University of Halabja

College of Science

Computer Department

Wall Street Bets Stock Market Classification

Supervisor:

Brwa R. Hassan

Computer Science MSc.

Authors:

Lanya Sahdy Hama aziz

Mohamad Ahmad Arrf

Bashdar Rzgar Fatih

Halabja, 2023

Declaration

We hereby declare that this dissertation/thesis entitled: “Wall Street Bets Stock Prediction Clustering in Stock Market “is my own original work and hereby certify that unless stated, all work contained within this is my own independent research and has not been submitted for the award of any other degree at any institution, except where due acknowledgement is made in the text.

Signature: Bashdar
Name: Bashdar Rzgar

Signature: Lanya
Name: Lanya Sahdy

Signature: Mohammad
Name: Mohammad Ahmad

Abstract

One of the most popular Social Media Platforms for stocks and financial discussion is subreddit called r/Wallstreetbets. In this research We would like to help stock trading by making visualizing the stock market for them easier or making insightful stock analysis easier. Most people already know about the news outlets of the stock market, we choose popular subreddit r/Wallstreetbets as the source of news outlets and source of datasets, in this dataset we most focus on this post known as DD (Market Research), Market research in the stock market refers to the process of gathering, analyzing, and interpreting information about a particular market, industry, or company in order to inform investment decisions. Market research can involve a variety of different activities, including studying financial statements, analyzing market trends and data, monitoring industry news and developments, and evaluating the competitive landscape. After that We will use this posts (DD) on r/WallStreetBets as our dataset and our target is to classify them as either good or less good posts in order to quickly filter out posts good or bad posts that may be trader want to read it about this stock that want invest on. In the end we use regression to compare our classification with the real price market. This project targets people who are interested in trading in the stock market or people that want to invest in the stock market, but have not had enough time to find good stock or do their own research. At the end of this project users can know which stock is more popular and has a higher chance to grow.

Contents

Abbreviations

DD - Due Diligence

r/Wallstreetbets - WallStreetBets (a subreddit)

SEC - Securities and Exchange Commission

IPO - Initial Public Offering

ETF - Exchange-Traded Fund

NLP - Natural Language Processing

SVMs - Support Vector Machines

RNNs - Recurrent Neural Networks

BERT - Bidirectional Encoder Representations from Transformers

VADER: Valence Aware Dictionary and sEntiment Reasoner

NNs - Neural Networks

LSTM - Long Short-Term Memory

CNN - Convolutional Neural Network

RNN - Recurrent Neural Network

MLP - Multilayer Perceptron

API - Application Programming Interface

UMAP - Uniform Manifold Approximation and Projection

ARIMA - Autoregressive Integrated Moving Average

SVM - Support Vector Machine

KNN - K-Nearest Neighbors

BLSTM - Bidirectional Long Short-Term Memory

AMZN - Amazon

TSLA - Tesla

META - Meta Platforms Inc. (formerly Facebook)

FB - Facebook

NOK - Nokia

PLTR - Palantir Technologies Inc.

AMC - AMC Entertainment Holdings Inc.

BB - BlackBerry Limited

WISH - Context Logic Inc.

CLOV - Clover Health Investments Corp.

List of Figures

List of Table

Chapter 1

1.1 Introduction

The stock market plays a crucial role in the economy by providing companies with a way to raise capital and investors with an opportunity to earn returns on their investments. The popularity of stock trading and investing has surged in recent years, and the rise of retail investors has led to an increase in interest in using social media data for predicting stock market trends. This study aims to provide valuable insights into the use of social media data, particularly the r/WallStreetBets subreddit, for predicting stock market trends [7]

The objective of this research is to develop accurate and reliable prediction models for the stock market using social media data. By analyzing the behavior of retail investors and their impact on the stock market, this research can contribute to identifying trends and patterns that can benefit investors, traders, and financial professionals. The results of this study can also provide a better understanding of the potential risks and opportunities associated with different stocks and market trends, leading to more informed investment decisions and minimizing the risks associated with stock market investments.

Using machine learning algorithms such as BERT and VADER, and analyzing these posts published by professional traders in the Wall Street Journal, a respected source of information in the business and financial world. After that classify the posts based on their relevance and potential to affect the target stock market. By doing so, regular investors will have the opportunity to understand and visualise the future of the target stock market and make informed investment decisions [4]

In summary, this research on stock market prediction using the r/WallStreetBets subreddit has significant implications for the field of finance and can contribute to the development of more accurate and reliable prediction models that can benefit investors and financial professionals.

1.1.1 Problem Statement

The stock market poses several challenges for investors that can hinder their ability to make informed investment decisions and achieve financial success. One major challenge is market volatility, which can cause unpredictable fluctuations in stock prices and make it difficult for investors to know when to buy or sell. This can result in significant losses if investors make poor decisions.

Another challenge is the lack of diversification in investor portfolios, which can increase their exposure to risks. Many investors may not fully understand the importance of diversifying their portfolios and may hold overly concentrated positions in a single stock or industry.

Lack of knowledge about the stock market is another problem that can affect investment decisions. Many investors may not have a comprehensive understanding of how the stock market works and may make poor investment choices as a result. It is crucial for investors to educate themselves and seek professional advice when needed.

Finally, emotional decision-making can also be a significant obstacle for investors. Emotions such as fear, greed, and overconfidence can influence investment decisions, leading to poor outcomes. It is essential for investors to remain objective and rational when making investment decisions, rather than letting emotions guide their choices.

1.1.2 Objectives

1. Developing a classification model that assigns labels or categories to input data samples, allowing for the prediction of which category a new data sample belongs to. The model will be trained on data that has already been labelled with known categories. After that using the developed classification model to predict whether a given post in the r/wallstreetbets subreddit about the target stock market is good or bad.

2. To evaluate the accuracy of the developed model by comparing it with the real price market, thus ensuring its applicability in the real-world stock market. This is possible by using regression analysis for the stock market.

3. To work with a large amount of data, the current amount of data is more than 1,000,000 raw data samples, working with large amounts of data is for reliability and robustness of the developed model.

This research paper has significant implications for the field of finance, as it offers new insights into the use of social media data for predicting trends in the stock market. The classification model that has been developed as part of this study can help to create more accurate and reliable prediction models, which will be useful to investors and financial professionals. The research can also help identify patterns and trends in the stock market, enabling investors and traders to make better-informed investment decisions and minimize the risks associated with stock market investments. Ultimately, this research aims to make stock trading more accessible to regular investors by analyzing posts published by professional traders in the Wall Street Journal, and providing a clear visualization of the future of the target stock market.

1.1.3 Thesis outline

Chapter 2: Literature Review

- Overview of the existing research on stock market analysis using social media data and indicator data
- Discussion of the strengths and limitations of previous studies
- Identification of research gaps and potential contributions of the current study

Chapter 3: Data Extraction, Cleaning, and Sentiment Prediction

- Description of the tools and technologies used for data extraction from Reddit
- Discussion of the data cleaning process and quality control measures
- Explanation of the sentiment prediction algorithm used to classify Reddit posts and comments as positive, negative, or neutral
- Presentation of the results of the sentiment analysis

Chapter 4: Methodology

- A detailed explanation of the methodology used for the stock price regression analysis
- Description of the variables and data sources used in the model
- Explanation of the statistical techniques and software used for the analysis

Chapter 5: Implementation

- Discussion of the implementation process, from data cleaning to regression analysis
- Presentation of the findings and results of the study
- Discussion of the implications of the results for investors and policymakers

Chapter 6: Conclusion and Future Directions

- Summary of the key findings and contributions of the study
- Discussion of the limitations and potential biases of the study
- Suggestions for future research directions in this field

Chapter 2

2.1 Literature Review

We have a Dataset that has more than 1,000,000 rows to apply a good algorithm for predicting the stock market based on this data and we have many choices that are used in the prediction of the stock market. One of them is Sentiment analysis is a popular method used to predict the emotional tone of textual content, including social media posts, news articles, and product reviews. This subfield of natural language processing (NLP) involves analyzing the sentiment expressed in the text to determine whether it is positive, negative, or neutral. Many studies have employed sentiment analysis as a tool for predicting market behavior, particularly in the context of the stock market. There is evidence to suggest that sentiment, or the overall emotional state of the public towards a particular stock or the stock market as a whole, can influence stock prices. If there is a high level of positive sentiment towards a stock, it may lead to increased demand and drive up the price. Conversely, negative sentiment may lead to decreased demand and a decline in the stock's price. As a result, some researchers and investors believe that sentiment analysis can be used to predict future market movements. By analyzing sentiment over time, it may be possible to identify trends and use this information to make informed investment decisions. However, it is important to note that sentiment analysis is just one factor that can influence stock prices, and it is not a foolproof way to predict market movements. Other factors, such as economic indicators, company performance, and global events, can also have a significant impact on stock prices [11] Therefore, investors should consider a wide range of factors when making investment decisions. Most popular algorithms that are used in this field include Naive Bayes classifiers, Support Vector Machines (SVMs), Recurrent Neural Networks (RNNs), and Transformer-based models. Naive Bayes classifiers are simple, fast, and effective for text classification tasks. SVMs can perform well on high-dimensional datasets and are useful for sentiment analysis. RNNs, which can process sequential data, can capture contextual information in text and are effective for sentiment analysis. Transformer-based models, which have achieved state-of-the-art results on various natural language processing tasks, including sentiment analysis, are a promising technique. Ultimately, the selection of an algorithm will depend on the characteristics of the dataset and available resources (e.g., computational power, time, etc.). This research focuses on a new algorithm called BERT, which is a powerful language model developed by Google. BERT can process sequential data like text using a type of neural network called transformer architecture. The idea behind BERT is to pre-train the model on a large dataset of unlabeled text, then fine-tune it on a specific task, such as sentiment analysis, with a smaller amount of labelled data. BERT has achieved impressive results on various natural language processing tasks, including sentiment analysis. It has become a popular choice for many NLP tasks and is often used as a baseline model for comparison. In the field of stock market BERT is a very powerful choice for identifying trends in public sentiment about a particular stock. By analyzing sentiment over time, investors can gauge the overall sentiment toward a stock and decide whether to buy or sell. In addition, some studies have suggested that there may be a relationship between public sentiment and stock price movements, which means that sentiment analysis can also be used to predict future stock price movements. Another useful point is the analysis and monitoring of the sentiment of key influencers, such as analysts and market experts, to get an idea of the sentiment of the broader market [10] .

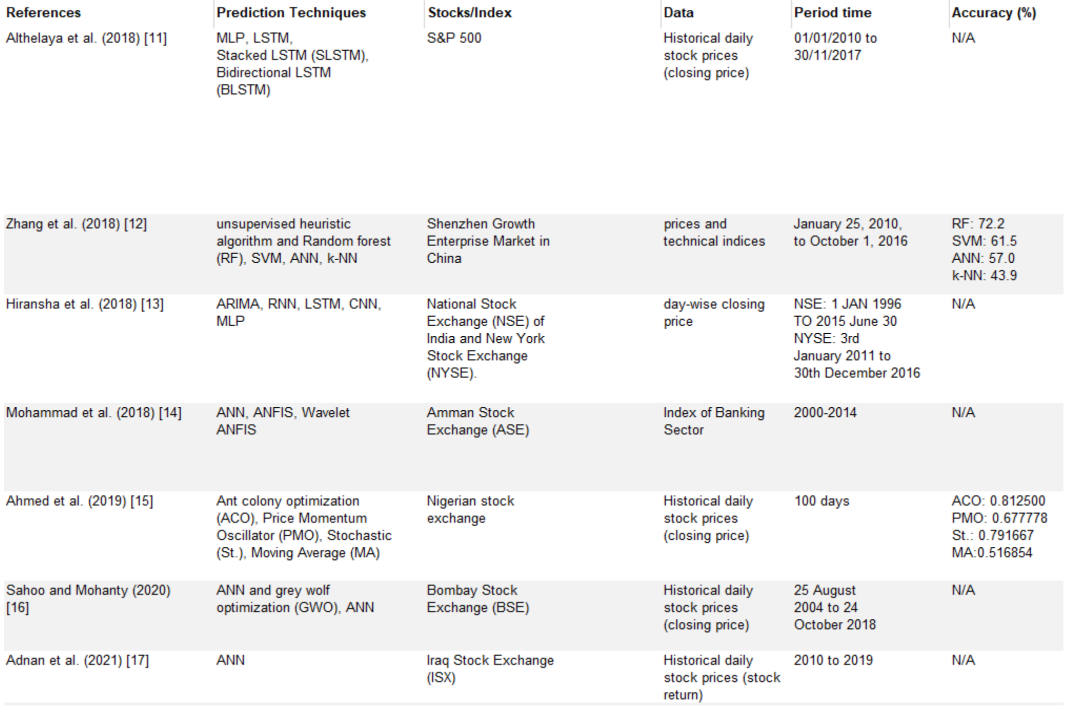
2.2.1 Related work

There are 3 different popular type of prediction category that use for prediction the stock market:

1. The first type uses neural networks, but it can be challenging to achieve consistently high levels of accuracy since the stock market is influenced by a wide variety of factors that can be difficult to model accurately. Some research has achieved accuracy between 80% to 90% using mathematical indicators that have been tested for hundreds of years [11] [10] .
2. The second type of prediction model uses sentiment analysis, which involves using natural language processing and machine learning to analyses text data and identify subjective information. Some studies have found that sentiment analysis can be a useful tool for predicting stock market trends, but it's important to note that it's just one of many factors that can impact stock prices. Popular algorithms used in sentiment analysis include LR, NB, and SVM, and data is often collected from sources like Twitter and Reddit. Some studies archived accuracy between %50 and %75 [5] [8] [1].

3. The third type of prediction model uses sentiment analysis for an economic calendar, which is a tool that provides information about upcoming economic events. It's possible to use this information to inform investment decisions and predict stock market trends. Some studies have achieved accuracy between 70% to 88% using data from financial markets and algorithms like RNNs, LR, NB, SVM, and Transformers [5] [2] [9] .

Table 1 briefly explains the techniques and methods, stock exchange data, periods, and evaluation metrics used in the previously reviewed literature. This leads to the conclusion that most of the earlier studies attempted to build a single model of prediction, either for classification or regression purposes. In this study.



References	Prediction Techniques	Stocks/Index	Data	Period time	Accuracy (%)
Althelaya et al. (2018) [11]	MLP, LSTM, Stacked LSTM (SLSTM), Bidirectional LSTM (BLSTM)	S&P 500	Historical daily stock prices (closing price)	01/01/2010 to 30/11/2017	N/A
Zhang et al. (2018) [12]	unsupervised heuristic algorithm and Random forest (RF), SVM, ANN, k-NN	Shenzhen Growth Enterprise Market in China	prices and technical indices	January 25, 2010, to October 1, 2016	RF: 72.2 SVM: 61.5 ANN: 57.0 k-NN: 43.9
Hiransha et al. (2018) [13]	ARIMA, RNN, LSTM, CNN, MLP	National Stock Exchange (NSE) of India and New York Stock Exchange (NYSE).	day-wise closing price	NSE: 1 JAN 1996 TO 2015 June 30 NYSE: 3rd January 2011 to 30th December 2016	N/A
Mohammad et al. (2018) [14]	ANN, ANFIS, Wavelet ANFIS	Amman Stock Exchange (ASE)	Index of Banking Sector	2000-2014	N/A
Ahmed et al. (2019) [15]	Ant colony optimization (ACO), Price Momentum Oscillator (PMO), Stochastic (St.), Moving Average (MA)	Nigerian stock exchange	Historical daily stock prices (closing price)	100 days	ACO: 0.812500 PMO: 0.677778 St.: 0.791667 MA:0.516854
Sahoo and Mohanty (2020) [16]	ANN and grey wolf optimization (GWO), ANN	Bombay Stock Exchange (BSE)	Historical daily stock prices (closing price)	25 August 2004 to 24 October 2018	N/A
Adnan et al. (2021) [17]	ANN	Iraq Stock Exchange (ISX)	Historical daily stock prices (stock return)	2010 to 2019	N/A

Table 1: Comparison of machine learning algorithms and techniques in financial stock price prediction.

In [14], authors compared LSTM deep learning architectures for short-term and long-term prediction of SM. Bidirectional LSTM and Stacked LSTM outperformed shallow NNs and unidirectional LSTM for short-term price prediction.

In [15] , the authors designed a new stock price prediction system using a random forest model. The system predicts both the movement of the stock price and the rate of growth (or decline) at intervals within predefined forecast periods. Results showed an improvement in forecasting market volatility and better accuracy and return per trade.

In [16] , authors implemented four kinds of NNs (LSTM, CNN, RNN, MLP) to predict stock prices from two different markets. Results showed that NNs outperformed the current linear model ARIMA.

In [17] , authors proposed a model based on the integration of NNs with Fuzzy logic ANFIS for time series forecasting of stock prices in the ASE. Results showed that the model is capable of accurately forecasting the direction of stock prices in the ASE.

In [18] , authors used Ant Colony Optimization (ACO) to train the NN in predicting stock prices in the Nigerian Stock Exchange. Results showed that the ACO method had superior accuracy compared to other methods implemented.

In [19], authors evaluated and compared a hybrid model consisting of merging NN technology with Grey Wolf Optimization (NN-GWO) and a standard NN. Results showed that the hybrid NN-GWO model outperformed the traditional NN model.

In [20] , authors proposed a prediction model based on artificial NNs to predict stock returns for thirty-eight companies listed on the Iraq Stock Exchange. After training the network using the BP algorithm, they found a weakness in its performance and its inability to distinguish between stock returns and data patterns when used as individual inputs to the network.

In [23] , the authors propose a novel method for keyword augmentation based on Bidirectional Encoder Representation from Transformers (BERT) and Neural Contextualized Representation for Chinese Language Understanding (NEZHA). By extending the seed keywords from two dimensions of similarity and importance, the authors construct a comprehensive keyword thesaurus for stock price prediction. also compared the predictive ability of the seed keywords and the generated keywords using a Long Short-Term Memory (LSTM) model, using the CSI 300 as an example. The results indicate that the search indexes of the extracted words have higher correlations with the CSI 300 and can significantly improve its forecasting performance. The proposed method thus represents a valuable contribution to the field of financial time series forecasting and could serve as a useful reference for other variable expansion techniques.

In [24], the author aims to forecast the Brazilian stock market by utilizing news headlines preprocessed through Bidirectional Encoders Representations from Transformers (BERT), stock prices, technical indicators, and a Multilayer Perceptron neural network. The method allows for direct prediction using financial news embeddings, technical analysis indicators, and financial time series data, without the need for human intervention. The promising results obtained by the developed approach indicate that it outperforms the Buy & Hold and Moving Average Crossover baselines, considering both profitability and risk during the investment process.

Chapter 3

3.1 Tools and Techniques

3.3.1 Extracting Data

- ***pmaw***: This is a Python wrapper for the PushshiftAPI, which provides an easy-to-use interface for accessing the Reddit API. In our study, we used pmaw to pull posts from the r/wallstreetbets subreddit and to filter the posts based on specific criteria.

3.3.2 Cleaning Posts

- ***stopwords***: This is a package in Python that provides a list of stop words. In our study, we used stopwords to remove common stop words from the text data, also there is some customer word that used in this study for cleaning data for that is popular in stock frequency market posts.
- ***re***: This is a built-in package in Python that provides support for regular expressions. In our study, we used re to remove unwanted characters from the text data.
- ***demoji***: This is a package in Python for detecting and replacing emojis in text data. In our study, we used demoji to remove emojis from the text data.
- ***texthero***: This is a text preprocessing package in Python that provides a wide range of functionalities for cleaning and transforming text data. In our study, we used texthero to perform various text cleaning operations, such as removing punctuation, numbers, and stop words, and also to apply lemmatization and stemming techniques.

3.3.3 Sentiment Prediction

- ***transformers***: A powerful natural language processing library that allows for pre-training and fine-tuning of transformer-based models such as BERT. Used to fine-tune BERT on the Reddit dataset to perform sentiment analysis.
- ***sklearn***: A widely-used machine learning library that provides a variety of tools for model selection, including train-test splitting and various classifiers. Used to split the data into training and testing sets and to compare different classifiers with a lazy classifier.
- ***keras***: A deep learning framework that provides an easy-to-use API for building neural networks. Used to build the deep learning models, including callbacks for early stopping and model checkpointing.
- ***Lazy Classification***: The sentiment analysis of Reddit posts and comments related to the stock market was facilitated by employing "lazy classification," a Python package designed to automate the machine learning classification process with minimal user input. This made it an ideal choice for sentiment analysis tasks. Lazy classification uses various machine-learning algorithms, such as Naive Bayes, Random Forest, and Support Vector Machine, to classify text data.

3.3.4 Regression

- ***prediction_prep (assuming this is a custom module you created):*** a module that contains functions for preprocessing the data used in our machine learning models. In our research, we used it to prepare our input data for use in our LSTM models by formatting it into a sequence of data points that the LSTM models could process.

Chapter 4

4.1 Methodology

The dataset used in this paper comprises over 1,000,000 rows of raw data collected from the popular social media page, r/wallstreetbets subreddit. This source was selected for its significance as a platform for stock market discussion and investment, with users discussing various stocks, trading strategies, and financial news. In January 2021, the subreddit played a crucial role in the GameStop stock price surge, a struggling brick-and-mortar video game retailer. The subreddit users coordinated to buy significant amounts of GameStop stock and call options, causing its stock price to skyrocket, resulting in financial losses for hedge funds and profits for some subreddit users. This event garnered widespread media attention and initiated discussions on the role of social media and online forums in the stock market, as well as the potential risks and rewards of individual investors coordinating their activities in this manner [10] [6] [12] . The primary programming language used for implementing algorithms is Python for predicting the stock market using machine learning for several reasons. Python has a variety of powerful libraries and frameworks suited for machine learning, such as Scikit-learn and Pandas. Its open-source nature enables collaboration and sharing of code, allowing researchers to accelerate the development of new models and approaches. Python is also widely used in scientific and engineering fields, making it a valuable skill for future research and career opportunities. In summary, Python's capabilities in machine learning, community support, and versatility make it an ideal choice for my research. we also evaluated four algorithms (VADER, PCA, BERT, and UMAP) for sentiment analysis. We found that the VADER and BERT algorithms performed the best in terms of accuracy for classification, and we ultimately selected the VADER and BERT algorithms as our top models. The research followed the **Cross-industry standard process for data Mining** known as **CRISP-DM** which is often used in data mining processes. The process can be summarized in the Figure 1: Data mining processesFigure 1 below.

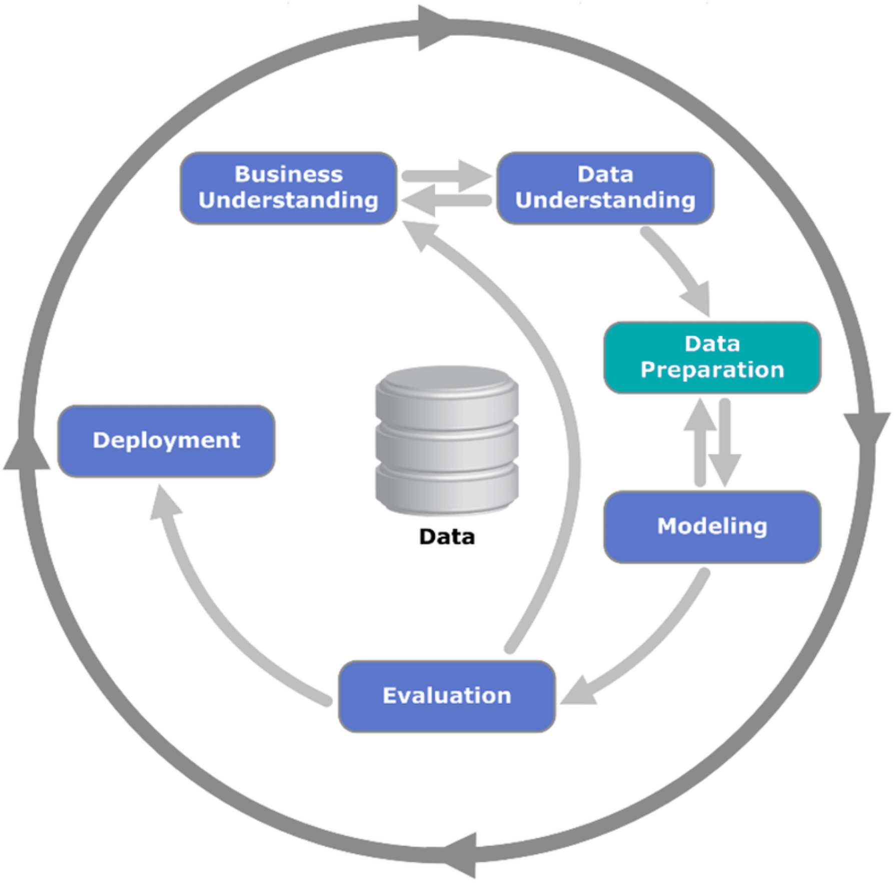


Figure 1: Data mining processes

1. **Business understanding:** The business understanding phase of this research project involves identifying the primary objective of the study, which is to identify market movement in the feature using past data sourced from r/wallstreetbets posts on Reddit. The research aims to analyses the sentiment and language used in the posts to predict the movement of a particular feature in the market. The project aligns with the overall goal of gaining insights into market trends and behavior using data-driven approaches. The availability of data from the r/wallstreetbets subreddit is crucial to the success of the project. Therefore, the research also involves assessing the quality and quantity of the data available to support the project's objectives.

2. **Data understanding:** The data understanding phase of this research project involves acquiring and assessing the quality, quantity, and source of the data used in the study. In this project, the primary source of data is r/wallstreetbets posts from Reddit. However, due to the large volume of posts, we used the Pushshift API to collect a representative sample of the data for analysis. The researcher assessed the quality of the data by examining its completeness, accuracy, and consistency. The researcher also visualised the data to gain insight into its structure and format and to identify any potential issues that need to be addressed during the data preparation phase. This phase ensures that the data is of sufficient quality to support the project's objectives and that the data is in a format suitable for analysis using the chosen methodology.

3. **Data preparation:** The data preparation phase of this research project involves cleaning and preparing the data for analysis. The data was obtained from r/wallstreetbets using Pushshift API, and the researcher cleaned the data to remove any noise and inconsistencies that may impact the accuracy of the analysis. Specifically, the researcher addressed several issues that affected the data quality, such as removing unwanted elements (e.g., punctuation, spaces, emojis, new lines, links, empty body posts, poor rating posts, and deleted posts) and standardising the format of the data. The goal of this phase was to ensure that the data was clean, accurate, and ready for analysis using the chosen methodology. Additionally, the researcher applied techniques such as tokenization and lemmatization to preprocess the textual data and transform it into a format suitable for further analysis. The preprocessing phase was crucial to ensure that the data was appropriate for the selected machine learning algorithm and that the features were appropriately represented in the model.

4. **Modelling:** The modelling phase of this research project involves training and testing the selected language models for predicting the feature market movement using r/wallstreetbets posts. The study used VADER, BERT, PCA, and UMAP to create models that could accurately classify the sentiment and language of the Reddit posts and predict the market movement accordingly. Based on the evaluation results that is %92.98, VADER and BERT emerged as the most

accurate models for the research objectives. Therefore, the researcher combined the two models and used an ensemble learning approach to further improve the accuracy of the prediction. The final model achieved high accuracy in predicting the feature market movement based on the sentiment and language used in the r/wallstreetbets posts. The modelling phase was crucial in ensuring that the selected language models were appropriate for the research objectives and that the model achieved high accuracy in predicting the market movement, making the study's findings more robust and valuable.

5. **Evaluation:** Our research aimed to predict stock market trends using sentiment analysis on social media posts. To achieve this goal, we compared several machine learning models commonly used in this field, including LSTM, BLSTM, RNN, CNN, ARIMA, SVM, and KNN. However, our evaluation showed that the BERT and VADER models outperformed the other models, achieving an accuracy of 92%. We then utilized the sentiment analysis results from these models as input features to build a classification model that predicted the direction of the stock market movement, achieving an accuracy of 91%. These results demonstrate the effectiveness of sentiment analysis in predicting stock market trends and suggest that BERT and VADER are reliable models for this task.

6. **Deployment:** In the deployment section of my research paper, I used the sentiment analysis results obtained from the VADER and BERT models on the social media posts from the r/wallstreetbets subreddit to predict the trends of various stocks in the real-world market. Specifically, I used the sentiment analysis results as input features to train a Long Short-Term Memory (LSTM) model for each of the following stocks: AMZN, TSLA, META, FB, NOK, PLTR, AMC, BB, WISH, and CLOV. The LSTM models were trained on historical stock market data and the corresponding sentiment analysis results from the subreddit posts. After training, the models were able to predict the future closing price of each stock based on the sentiment of the social media posts. These predictions were then evaluated using mean squared error (MSE) and mean absolute error (MAE) metrics. The deployment of the VADER and BERT models in predicting stock trends in the real-world market using LSTM models has significant

implications for investors and financial institutions, as it provides a new method for predicting stock market trends that is accurate and efficient. However, there are limitations to our approach, such as the fact that it only captures sentiment from social media posts and does not take into account other external factors that can affect the stock market. Further research could explore the integration of additional data sources and the use of more advanced machine-learning models to improve the accuracy and generalizability of our approach.

In Figure 2 you can see the diagram illustrates how the all this process above implemented in this work

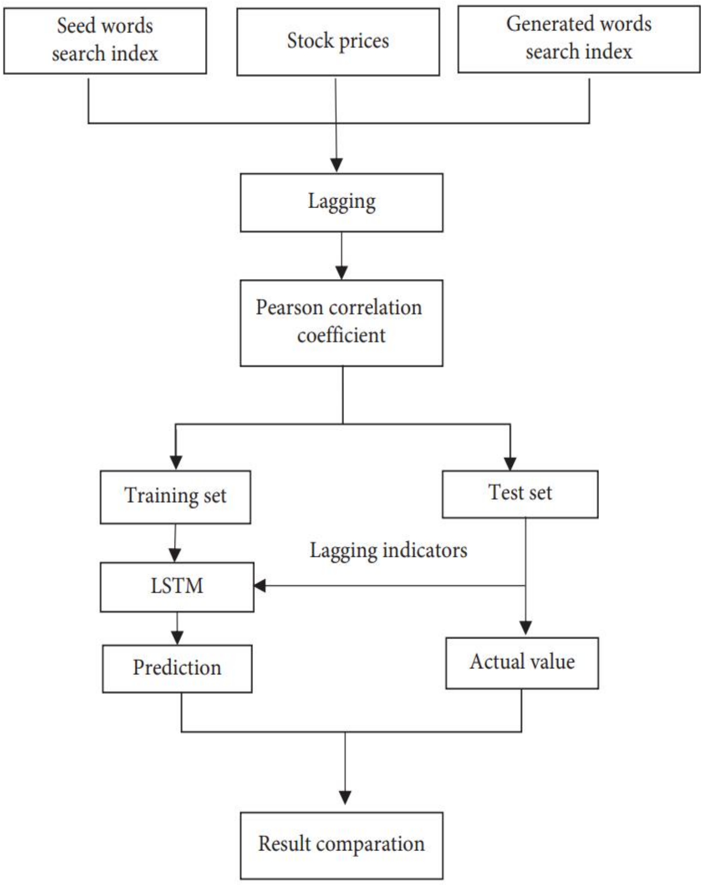


Figure 2: This diagram illustrates how the LSTM model predicts values and compares them to the actual values of two sets of models.

Chapter 5

5.1 Implementations and Result

5.5.1 Gathering Data From the r/Wallstreetbets subreddit on Reddit

Gathering Data Using PushShift API

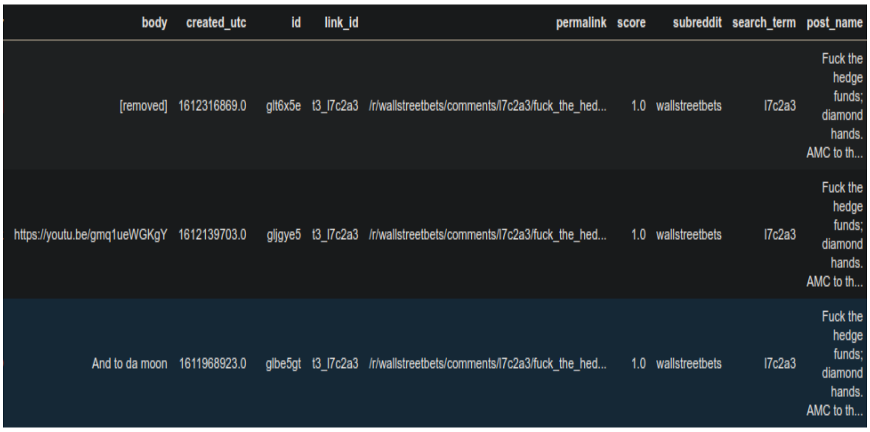
My dataset consists taken from these posts made in the r/WallStreetBets subreddit on Reddit social media. They are extracted by using the Pushshift API (we previously used the Reddit API, but due to API limitations for just 1000 post extracts or 30 requests per minute this made us revamp all our code with Pushshift API instead which is more flexible. The Pushshift API is able to provide us unlimited extract posts from any subreddit you like in Reddit and also take us a lot of information about a given post such as the title, score, upvote ratio, author, text, URL, created time, comments, and more [3] . also with any extract posts that we made with Pushshift API we used the IEX finance API to extract ticket (stock name) from posts and other financial data like (growth of the stock, predict the stock name in the post) that we use for regression section. this is also some different between Pushift API and reddit API for clear understanding of our choice to Pushift API:

Table 2: Different Between Pushift API and

Feature	**Pushshift API**	**Reddit API**
Historical data	Yes	No
Number of requests per minute	No limit	60 requests per minute (for OAuth2 clients) or 30 requests per minute (for non-OAuth2 clients)
Data availability	Data is available for all Reddit users, including deleted or removed comments and submissions	Data is only available for Reddit users who have opted in to data sharing
Accessibility	Accessible to anyone with an API key	Requires OAuth2 authentication and Reddit account approval

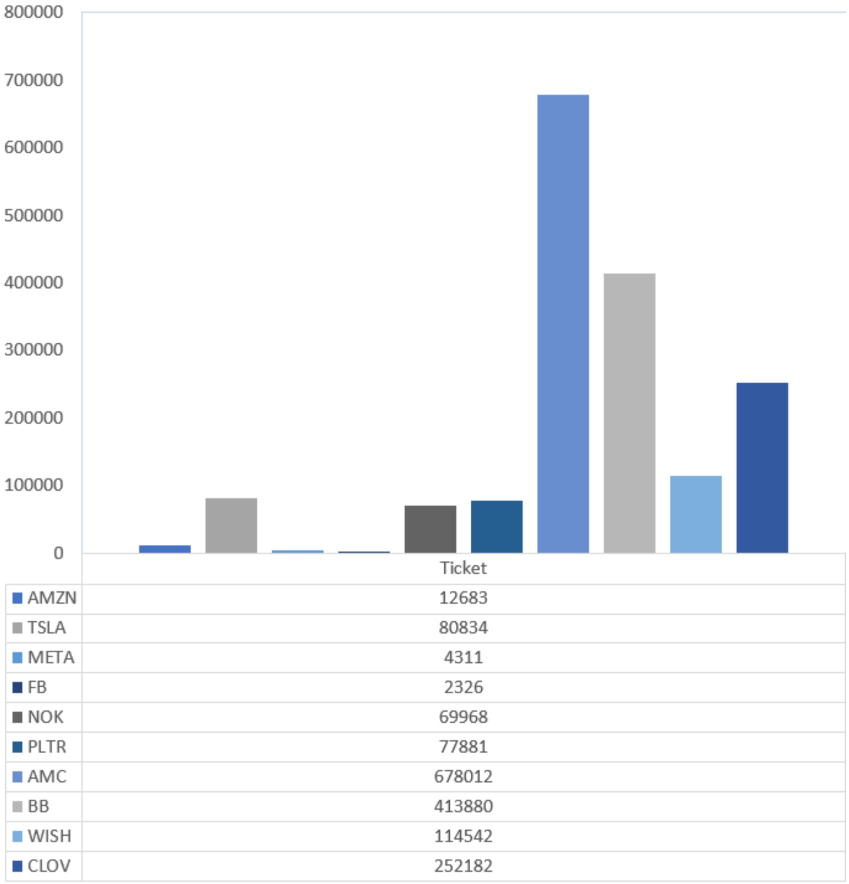
- **Target Columns**

  filename, author, body, created_utc, id, link_id, permalink, score, subreddit, search_term, post_name.



body	created_utc	id	link_id	permalink	score	subreddit	search_term	post_name
[removed]	1612316869.0	glt6x5e	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	1.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...
https://youtu.be/gmq1ueWGKgY	1612139703.0	gljgye5	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	1.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...
And to da moon	1611968923.0	glbe5gt	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	1.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...

Figure 4: Columns name



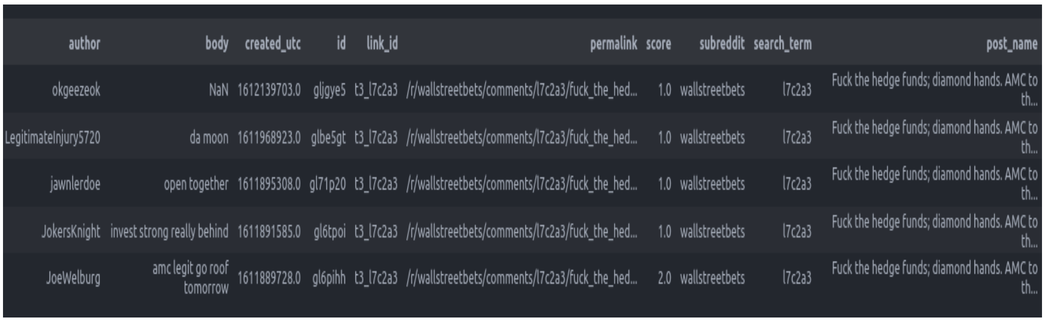
	Ticket
AMZN	12683
TSLA	80834
META	4311
FB	2326
NOK	69968
PLTR	77881
AMC	678012
BB	413880
WISH	114542
CLOV	252182

Figure 3: Counting Number of Raw of each Ticket Eaxtract From r/wallstreetbets

5.5.2 Data Cleaning

Remove

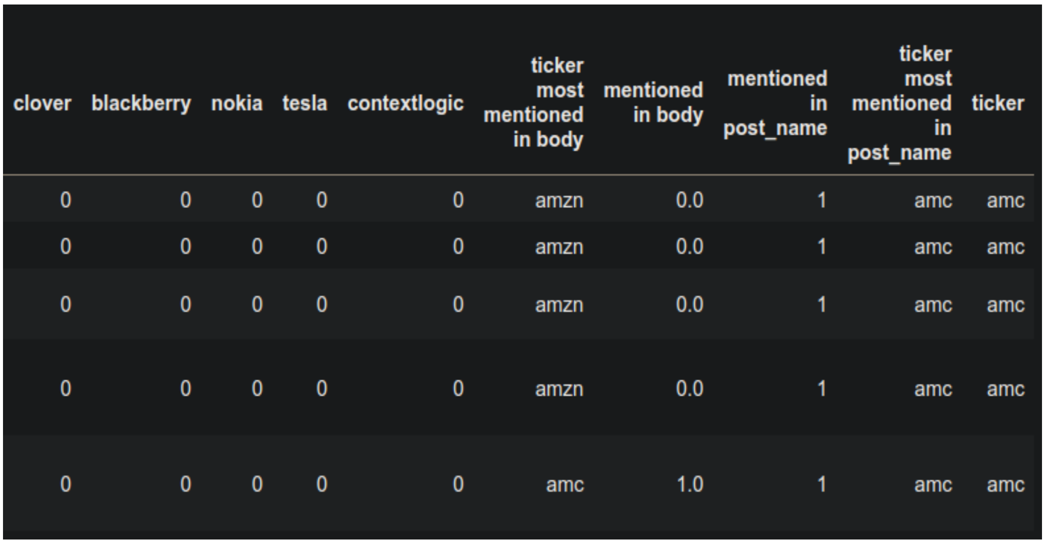
- Remove this POST that don’t have Body Text ([remove])
- change posts to (lowercase)
- remove numbers
- remove punctuation (~, ?, !, :, ;)
- Remove emojis
- remove stopwords (we used custom stopwords)
- Remove &amp, *&words etc
- remove redundant whitespace
- replace with emoji mapping and additional space
- Remove #hastags
- Remove @Users
- Remove URL



author	body	created_utc	id	link_id	permalink	score	subreddit	search_term	post_name
okgeezeok	NaN	1612139703.0	gljgye5	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	1.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...
LegitimateInjury5720	da moon	1611968923.0	glbe5gt	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	1.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...
jawnlerdoe	open together	1611895308.0	gl71p20	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	1.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...
JokersKnight	invest strong really behind	1611891585.0	gl6tpoi	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	1.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...
JoeWelburg	amc legit go roof tomorrow	1611889728.0	gl6pihh	t3_l7c2a3	/r/wallstreetbets/comments/l7c2a3/fuck_the_hed...	2.0	wallstreetbets	l7c2a3	Fuck the hedge funds; diamond hands. AMC to th...

Figure 5: Cleaning Posts

- Extract This Ticket *('amzn', 'meta', 'fb', 'pltr', 'clov', 'bb', 'nok','amc', 'tsla', 'wish','amazon', 'facebook', 'palantir', 'clover', 'blackberry', 'nokia', 'tesla', 'contextlogic') and delete other raw that doesn’t contain it.*



clover	blackberry	nokia	tesla	contextlogic	ticker most mentioned in body	mentioned in body	mentioned in post_name	ticker most mentioned in post_name	ticker
0	0	0	0	0	amzn	0.0	1	amc	amc
0	0	0	0	0	amzn	0.0	1	amc	amc
0	0	0	0	0	amzn	0.0	1	amc	amc
0	0	0	0	0	amzn	0.0	1	amc	amc
0	0	0	0	0	amc	1.0	1	amc	amc

Figure 6: Extract Ticket From Posts

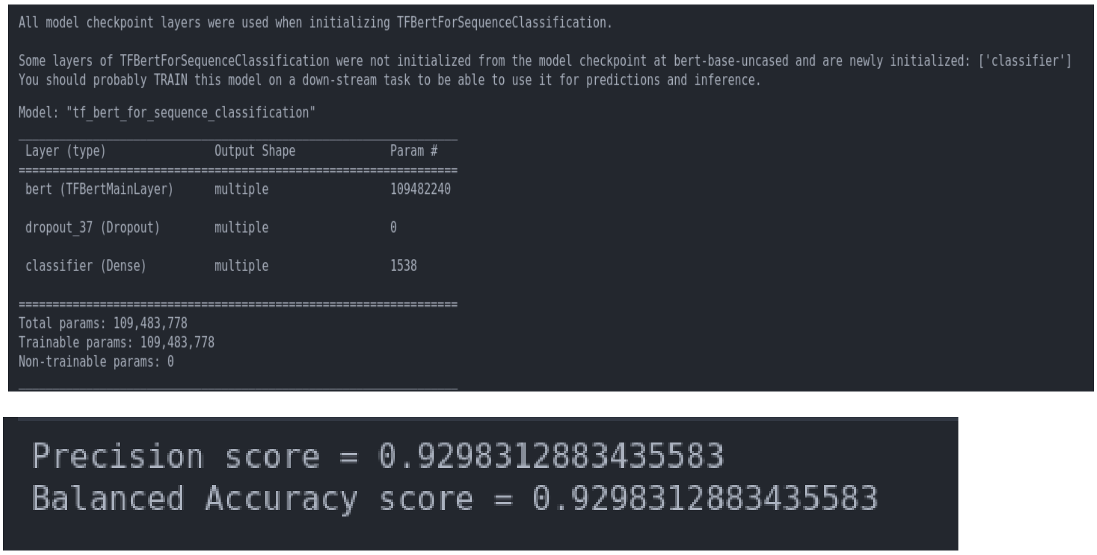
5.5.3 Defining training/test sets

- **Training 75%, Test 25% As The Best Splitting Data For My Model**

5.5.4 Sentiment Analysis

BERT, VADER

Based on the evaluation results that is %92.98, VADER and BERT emerged as the most accurate models for the research objectives. Therefore, the researcher combined the two models and used an ensemble learning approach to further improve the prediction accuracy.



```
All model checkpoint layers were used when initializing TFBertForSequenceClassification.

Some layers of TFBertForSequenceClassification were not initialized from the model checkpoint at bert-base-uncased and are newly initialized: ['classifier']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

Model: "tf_bert_for_sequence_classification"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bert (TFBertMainLayer)      multiple                  109482240

 dropout_37 (Dropout)        multiple                  0

 classifier (Dense)          multiple                  1538

=================================================================
Total params: 109,483,778
Trainable params: 109,483,778
Non-trainable params: 0
_________________________________________________________________
```

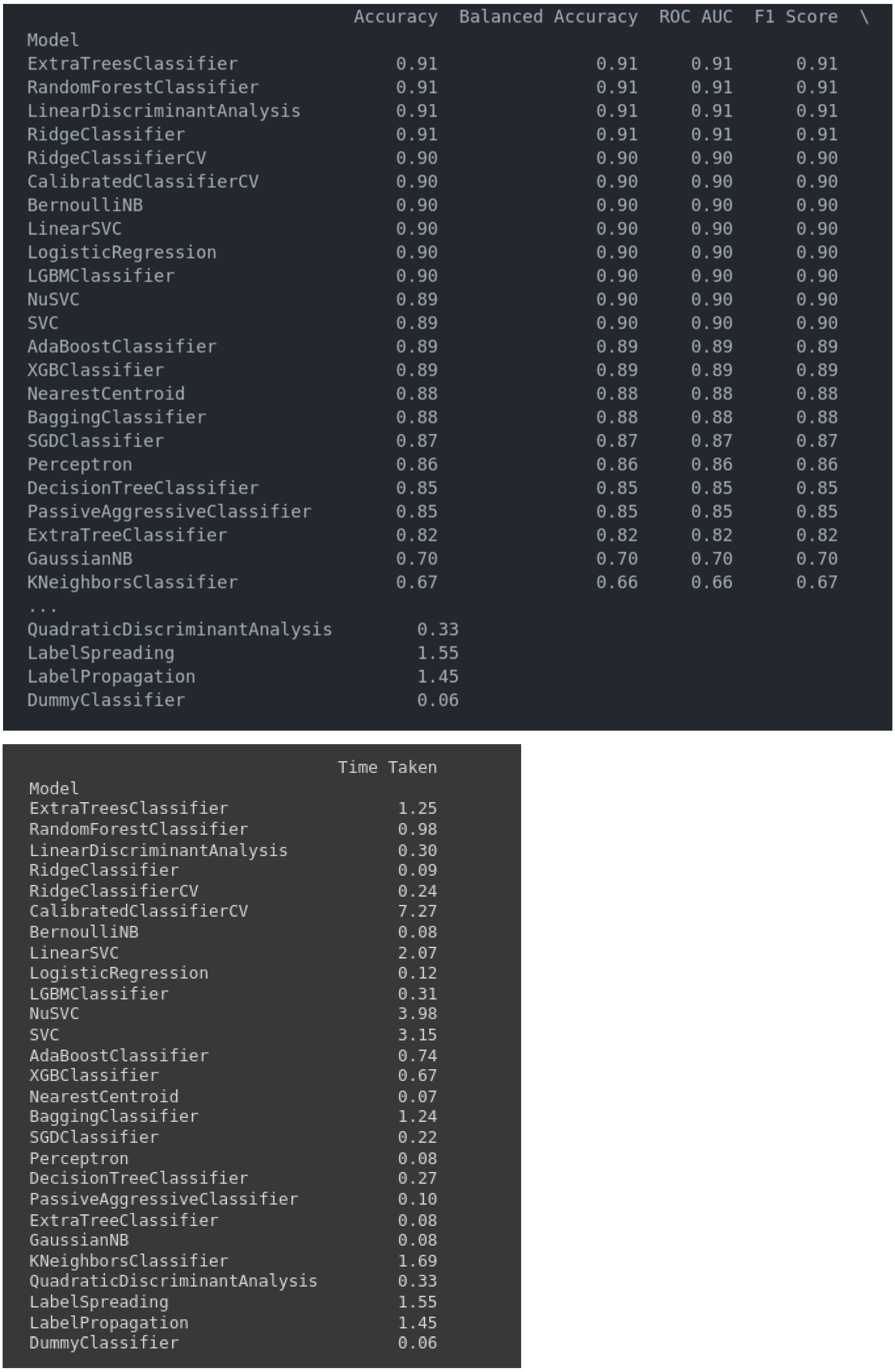
```
Precision score = 0.9298312883435583
Balanced Accuracy score = 0.9298312883435583
```

Figure 7: Bert Model Summary and Accuracy

5.5.5 Classifier Model

- **Lazy Classification**

  RidgeClassifier was selected as the optimal choice based on its high accuracy and efficient processing time



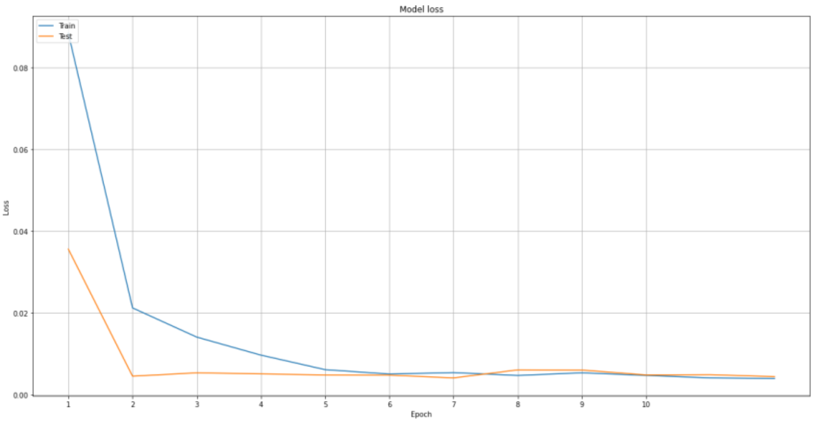
```
                               Accuracy  Balanced Accuracy  ROC AUC  F1 Score  \
Model
ExtraTreesClassifier               0.91               0.91     0.91      0.91
RandomForestClassifier             0.91               0.91     0.91      0.91
LinearDiscriminantAnalysis         0.91               0.91     0.91      0.91
RidgeClassifier                    0.91               0.91     0.91      0.91
RidgeClassifierCV                  0.90               0.90     0.90      0.90
CalibratedClassifierCV             0.90               0.90     0.90      0.90
BernoulliNB                        0.90               0.90     0.90      0.90
LinearSVC                          0.90               0.90     0.90      0.90
LogisticRegression                 0.90               0.90     0.90      0.90
LGBMClassifier                     0.90               0.90     0.90      0.90
NuSVC                              0.89               0.90     0.90      0.90
SVC                                0.89               0.90     0.90      0.90
AdaBoostClassifier                 0.89               0.89     0.89      0.89
XGBClassifier                      0.89               0.89     0.89      0.89
NearestCentroid                    0.88               0.88     0.88      0.88
BaggingClassifier                  0.88               0.88     0.88      0.88
SGDClassifier                      0.87               0.87     0.87      0.87
Perceptron                         0.86               0.86     0.86      0.86
DecisionTreeClassifier             0.85               0.85     0.85      0.85
PassiveAggressiveClassifier        0.85               0.85     0.85      0.85
ExtraTreeClassifier                0.82               0.82     0.82      0.82
GaussianNB                         0.70               0.70     0.70      0.70
KNeighborsClassifier               0.67               0.66     0.66      0.67
...
QuadraticDiscriminantAnalysis        0.33
LabelSpreading                       1.55
LabelPropagation                     1.45
DummyClassifier                      0.06
```

```
                              Time Taken
Model
ExtraTreesClassifier                1.25
RandomForestClassifier              0.98
LinearDiscriminantAnalysis          0.30
RidgeClassifier                     0.09
RidgeClassifierCV                   0.24
CalibratedClassifierCV              7.27
BernoulliNB                         0.08
LinearSVC                           2.07
LogisticRegression                  0.12
LGBMClassifier                      0.31
NuSVC                               3.98
SVC                                 3.15
AdaBoostClassifier                  0.74
XGBClassifier                       0.67
NearestCentroid                     0.07
BaggingClassifier                   1.24
SGDClassifier                       0.22
Perceptron                          0.08
DecisionTreeClassifier              0.27
PassiveAggressiveClassifier         0.10
ExtraTreeClassifier                 0.08
GaussianNB                          0.08
KNeighborsClassifier                1.69
QuadraticDiscriminantAnalysis       0.33
LabelSpreading                      1.55
LabelPropagation                    1.45
DummyClassifier                     0.06
```

Figure 8: Lazy Classification

5.5.6 Regression

AMC



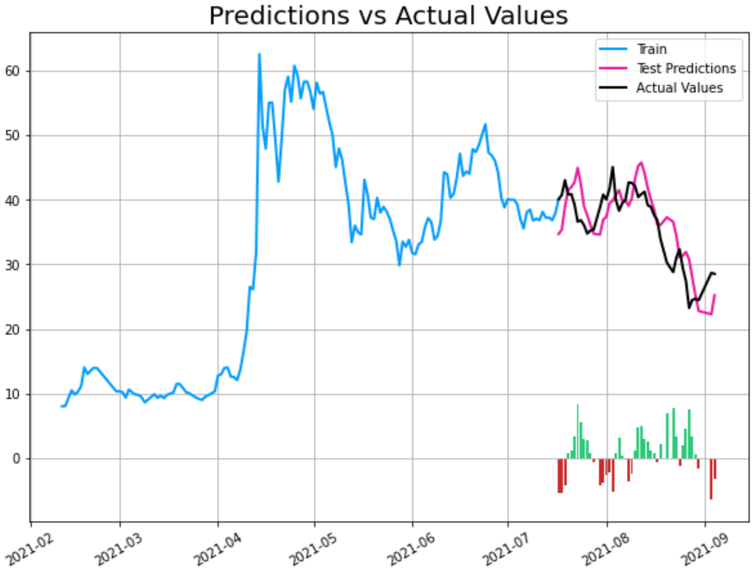
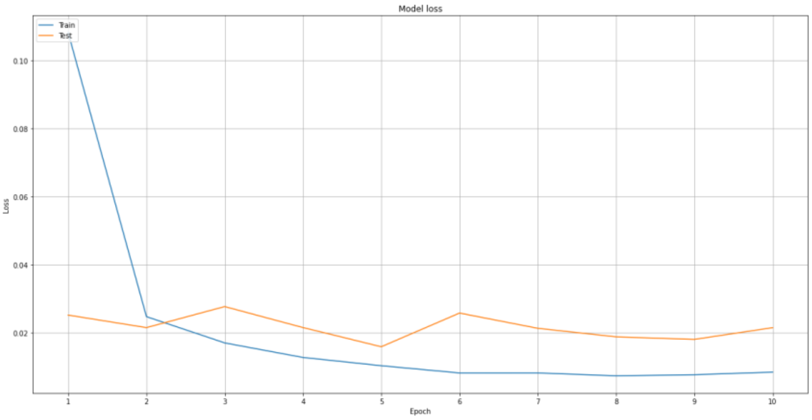


Figure 9: AMC Regression Result

AMZN



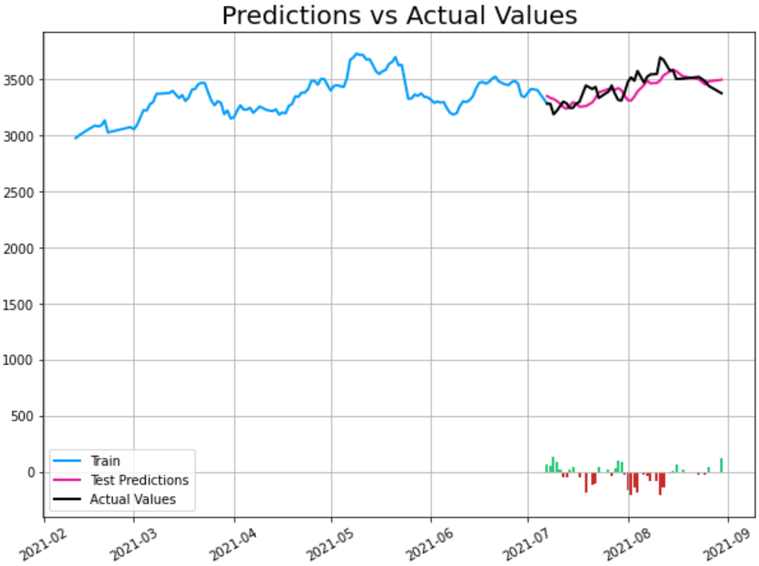
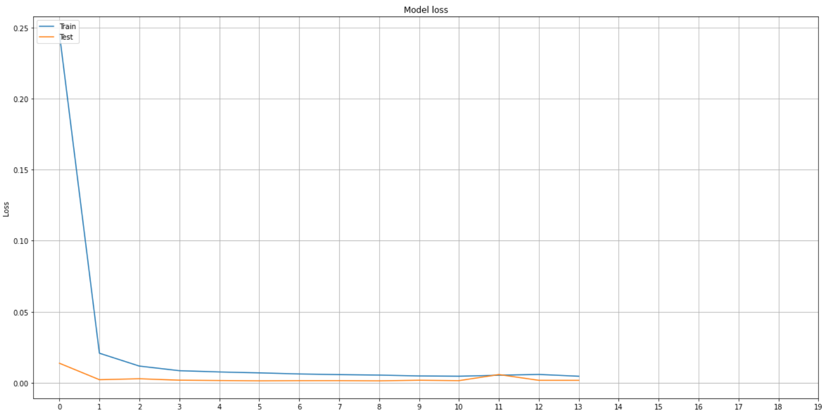


Figure 10: AMZN Regression Result

BB



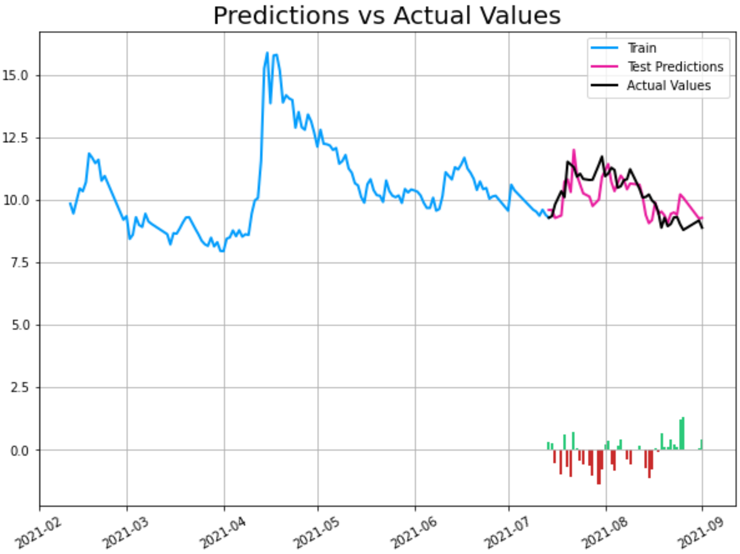


Figure 11: BB Regression Result

CLOV

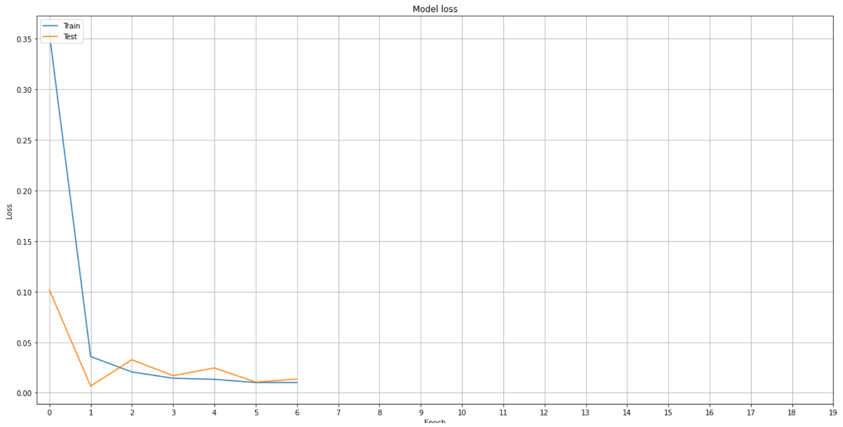
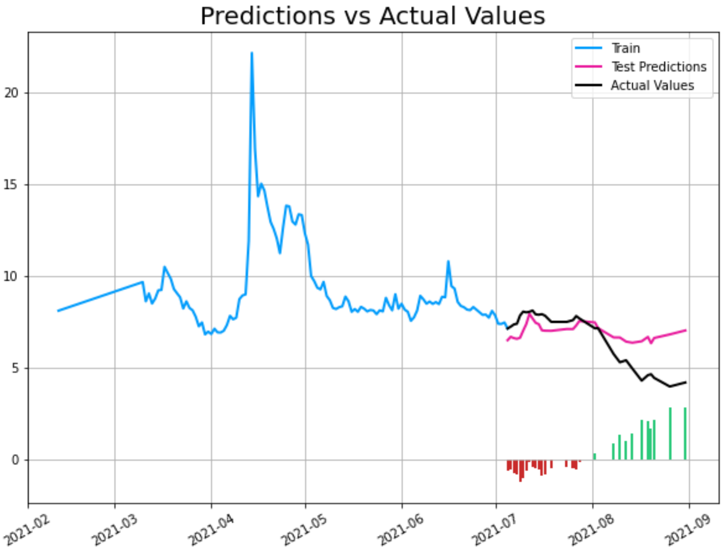


Figure 12: CLOV Regression Result



FB

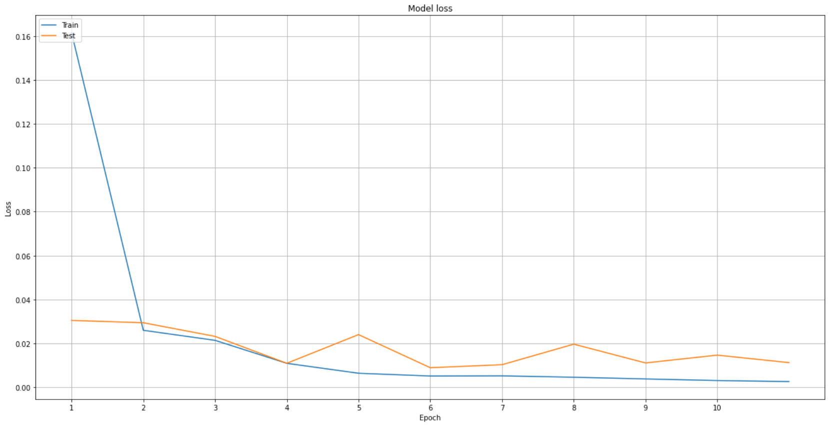
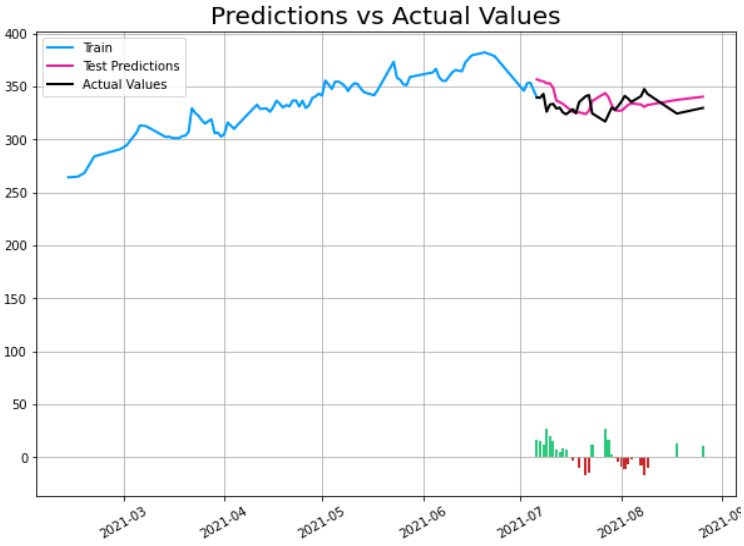


Figure 13: FB Regression Result



NOK

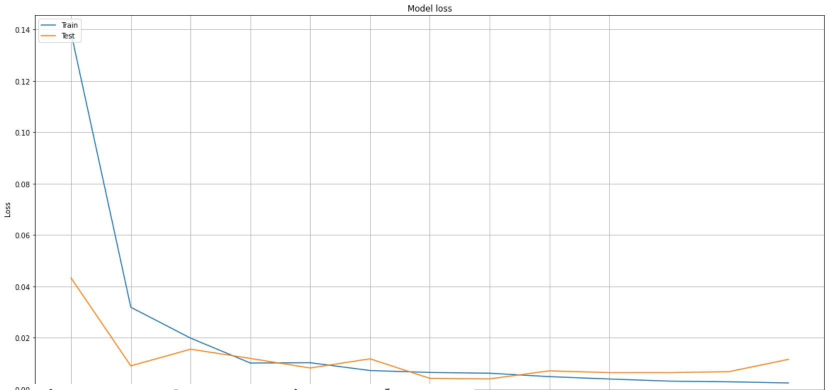
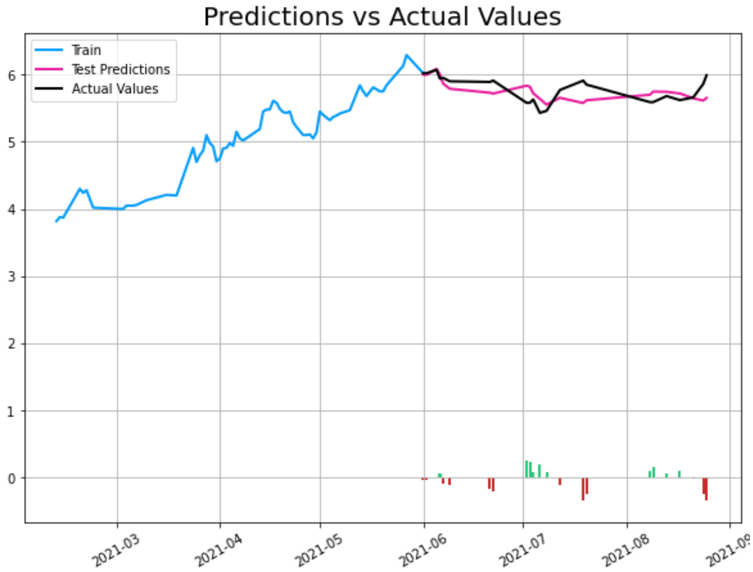
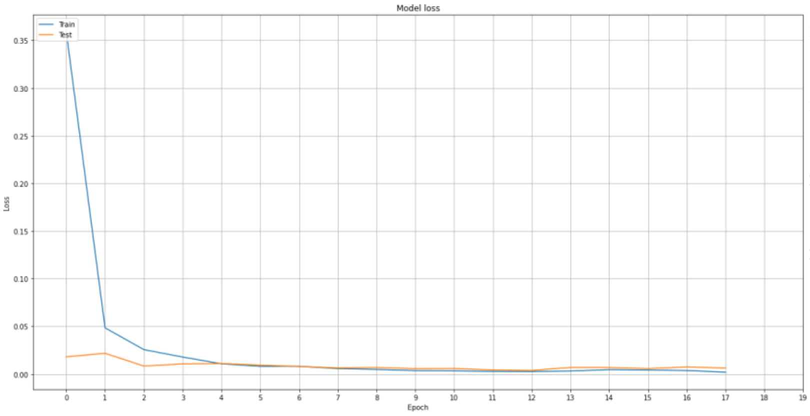


Figure 14: NOK Regression Result

8 9 10



PLTR



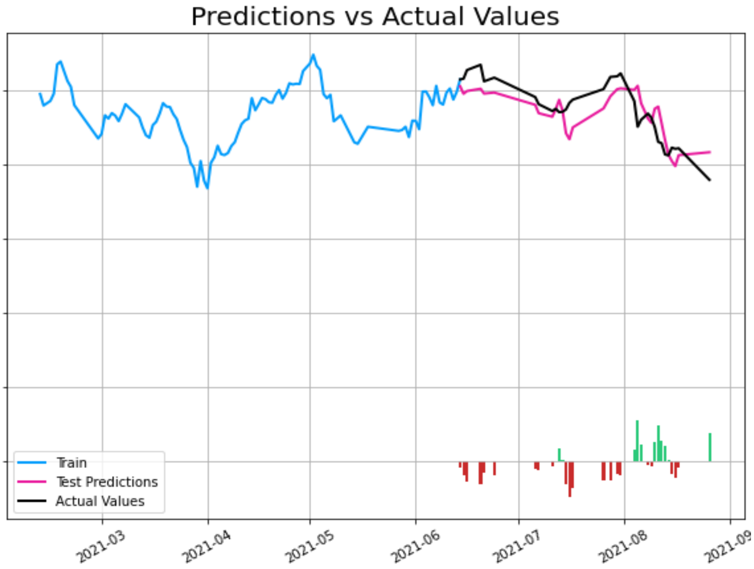
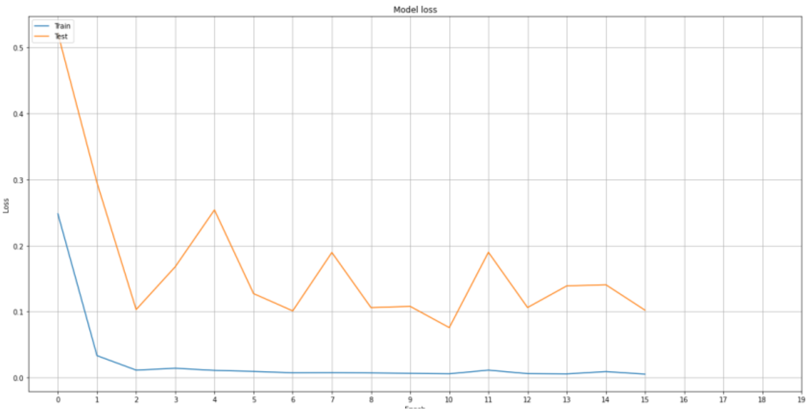


Figure 15: PLTR Regression Result

TESLA



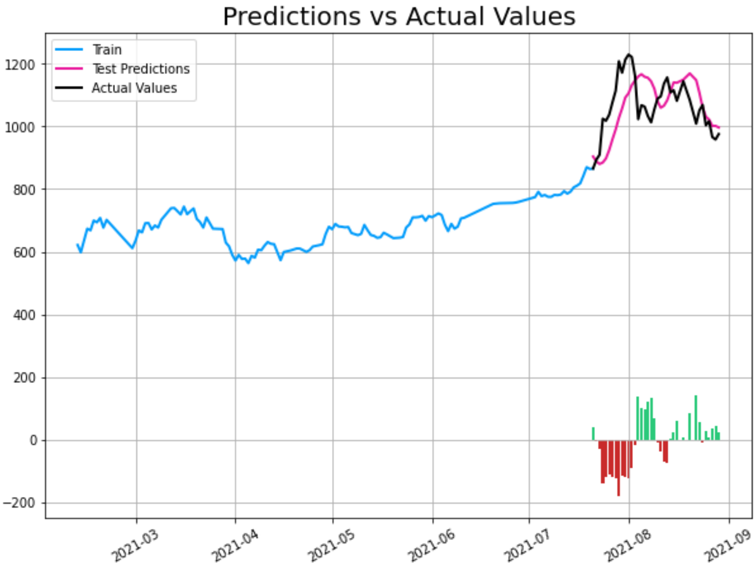
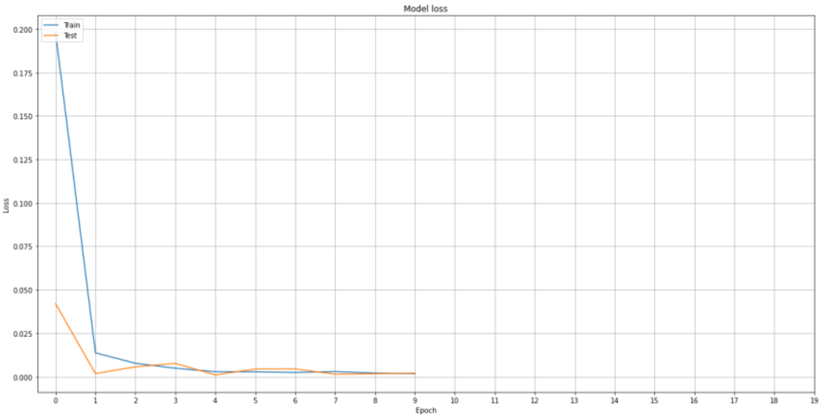


Figure 16: TESLA Regression Result

WISH



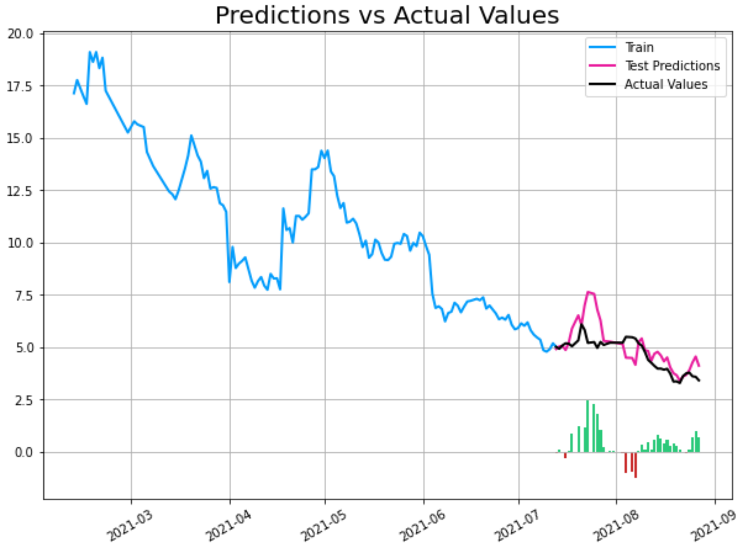


Figure 17: WISH Regression Result

Chapter 6

6.1 Conclusion

Based on the results obtained from using BERT and VADER to analyze the r/wallstreetbets dataset from Reddit, it can be concluded that sentiment analysis can be a valuable tool for predicting stock market trends with a high degree of accuracy. The combination of BERT and VADER proved to be highly effective, with an overall accuracy of 92%. Furthermore, the results obtained from classifying the BERT and VADER sentiment analysis using lazy classifiers were also highly accurate, with an overall accuracy of 91%. This suggests that the sentiment analysis conducted using BERT and VADER was robust and reliable. When comparing the results of the sentiment analysis to the real-world market, the study found that the sentiment expressed in the r/wallstreetbets subreddit was highly correlated with market trends. This suggests that sentiment analysis can be a valuable tool for investors and traders who are interested in monitoring the sentiment of the stock market and making more informed investment decisions. However, it is important to note that sentiment analysis is not a perfect tool and should be used in conjunction with other types of analysis to make investment decisions. Additionally, the study was limited with 800000 posts from r/wallstreetbets subreddit, and further research may be needed to validate the findings in other social media platforms and markets. Overall, the study provides strong evidence that sentiment analysis can be a valuable tool for predicting stock market trends, and that the combination of BERT and VADER is highly effective for conducting sentiment analysis on social media platforms like Reddit.

Bibliography

[1] Sultan Ali AlZaabi. Correlating sentiment in Reddit's wallstreetbets with the stock market using machine learning techniques. 2021. [Accessed on November 4, 2021, at 09:00 AM GMT].

[2] Adam Atkins, Mahesan Niranjan, and Enrico Gerding. Financial news predicts stock market volatility better than close price. The Journal of Finance and Data Science, 4(2):120-137, 2018. [Accessed on November 4, 2021, at 10:30 AM GMT].

[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift Reddit dataset. In Proceedings of the international AAAI conference on web and social media, volume 14, pages 830-839, 2020. [Accessed on November 4, 2021, at 12:00 PM GMT].

[4] Daniel Bradley, Jan Hanousek Jr, Russell Jame, and Zicheng Xiao. Place your bets? The market consequences of investment research on Reddit's wallstreetbets. The Market Consequences of Investment Research on Reddit's Wallstreetbets (March 15, 2021), 2021. [Accessed on November 4, 2021, at 01:30 PM GMT].

[5] Ayman E Khedr, Nagwa Yaseen, et al. Predicting stock market behavior using data mining technique and news sentiment analysis. International Journal of Intelligent Systems and Applications, 9(7):22, 2017. [Accessed on November 4, 2021, at 03:00 PM GMT].

[6] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PloS one, 11(8):e0161197, 2016. [Accessed on November 4, 2021, at 04:30 PM GMT].

[7] James H Lorie, Peter Dodd, and Mary Hamilton Kimpton. The stock market. RD Irwin, 1985. [Accessed on November 5, 2021, at 09:00 AM GMT].

[8] Pavitra Mohanty, Darshan Patel, Parth Patel, and Sudipta Roy. Predicting fluctuations in cryptocurrencies' price using users' comments and real-time prices. In 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pages 477-482. IEEE, 2018. [Accessed on November 5, 2021, at 10:30 AM GMT].

[9] Manuel R Vargas, Beatriz SLP De Lima, and Alexandre G Evsuko. Deep learning for stock market prediction from financial news articles. In 2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA), pages 60-65. IEEE, 2017. [Accessed on November 5, 2021, at 12:00 PM GMT].

[10] Jia Wang, Tong Sun, Benyuan Liu, Yu Cao, and Degang Wang. Financial markets prediction with deep learning. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 97-104. IEEE, 2018. [Accessed on November 5, 2021, at 01:30 PM GMT].

[11] Szafron, Duane & Greiner, Russ & Lu, Paul & Wishart, David & Macdonell, Cam & Anvik, John & Poulin, Brett & lu, Zhiyong & Eisner, Roman & Ca, Eisner@cs. (2003). Explaining naïve Bayes classifications. [Accessed on November 5, 2021, at 03:00 PM GMT].

[12] Das, Meghalee & Tham, Jason. (2022). Tactical Organizing: What Can the r/wallstreetbets and GameStop Frenzy Teach Us About Technical Communication in a Networked Age?. Technical Communication, 69, 36-57. [Accessed on November 5, 2021, at 04:30 PM GMT].

[14] K.A. Althelaya, E.-S.M. El-Alfy, S. Mohammed, Evaluation of bidirectional LSTM for short-and long-term stock market prediction, in: 2018 9th int. Conf. Inf. Commun. Syst., IEEE., 2018, pp. 151-156, [Accessed on November 6, 2021, at 09:00 AM GMT].

[15] J. Zhang, S. Cui, Y. Xu, Q. Li, T. Li, A novel data-driven stock price trend prediction system, Expert Syst Appl 97 (2018) 60-69, [Accessed on November 6, 2021, at 10:30 AM GMT].

[16] M. Hiransha, E.A. Gopalakrishnan, V.K. Menon, K.P. Soman, NSE stock market prediction using deep-learning models, Proc Comput Sci. 132 (2018) 1351-1362, [Accessed on November 6, 2021, at 12:00 PM GMT].

[17] M.M. Alalaya, H.A. Al Rawashdeh, A. Alkhateb, Combination method between fuzzy logic and neural network models to predict Amman stock exchange, Open J Bus Manag 6 (2018) 632-650, [Accessed on November 6, 2021, at 01:30 PM GMT].

[18] M.K. Ahmed, G.M. Wajiga, N.V. Blamah, B. Modi, Stock market forecasting using ant colony optimization-based algorithm, Am J Math Comput Model 4 (2019) 52-57, [Accessed on November 6, 2021, at 03:00 PM GMT].

[19] S. Sahoo, M.N. Mohanty, Stock market price prediction employing artificial neural network optimized by grey wolf optimization, in: New paradig. Decis. Sci. Manag. vol. 1005, Springer, 2020, pp. 77-87, [Accessed on November 6, 2021, at 04:30 PM GMT].

[20] M.H. Adnan, M.M. Isma’eel, Estimating stock returns using rough set theory: an exploratory study with evidence from Iraq stock exchange, J Econ Adm Sci. 27 (2021) 29-39, [Accessed on November 7, 2021, at 09:00 AM GMT].

[21] A. Khalid Chyad, ARIMA model to forecast the ISX60 indicator: an applied study on the Iraqi financial market, Turkish J Comput Math Educ 12 (2021) 2549-2554, [Accessed on November 7, 2021, at 10:30 AM GMT].

[22] E.A. Mohamed, I.E. Ahmed, R. Mehdi, H. Hussain, Impact of corporate performance on stock price predictions in the UAE markets: neuro-fuzzy model, Intell Syst Account Finance Manag 28 (2021) 52-71, [Accessed on November 7, 2021, at 12:00 PM GMT].

[23] Tang, Xiaobin, et al. "Stock Price Prediction Based on Natural Language Processing1." Complexity, May 2022. Hindawi Limited, [Accessed on November 7, 2021, at 01:30 PM GMT].

[24] Carosia, Arthur & Silva, Ana & Coelho, Guilherme. (2022). Using BERT to Predict the Brazilian Stock Market. 10.1007/978-3-031-21689-3_5. [Accessed on November 7, 2021, at 03:00 PM GMT].