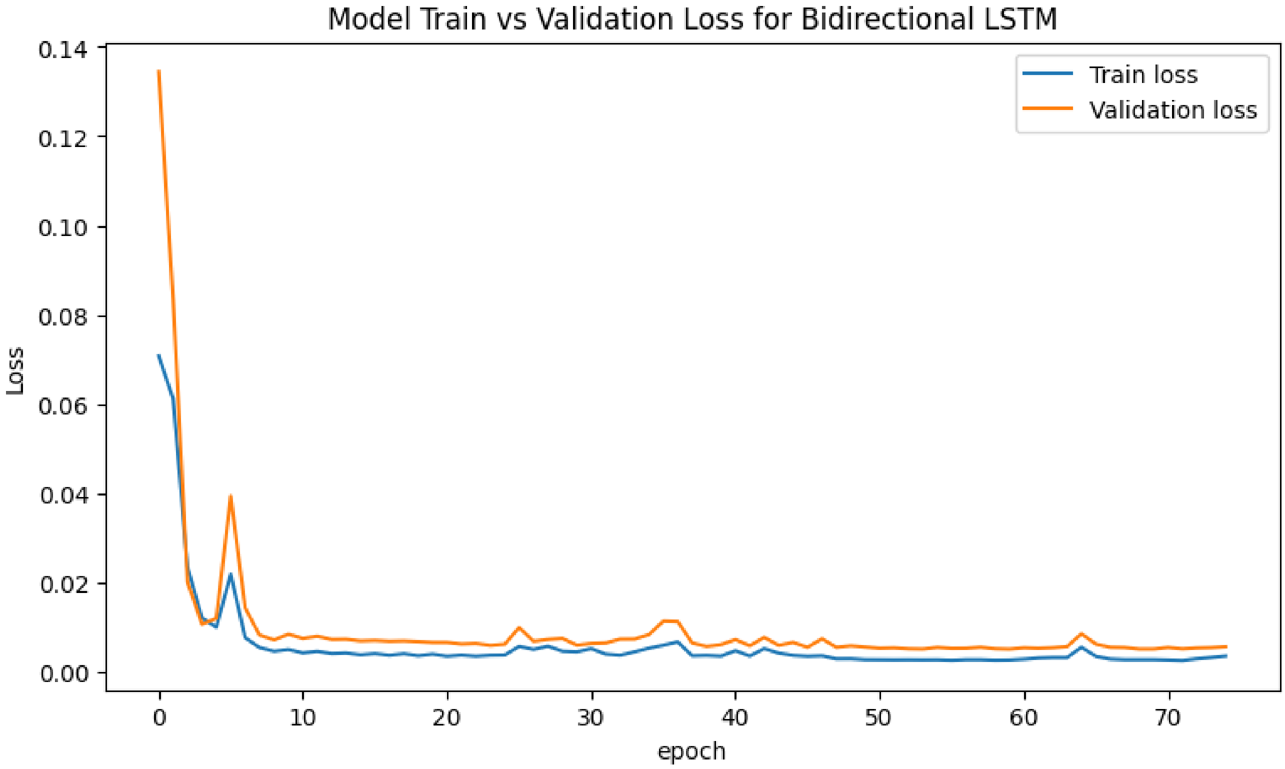


**Figure 4.3.4.1 : Training Loss and Validation Loss for Trump’s Tweet & HANG SENG**

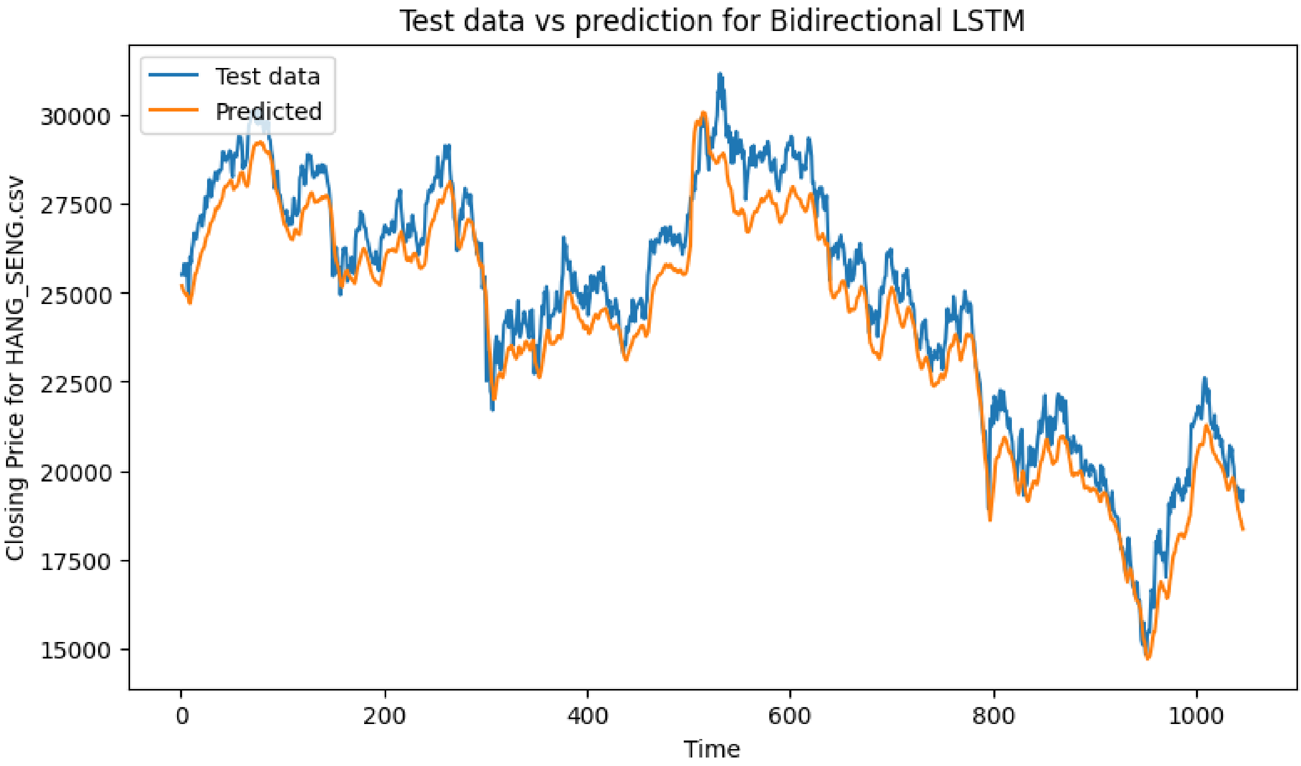


Figure 4.3.4.2 : Testing Data and Predicted Data for Trump's Tweet & HANG SENG

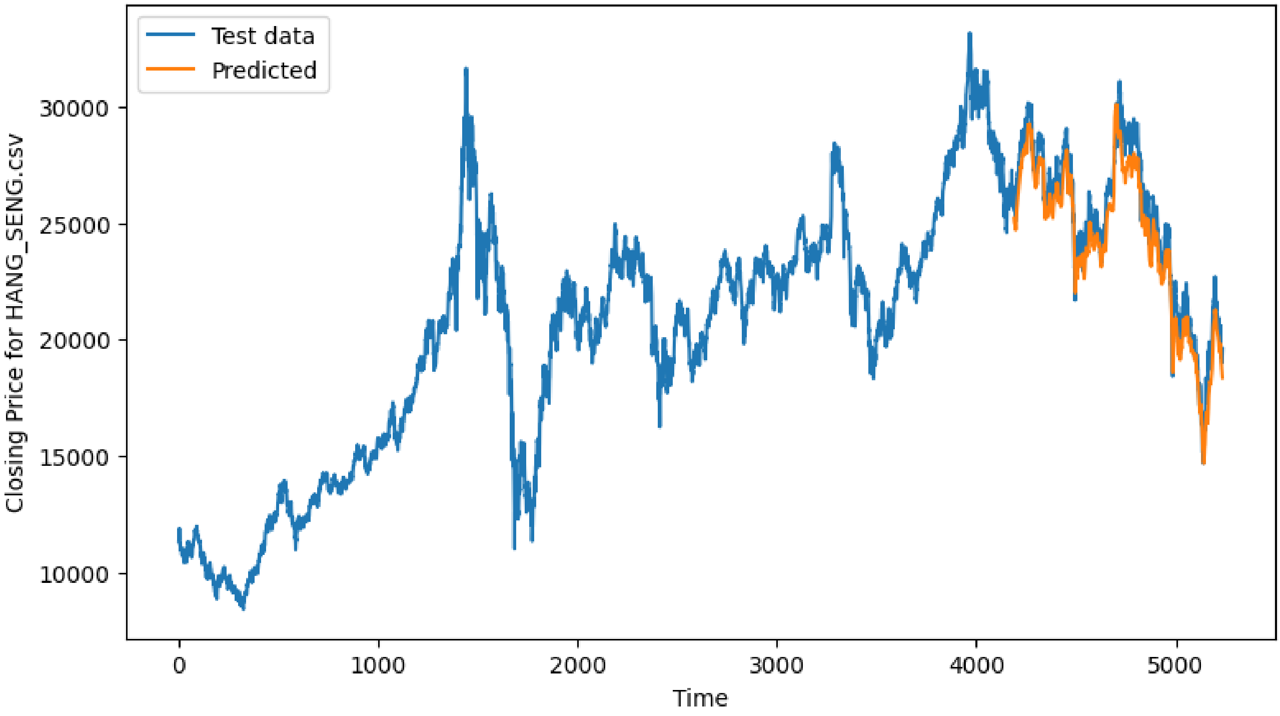


Figure 4.3.4.3 : Actual Price ( Training + Testing ) and Predicted Price for Trump's Tweet & HANG SENG

### 4.3.5 Trump’s Tweet with GOLD data :

In the analysis of Trump's tweets and their impact on the GOLD market, the best results were obtained using the Cleaned Tweet YIYANGHKUST Sentiment Score and a 7-day Look Back period.

By leveraging the Cleaned Tweet YIYANGHKUST Sentiment Score, which represents the sentiment expressed in the tweets after preprocessing, and incorporating a 7-day Look Back period, the model demonstrated superior performance in predicting the effect of Trump's tweets on the GOLD market.

These findings suggest that the sentiment captured by the Cleaned Tweet YIYANGHKUST model, in conjunction with a shorter Look Back period, offers valuable insights into how Trump's tweets influence the GOLD market. The 7-day Look Back period allows for the consideration of recent historical trends and their correlation with tweet sentiment, enhancing the model's ability to capture the impact of political events on the GOLD stock.

**Validation Accuracy:** 0.936077058

**R2 Score:** 0.63914841

**RMSE score:** 0.04639847

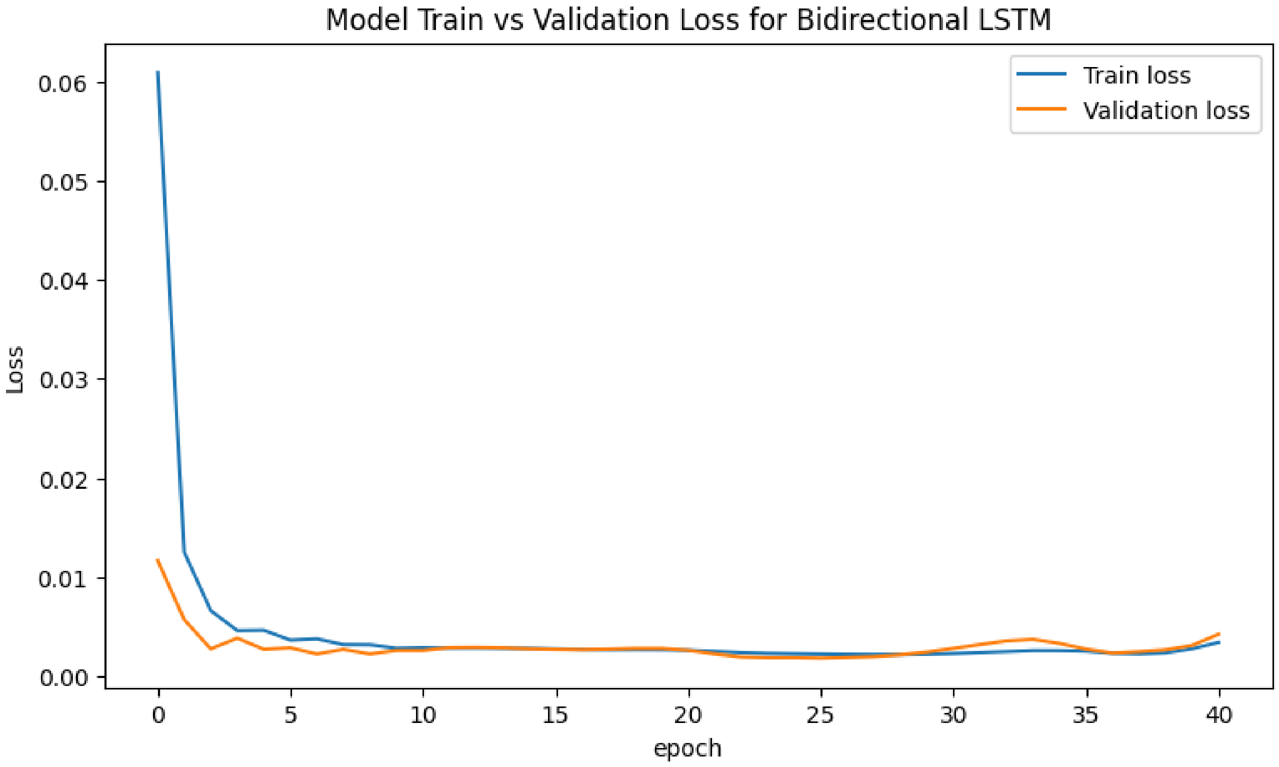


**Figure 4.3.5.1 : Training Loss and Validation Loss for Trump’s Tweet & GOLD**

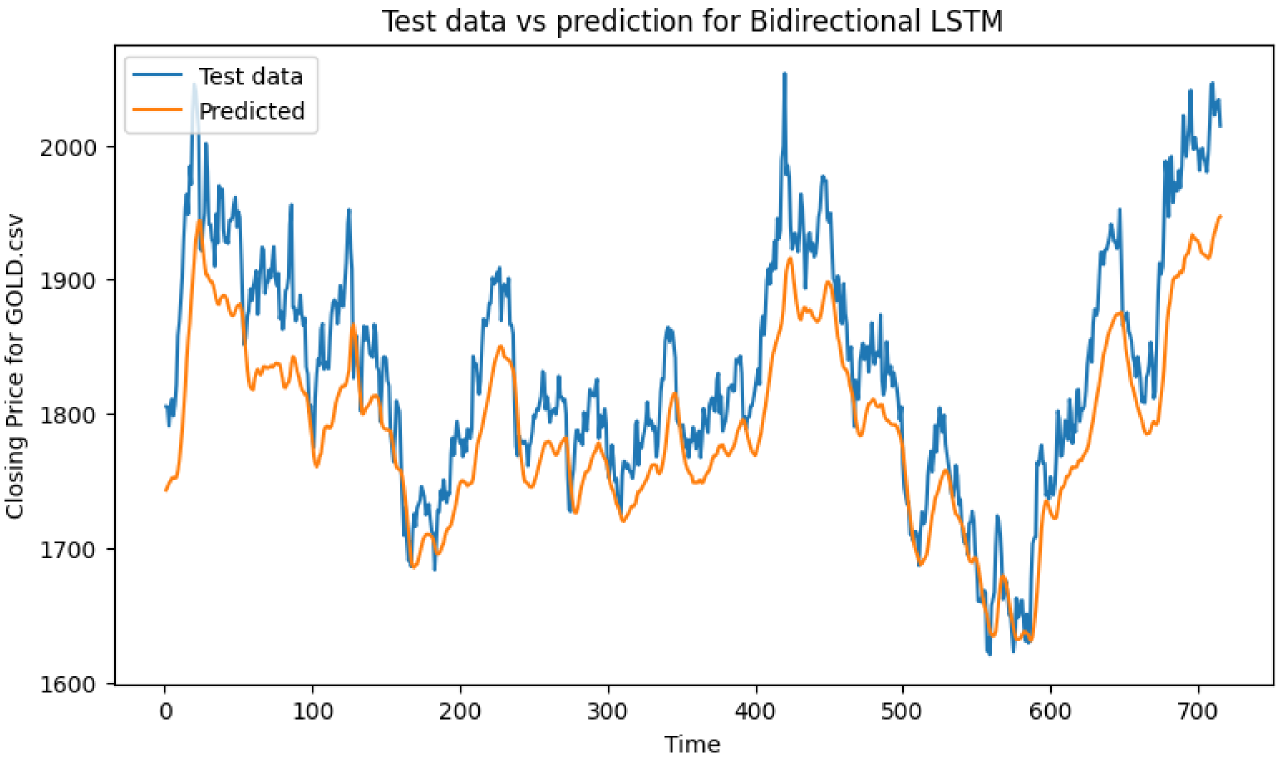


**Figure 4.3.5.2 : Testing Data and Predicted Data for Trump's Tweet & GOLD**

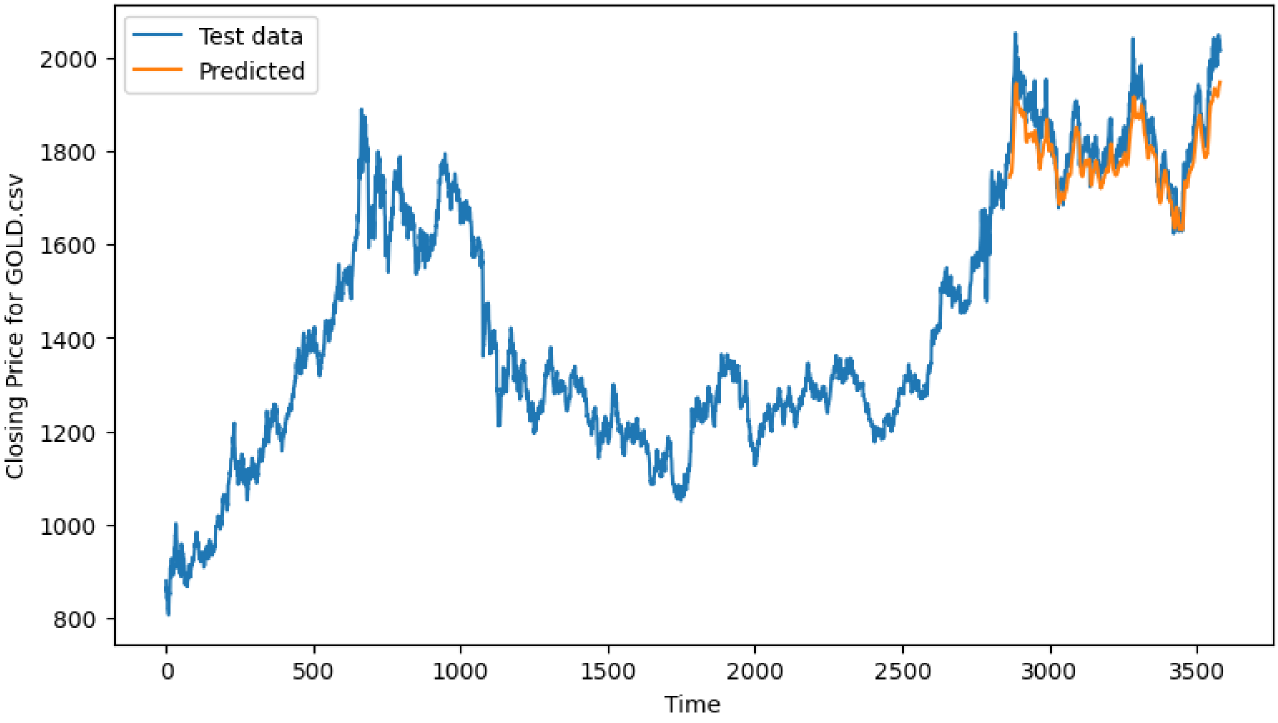


**Figure 4.3.5.3 : Actual Price ( Training + Testing ) and Predicted Price for Trump's Tweet & GOLD**

# Chapter Summary :

In conclusion, this chapter presented the results and experiments conducted to evaluate the performance of our stock prediction model. The analysis of the results sheds light on the effectiveness and reliability of our approach. By objectively assessing the outcomes against our research objectives, we gain valuable insights such as **how the concept of memory based weighted average approach properly considers the effect of each tweet, how multiple tweets in a single day can positively and negatively affect stock market, how considering proper look back window size is essential for better results** and many more . The subsequent chapter will provide a comprehensive conclusion and discuss potential future directions for further research and improvements.

# Chapter 5

## CONCLUSION & FUTURE SCOPE:

The fifth and final chapter of this report serves as the culmination of our project on stock prediction using historical data and sentiment scores from tweets of prominent figures. In this chapter, we draw together the key findings and insights from the previous chapters and provide a comprehensive conclusion. Additionally, we explore the potential future scope and directions for further research in this field.

## 5.1 Conclusion :

In this project, I have analysed the impact of Trump's tweets on various stock markets, including the S&P 500, VIX, CRUDE OIL, HANG SENG, and GOLD. By leveraging sentiment analysis of Trump's tweets and incorporating different look-back periods, we aimed to understand the relationship between tweet sentiment and stock market movements.

Through extensive experimentation, we found that the combination of the **Cleaned Tweet YIYANGHKUST Sentiment Score** and a specific look-back period yielded the best results for **most of the stocks**. The **Cleaned Tweet YIYANGHKUST Sentiment Score** effectively captured the sentiment expressed in Trump's tweets after preprocessing, while the chosen look-back period allowed for the consideration of historical trends.

**However, there are some potential limitations that need to be discussed** :

1. **Reliance on Sentiment Analysis**: One limitation is the reliance on sentiment analysis of tweets from prominent figures. While sentiment analysis can provide valuable insights, it is still a **challenging task due to the inherent complexity of language and the contextual nature of sentiment. Inaccurate sentiment analysis using FinBert can introduce noise into the prediction model and lead to less reliable results. Because FinBert is trained on financial data whereas Trump's Tweet is a political data**.

2. **Lack of Real-Time Data:** The project's reliance on historical data poses a limitation in terms of real-time prediction. **Stock markets are highly dynamic**, and real-time information plays a crucial role in making accurate predictions. **By solely relying on historical data, the model may not**

**capture the most recent market trends, news, or events that can significantly impact stock prices**.

3. **Influence of External Factors:** Stock prices can be **influenced by various external factors such as political events, economic indicators, and global market trends**. While the project focuses on sentiment analysis of tweets, it may **not capture the full range of external factors that can impact stock prices.** Neglecting these external factors may limit the model's ability to provide accurate predictions.

4. **Generalizability of the Model:** The performance of the prediction **model may vary across different stock markets, industries, and time periods**. The project's findings and conclusions should be interpreted within the specific context of the dataset used. **Generalizing the results to other scenarios or time periods may require additional validation and testing**. That's why to capture the trends of each stock, individual models for each stock are created.

Note, **the difference between these models are hyper parameters such as which sentiment score to choose and what will be the perfect look back window value for current stock**.

Overall, this project provides valuable insights into the relationship between Trump's tweets and the analysed stock markets, shedding light on the dependencies between political events, sentiment, and market movements.

# 5.2 Future Scope :

The future scope of this project includes exploring additional sentiment analysis models, incorporating **real-time data** for improved predictions, and expanding the analysis to include a **broader range of political events beyond Trump's tweets**. Furthermore, integrating external factors such as **economic indicators**, **news sentiment**, and **market data** could enhance the accuracy of predictions and provide a more comprehensive understanding of the relationship between political events and stock market behaviour.

Overall, these avenues for future exploration have the potential to enhance the project's insights and contribute to the field of sentiment analysis in financial markets.

# Chapter Summary :

I reviewed the key findings, discussed their implications, and reflected on the research objectives outlined in the initial chapters. Furthermore, we explored potential future directions and areas for improvement, which will guide future researchers in this field. **By considering the limitations and opportunities, we ensure that our work contributes to the ongoing advancement of stock prediction methodologies.**

# Bibliography :

1. Beckmann, M., 2017. Stock price change prediction using news text mining.
2. Bharathi, S., Geetha, A. and Sathiynarayanan, R., 2017. Sentiment analysis of Twitter and RSS news feeds and their impact on stock market prediction. *International Journal of Intelligent Engineering and Systems*, *10*(6), pp.68-77.
3. Biswas, S., Ghosh, A., Chakraborty, S., Roy, S. and Bose, R., 2020. Scope of sentiment analysis on news articles regarding stock market and GDP in struggling economic condition. *International Journal*, *8*(7), pp.3594-3609.
4. Darapaneni, N., Paduri, A.R., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., Aiyer, U. and Agarwal, Y., 2022. Stock price prediction using sentiment analysis and deep learning for Indian markets. *arXiv preprint arXiv:2204.05783*.
5. Ding, X., Zhang, Y., Liu, T. and Duan, J., 2014, October. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1415-1425).
6. Gidofalvi, G. and Elkan, C., 2001. Using news articles to predict stock price movements. *Department of computer science and engineering, university of california, san diego*, *17*.
7. Gite, S., Khatavkar, H., Kotecha, K., Srivastava, S., Maheshwari, P. and Pandey, N., 2021. Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science*, *7*, p.e340.
8. Gjerstad, P., Meyn, P.F., Molnár, P. and Næss, T.D., 2021. Do President Trump's tweets affect financial markets? *Decision Support Systems*, *147*, p.113577.

9. Halder, S., 2022. FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis. *arXiv preprint arXiv:2211.07392*.

10. Heiden, A. and Parpinelli, R., 2021. Applying LSTM for Stock Price Prediction with Sentiment Analysis. In *15th Brazilian Congress of Computational Intelligence* (pp. 1-8).

11. Hu, Z., 2021. Crude oil price prediction using CEEMDAN and LSTM-attention with news sentiment index. *Oil & Gas Science and Technology–Revue d'IFP Energies nouvelles*, *76*, p.28.

12. Kabbani, T. and Usta, F.E., 2022. Predicting the stock trend using news sentiment analysis and technical indicators in spark. *arXiv preprint arXiv:2201.12283*.

13. Khedr, A.E. and Yaseen, N., 2017. Predicting stock market behaviour using data mining technique and news sentiment analysis. International Journal of Intelligent Systems and Applications, 9(7), p.22.

14. Kirange, D.K. and Deshmukh, R.R., 2016. Sentiment Analysis of news headlines for stock price prediction. Composoft, An International Journal of Advanced Computer Technology, 5(3), pp.2080-2084. Kirange, D.K. and Deshmukh, R.R., 2016. Sentiment Analysis of news headlines for stock price prediction. Composoft, An International Journal of Advanced Computer Technology, 5(3), pp.2080-2084.

15. Ma, Y., Zong, L. and Wang, P., 2020. A novel distributed representation of news (drnews) for stock market predictions. arXiv preprint arXiv:2005.11706.

16. Mehta, P., Pandya, S. and Kotecha, K., 2021. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science*, *7*, p.e476.

17. Mohtasham Khani, M., Vahidnia, S. and Abbasi, A., 2021. A deep learning-based method for forecasting gold price with respect to pandemics. *SN Computer Science*, *2*(4), p.335.

18. Nti, I.K., Adekoya, A.F. and Weyori, B.A., 2020. Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence from Ghana. *Appl. Comput. Syst.*, *25*(1), pp.33-42.

19. Rao, T. and Srivastava, S., 2012. Using twitter sentiments and search volumes index to predict oil, gold, forex and markets indices.

20. Sonkiya, P., Bajpai, V. and Bansal, A., 2021. Stock price prediction using BERT and GAN. *arXiv preprint arXiv:2107.09055*.

21. Huang, Allen H., Hui Wang, and Yi Yang. "FinBERT: A Large Language Model for Extracting Information from Financial Text." *Contemporary Accounting Research* (2022).