University of Southampton

Faculty of Engineering and Physical Sciences

Electronics and Computer Science

Sentiment Analysis for Predicting S&P 500 Movements

by

Hartej Singh Haer

13/09/21

Supervisor: Dr Enrico Gerding

Second Examiner: Dr Igor Golosnoy

A dissertation submitted in partial fulfilment of the degree of MSc Data Science

Abstract

Stock price and market trend forecasting is a challenging issue due to the many variables, both known and unknown, where a time series element adds another layer of complexity. Investors are constantly looking for an edge in the market for improved predictions to maximise profits while academics look to provide evidence for different stock prediction theories.

More recently, with the rise of social media platforms, the potential of sentiment analysis has been explored further as a predictor of stock price movements. Specifically, the aims of this study were to investigate if sentiment analysis can improve stock trend prediction and to compare Reddit and Twitter to the standard benchmark, Financial News, as a source of sentiment analysis.

Furthermore, the purpose of the thesis was to investigate the classification performance of Logistic Regression, Support Vector Machine and Extreme Gradient Boosting in predicting the daily trend movement of the S&P 500, with and without the addition of sentiment analysis from Reddit, Twitter and News as a feature. Moreover, the performance metrics indicated a positive difference between the models' using sentiment and not using sentiment. The average accuracy with sentiment for News, Twitter, and Reddit, across all three models, are 55.2%, 63% and 62.2%, respectively. The Mathews Correlation Coefficient reported that News attains a score of 0.122, Twitter at 0.063 and Reddit at 0.125. With sentiment applied, the error rate improved by 16% for News, 21% for Twitter and 13% for Reddit.

Overall, the work provided evidence that sentiment is a viable and valuable variable for stock prediction. Reddit and Twitter have shown to have an advantage over News as a source of textual data. However, the work comparing Reddit vs. Twitter was not conclusive, but Reddit seemed to have an edge in this study due to its unrestrictive word length. The work enhanced the field of stock prediction and gave evidence to disprove that the stock market can only be predicted at random accuracy.

Statement of Originality

- I have read and understood the ECS Academic Integrity information and the University's Academic Integrity Guidance for Students.
- I am aware that failure to act in accordance with the Regulations Governing Academic Integrity may lead to the imposition of penalties which, for the most serious cases, may include termination of programme.
- I consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

You must change the statements in the boxes if you do not agree with them.

We expect you to acknowledge all sources of information (e.g. ideas, algorithms, data) using citations. You must also put quotation marks around any sections of text that you have copied without paraphrasing. If any figures or tables have been taken or modified from another source, you must explain this in the caption and cite the original source.

I have acknowledged all sources, and identified any content taken from elsewhere.

If you have used any code (e.g. open-source code), reference designs, or similar resources that have been produced by anyone else, you must list them in the box below. In the report, you must explain what was used and how it relates to the work you have done.

I have not used any resources produced by anyone else.

You can consult with module teaching staff/demonstrators, but you should not show anyone else your work (this includes uploading your work to publicly-accessible repositories e.g. Github, unless expressly permitted by the module leader), or help them to do theirs. For individual assignments, we expect you to work on your own. For group assignments, we expect that you work only with your allocated group. You must get permission in writing from the module teaching staff before you seek outside assistance, e.g. a proofreading service, and declare it here.

I did all the work myself, or with my allocated group, and have not helped anyone else.

We expect that you have not fabricated, modified or distorted any data, evidence, references, experimental results, or other material used or presented in the report. You must clearly describe your experiments and how the results were obtained, and include all data, source code and/or designs (either in the report, or submitted as a separate file) so that your results could be reproduced.

The material in the report is genuine, and I have included all my data/code/designs.

We expect that you have not previously submitted any part of this work for another assessment. You must get permission in writing from the module teaching staff before re-using any of your previously submitted work for this assessment.

I have not submitted any part of this work for another assessment.

If your work involved research/studies (including surveys) on human participants, their cells or data, or on animals, you must have been granted ethical approval before the work was carried out, and any experiments must have followed these requirements. You must give details of this in the report, and list the ethical approval reference number(s) in the box below.

My work did not involve human participants, their cells or data, or animals.

ECS Statement of Originality Template, updated August 2018, Alex Weddell aiofficer@ecs.soton.ac.uk

Acknowledgements

I would like to express my appreciation to all the people that helped me this year. Most notably my supervisor, Dr Enrico Gerding for his insightful comments and guidance throughout this project. I would also like to gratefully thank Dr Igor Golosnoy who gave valuable feedback based upon my demonstration. Special thanks to the department of Electronics and Computer Science at Southampton University who helped me when I fell ill with COVID. Finally, I would like to express my deepest gratitude to my family and friends who supported me through this thesis.

Table of Contents

Chapter 1

Introduction

1.1 Problem Statement

The Efficient Market Hypothesis (EMH) suggests financial market trends reflect news, events, and key information. These variables are unpredictable in nature and cannot be timed, implying that stock prices are difficult to predict with more than 50% accuracy as they follow a random walk pattern [1]. Expanding on this theory, all investors are assumed to have the same information, and this information is reflected in the price the moment it is known meaning investors cannot beat the market and there is no incentive to predict prices [2].

On the other hand, some studies suggest that the stock market does not follow the EMH and can be predicted to some degree [3] and [4].

An opposing claim is the behavioural finance theory, shown in [5], which states that investors are irrational, and their psychology can affect the market causing them to overreact to key information, whether it is overly optimistic or pessimistic, regarding a stock's valuation. It theorises that markets are driven by human emotion, such as fear and greed, which supersedes investor logic [6]. Investors have access to different levels of information and they react to this information in various ways; reflected in stock prices. This theory highlights which emotions can be a factor when deciding a stock's price, and consequently market sentiment needs to be studied in further depth for more accurate predictions.

1.2 Motivation

Investors and hedge fund managers are constantly trying to refine their edge and beat the market by making predictions of stocks, which is a significant area of interest for them. Stock price and market trend forecasting is a challenging issue due to the many variables involved, both known and unknown, for example, volume, prices, and sentiment etc. Financial trends are a time series problem, adding another layer of complexity to predicting market conditions which are in a constant flux [7]. Accurate prediction of assets would lead to huge opportunities for profit, whereas a poor prediction can incur major losses, leading the motivation for this area of work [8].

If investors and hedge fund managers can accurately forecast stock price movement, they can capitalise on this by selling or buying this stock, gaining profit. Funds are already being developed to include sentiment from online sources, such as the recently released Buzz Exchange Traded Fund (ETF) which uses a proprietary AI algorithm to pick stocks based upon positive investor sentiment online. Furthermore, Algorithmic Trading and Quantitative Trading [9], [10] are being applied to real stock market situations to generate a profit.

1.3 Social Media and its' Relationship With Stocks

Previous research has indicated two main areas of analysis that can help investors and fund managers predict stock price movement: Technical Analysis and Fundamental Analysis. Technical analysis uses patterns and tools to analyse charts to predict a stock's price. While some research has shown this approach to work [11], the results were mediocre as it only focuses on structured data. Fundamental analysis uses economic and financial factors, such as earnings and reports to predict stock prices. This approach incorporates sentiment analysis.
More recently, with the rise of social media platforms, the potential of sentiment analysis has been explored further as a predictor of stock price movements. Sentiment analysis can be described as the public's perception or mood towards a particular stock or company. It classifies the polarity of the text into positive, negative, or neutral. The aggregated opinion of a stock is a significant variable that is becoming

more influential as a market driver, with the increased use of social media platforms. Analysing textual data from social media platforms and financial news reflects investors' psychology, and intuitively could increase the predictive power of analytical models for stock price prediction. Previous research proves that sentiment is highly correlated with the price of a stock [12] and [13]. This thesis aims to extend the growing body of work in this subject area.

1.4 **Aims**

This research focuses on two aims:

1. To investigate if sentiment analysis can improve stock trend prediction.
2. To compare and analyse Reddit and Twitter data to the standard benchmark, Financial News, as a source of sentiment analysis.

The research questions led to the following hypothesis: implementing sentiment analysis can predict the S&P 500 trend with a higher than random chance.

The purpose of this thesis is to investigate the classification performance of three different machine learning algorithms in predicting the daily trend movement of the US stock market index, Standard and Poor's 500 (S&P 500), with and without the addition of sentiment analysis from Reddit, Twitter, and Financial News as a feature.

The problem is expressed as a supervised learning binary problem, where a value of +1 represents bullish market behaviour (stock price increases) and a value of 0 which represents bearish market behaviour (stock price decreases or remains consistent) representing stock trend. As shown in [2] and [8], prediction of the actual price of a stock is less meaningful than the stock's direction because investors are better able to anticipate the market and generate a profit from the latter.

Once the data was collected, sentiments were extracted from the sources using VADER, and machine learning algorithms were applied to the problem, with and without sentiment, for each data source. The models chosen for prediction were Logistic Regression (LR), Support Vector Machine (SVM) and Extreme Gradient Boost (XGB). The results from these models were evaluated.

1.5 **Contribution**

The majority of studies in this area focus on one specific source of sentiment, either Twitter, Reddit or Financial News. This thesis contributes to the existing literature by investigating which of these three sources of sentiment are the better predictor for stock trends. The comparison of these sources has not been studied extensively in a machine learning stock context where all are viable sources of sentiment analysis.

The rest of this work is organised as follows. Chapter 2 reviews the current field of work for stock prediction with sentiment analysis. Chapter 3 covers the research methodology including pre-processing, sentiment analysis and models for stock prediction. The results are summarised in Chapter 4 while they are discussed in accordance with the aims of this thesis in Chapter 5. Chapter 6 evaluates the limitations of the research. The research is concluded with suggestions of future work in Chapter 7. Chapter 8 discusses how the project was managed and the problems that were overcome during the thesis.

Chapter 2

Literature Review

This chapter reviews the recent and key literature in the field of stock market prediction with sentiment analysis. A brief overview of sentiment analysis is discussed before moving on to the literature of stock prediction with both news and social media text data. Different algorithms and findings are collated to give an overview of the surrounding work; approving or disapproving the theories aforementioned in the Introductory chapter.

2.1 Sentiment Analysis

Yu et al. [14] uses sentiment from online stock market news articles to show expanded emotion words improve classification performance. The authors aim to build upon the technique of using pre-defined sentiment lexicons or dictionaries, such as "SentiWordNet" and "Harvard-IV-4", by creating a financial specific lexicon to improve performance.

This approach is improved by Sohangir et al. [15] who compares different lexicon-based approaches with different machine learning analysers. The models used were VADER, SentiWordNet, TextBlob, LR, SVM and Naive Bayes (NB) on StockTwits data, to investigate if they increase the accuracy of sentiment analysis. The best performing and fastest approach was VADER at 94.4% accuracy; a substantial improvement on the machine learning models. The authors state that the negatives of using machine learning based sentiment analysis, with high dimensional data, is that the training time is time consuming and computationally expensive.

Additionally, [16] further refines the VADER lexicon by adding new words with relevant sentiment to make the algorithm more finance specific for new articles from Finwiz. The authors observed that a positive change in sentiment correlated with a rise in the Tesla stock price and vice versa for a negative change in sentiment, highlighting which investors can use sentiment to improve investment decisions.

Similar methods were used by Lubitz [2] but with two textual datasets, Financial News and Reddit posts. The text datasets used for this research were from January 2008 and July 2017, from the S&P 500 index. Both dictionary-based and machine learning sentiment models were compared with a range of pre-processing and feature extraction techniques. Stemming was found to have no overall impact on the results, while weighted functions enhanced classifier accuracy. More text was classified as negative than positive, possibly because news outlets and social media platforms are more negatively biased due to the psychological behaviour of humans. The results suggested that the predictive power of Reddit is slightly better than using financial news at 56% accuracy. The work showcases Reddit as a potential source of sentiment analysis for predicting stock index movements.

2.2 Stock Prediction

Li. et al. [17] focuses on projecting news articles from a major Hong Kong vendor on to the sentiment space using the Harvard Psychological Dictionary and Loughran–McDonald Financial Sentiment Dictionary. Their work indicates that the two dictionaries can be used effectively for a market forecast task. While the results show a strong correlation between same day news and the market, they find sentiment does not provide useful predictions. This may be due to their poor choice of measuring stock trend where they find the percentage change between the opening and closing price on a trading day. The opening price can be subject to change by news releases, just before the market opens and pre-market and after-market trading.

In [18], this problem is rectified by finding the difference between today's close price and yesterdays, which accounts for the impact of news releases during market hours. The research uses TextBlob to identify sentiment from the news provider Techmeme for the Google stock. However, it illustrates a negative relationship between Google stock returns and sentiment by -4.3%, using a decision tree classifier. Since

this paper only utilises one stock it indicates sentiment may be better at predicting indexes as it considers a broader public sentiment. This work shows that prediction is worse than the 50% random walk.

Furthermore, [19] reinforces this view as the sentiment features are not statistically significant with stock price. Reddit platform is used as the textual data with VADER, as the sentiment analyser. However, the authors construct a profitable portfolio trading strategy based upon sentiment analysis of Reddit for predictions. They conclude that Reddit text data provides prediction power to models and can help investors generate a profit in the stock market.

The work in [20] finds that the EMH hypothesis can be disapproved. The study analyses the KOSPI (Korea Composite Stock Price Index), in addition to using an opinion mining sentiment dictionary for economic news to predict stock trend. The accuracy results range from 60% to 65%. The authors state accuracy scores depend on the news media used as they have their own characteristics. Only one year's worth of news data was used, limiting the reliability of the results. Additionally, Rahman et al. [21] aims to extract sentiment from financial news for five blue chip companies on the Malaysian stock exchange to classify stock trend. 15,000 news articles were used where the data was cleaned and pre-processed including tokenisation, stemming and stopword removal. The SVM model attained an average accuracy of 56% which is above the 50% random walk and was reported to have helped investors understand the probabilities of stock price movements. Similarly, [22] shows an increase in performance of up to 73% using an SVM model with similar pre-processing, using a time series approach with news data from May 2014 to April 2015 concerning five stocks from the VN30 index. This paper may achieve a higher accuracy than the former as a novel delta-TFIDF approach was used, increasing the importance of words with uneven distribution between positive and negative classes. The results find that news quality significantly impacts stock trend prediction and SVM performs poorly with few inputs. The authors highlight future work needs to incorporate technical analysis and numeric features alongside sentiment analysis to achieve even better performance.

Building upon this work, both [8] and [23] show the benefits of numeric features for achieving the highest accuracy scores. The work in [23] constructs a prediction model using a multitude of news sources. The sentiment of news is classified by a NB model resulting in a range of accuracy predictions of 73% to 86%. The authors use a k-nearest neighbours (KNN) classifier on three companies from the NASDAQ index. The original numerical data included opening, closing, high and low. Different techniques were used on the numeric features such as transforming the original values into discrete values to be positive, negative, or equal, based upon the difference with the previous day's closing price. This study finds that considering only news sentiment analysis produces accuracies from 59%-63%. Conversely, prediction accuracy is improved up to 89.80% for stock trend prediction when considering the numeric features. The paper not only shows that there is a strong relationship between stock news and changes in stock prices, but also the increase in accuracy when using numeric data. The work in [8] takes a different approach by using 10 technical indicators based on numeric data alongside news sentiment analysis to enhance stock prediction models. A dataset of 3,500 samples is collated from between 2006 and 2020 concerning the OMS30, the Swedish stock market index. Financial news was determined to be the best source of sentiment in this field, as social media data can be unreliable and biased. While VADER was considered as the sentiment analyser, the author chose to manually label the dataset as VADER but this does not work in the Swedish language. The author uses XGBoost as the classifier due to its reported popularity in Kaggle competitions. The results indicate that XGB has a good performance on this type of problem where 64% accuracy was achieved using only sentiment as a feature, and 73% was achieved using both sentiment and numeric features. From these results the author confirms that numerical features had a bigger impact on improving performance rather than sentiment features alone. The study mentions different datasets are needed for sentiment analysis including social media data to further boost predictive performance.

In contrast, [24] has an opposing view on numeric data and sentiment when predicting the trend of Amazon stock price. The authors state that only using numeric features results in a similar performance to a random walk model, around 50%, but when combined with sentiment analysis, performance is boosted by 10% to 61.2%, showing that choosing the right features is a key determinant in this area of research.

Bouktif et al. [24] finds sentiment counts, N-grams, feature lags, and polarity could be good predictors from Twitter data when using five supervised learning algorithms. The work finds XGB and Random Forest (RF) have the best scores for both accuracy and F1-score. The study supports the behavioural finance theory where investors' sentiment can affect the market through their psychology.

Another approach approving this theory is by Gupta et al. [25] who investigates the impact of sentiment analysis through StockTwits on stock price prediction. The text data is pre-processed, and features are extracted before implementing machine learning classifiers for sentiment. The text data is aggregated into daily sentiment with only positive and negative sentiment. Five companies were examined from the S&P 500 for 9 months' worth of stock price data. The paper finds that there is a high correlation between sentiment and stock price change in the same day, but this relationship declines as days pass. Moreover, a time series model is built which demonstrates that sentiment analysis positively impacts accuracy in all 5 companies up to 65.6%. The addition of sentiment causes a 0.6%-3.3% increase in accuracy where it is concluded that more tweets lead to higher confidence in sentiment signal having more impact on the overall accuracy. The limitation of this study is that the prediction model is constructed based upon data from the previous 5 days. A similar approach was taken by [26] which shows better results than the previous work, with accuracy scores of 69% for LR and 72% for SVM. The model uses aggregate sentiment values for a 3-day period and the dataset was 12 months long. The work finds a strong correlation between Twitter sentiment and stock price movement. The authors in [25] and [26] deduce that their work gives a significant advantage to investors as their models are considerably above the 50% random walk suggested by the EMH hypothesis.

Chapter 3

Methodology

The proposed methodology can be split into three parts. The first part involves collecting, cleaning, and pre-processing the data. The second part involves sentiment analysis and labelling the data. Lastly, the third part involves stock trend prediction using various machine learning models.

The data for Reddit, Twitter and Financial news was collected online while the financial data was extracted from Yahoo Finance. Using datasets found online was deemed more effective than retrieving the data myself, as scraping through data online is time consuming and word count restrictions for certain sources makes it complex to gather an adequate dataset in the timeframe for this thesis. The datasets were cleaned and pre-processed to remove undesirable data and to structure the text data. The stock price data was transformed into a supervised learning problem, and Vader was used to find the sentiment for each day only including positive and negative sentiment with values of '+1' and '-1' respectively. Using XGB, LR, and SVM a time series analysis of stock trend prediction was carried out for each data source with and without sentiment. The work was coded using Python where the main libraries included numpy, nltk, pandas, matplotlib, sklearn, and Vader Sentiment.

This approach aims to evaluate the predictive relationship between sentiment analysis and stock price movements, and to compare Reddit and Twitter to Financial News as a source of sentiment analysis.

The chapter is divided into the following subsections data, pre-processing of both textual and numeric data, sentiment analysis, stock prediction and performance criteria.

3.1 Data

3.1.1 Reddit

Reddit is a forum website where users can share news, content, and interact with each other. The content is segregated into mini communities known as 'subreddits', these allow users to read and contribute to specific subjects such as movies, stocks, cryptocurrency, and food. By the end of 2020, Reddit had recorded an average of 52 million users a day, [27] many of whom give their opinions on stocks on a daily basis, making Reddit a good source of unstructured text data. The work in [19] highlights Reddit's correlation with future stock price changes, giving credibility of its use as a data source in this thesis.

There are several benefits with Reddit compared to other companies such as Facebook and Twitter. The content is already filtered by subreddits making it easy to find appropriate information. As shown by Buntain and Golbeck [28] most users only interact with one subreddit which increases the value of the information as active users are more knowledgeable about their respective topic. Furthermore, the text data itself contains less spam and jargon as moderators for each subreddit ensure only relevant content is displayed. This not only makes sorting the data easier but also increases the reliability of the data. Nevertheless, collating the title, description and replies for a single post is very difficult, so much of the text scraped, is broken and lacks structure. It is therefore, hard for machine algorithms to understand the wider context of the textual data [19].

Leukipp provided the Reddit dataset on Kaggle [29] The data was scraped from Reddit using the Reddit API, PushShift API and Python reddit API wrapper with the task of fetching posts. The data includes 14 subreddits but only 4 were chosen due to their relevance with stock trend prediction. These subreddits were r/stocks, r/stockmarket, r/investing, r/finance and were filtered by using the key word 'S&P 500'. The datasets were aggregated into a single CSV (Comma-Separated Value) file with only date and time selected. The time period of the data was from 1/1/21 - 5/8/21 and in total contained 991 posts. By removing duplicates and null values the size of the dataset was reduced to 703 unique posts.

3.1.2 **Twitter**

Twitter is a social networking service where users interact and post their thoughts through tweets, comments, likes and retweets [30]. By the end of 2020, Twitter boasted 192 million daily active users [31]. Moreover, Twitter is widely accepted in the financial prediction community as illustrated in more detail in the following research [25] and [26].

Twitter has been considered the main source of social media sentiment analysis for stock prediction due to the vast quantity of information which can be filtered by topic, and through hashtags found on the site. Many companies interact with users to aid public relations, in turn generating more unstructured information about the company as more users will engage and post their thoughts. This builds up a network of data which can be aggregated to find the general sentiment about finance topics [32]. However, the quality of posts may be lower than Reddit, as there is no moderator to keep posts on topic, meaning much of the information may contain little value for analysis [33].

The Twitter dataset was provided by Bruno Taborda on iee-dataport [34]. The data was extracted from Twitter using the Twitter REST API search, known as Tweepy which allows the last 7 days to, filtered by Twitter tags and English. The CSV file downloaded was split into labelled and unlabelled data where the unlabelled data was disregarded as it contained too many posts compared to the other data sources. Sentiment was removed from the labelled dataset and the file was further filtered by using the key word 'S&P 500' so only the relevant information was available for analysis. The time period of the data was from 9/4/20 - 16/7/20 and in total contained 5000 posts. By removing duplicates and null values the size of the dataset was reduced to 4844 unique posts.

3.1.3 **Financial News**

The baseline data source for this study is Financial News, specifically articles from CNBC as it is an important source of information for many stock investors. CNBC is an American business news company where in 2020 the channel reached a record 84.9 million average monthly unique visitors [35]. The majority of stock prediction research has used sentiment analysis of news as a feature within their study as it is perceived to be more consistent and trustworthy than social media data. The work in [22] and [23] display news sentiments correlation with stock price.

Analysing sentiment from a news source such as CNBC is a useful contrast compared to sentiment analysis from social media data. The language in news articles is written in a more formal style, as no emojis, abbreviations or slang are used. Therefore, Financial News sources have less noise making it easier for text analysis. The articles' content can be deemed richer in value as the writers must back up points with reliable evidence, which is later checked by editors giving more credibility. However, text data from news sources can be overexaggerated as their primary role is to entertain the reader so the content may be biased.

The news dataset was provided by Lucas Pham on Kaggle [36]. The CNBC news dataset was selected as it contained information exclusively about the S&P 500. The text data was extracted from online articles using the package BeautifulSoup in Python. The data selected from the original dataset was the last updated data in the preview text of articles. The preview of articles contain more information and are less polarising than headlines, so was subsequently deemed as better quality for a sentiment analysis task. The time period of the data was from 17/12/17 - 19/7/20 and in total contained 3081 posts. By removing duplicates and null values the size of the dataset was reduced to 2619 unique posts.

3.1.4 **Stock Price Data**

The stock price data was extracted from Yahoo Finance [37] which is part of the Yahoo network that provides content exclusively about stocks, financial reports and earnings.

Specifically, the S&P 500 index was selected, representing the top 500 large-cap companies in America and can be considered the benchmark for investors. By the end of 2020 $4.6 trillion was invested into the index highlighting it as a core component in many investors' portfolios [38]. This index was chosen due to America's strong position in the finance world, and its exposure to many sectors giving a broader view of the market. However, the top ten largest businesses account for 26.4% of the market capitalization value, indicating that it is overly weighted towards bigger companies and hence may not be as representative of the overall market as initially assumed [38].

The data obtained from Yahoo Finance, where from three different time periods to correlate with the individual datasets above, 1/1/21 - 5/8/21, 9/4/20 - 16/7/20, and 17/12/17 - 19/7/20. The data included the date, open, close, high, low, adjusted close and volume for each day. To accomplish consistency for the machine learning models used for prediction, the close prices and volume of each day were used; in contrast to most literature which used the adjusted close price such as the work in [39]. The size of the data relates to the number of days extracted, where the News dataset included 644 days, 145 days for the Reddit dataset and 67 days for the Twitter dataset. While there is some discrepancy in the number of days between the datasets, limiting the reliability of future results; there is an overlap in terms of time period, volume of posts only data related to the S&P 500 index was extracted.

3.1.5 **Exploratory Data Analysis (EDA)**

In this section, text data is analysed focusing on the length of the text and words with low discriminatory power. This is a fundamental step in the methodology since it helps us to comprehend the data and apply algorithms in subsequent stages that will perform best, based on what is learnt.

3.1.5.1 **Length of Text**

Figure 1 shows histogram plots for text length of difference sources.

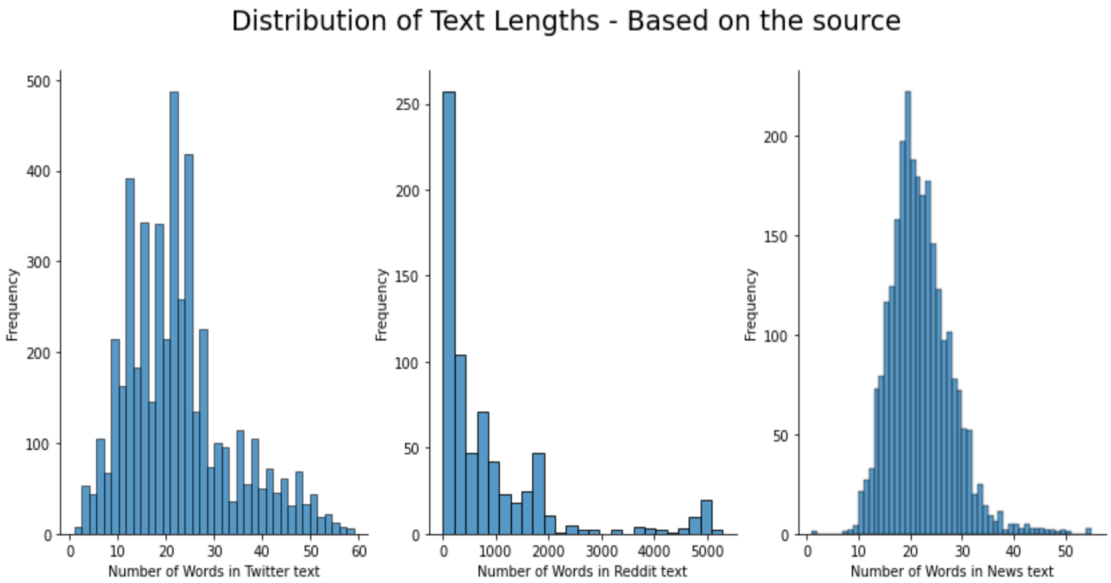

Figure 1: Histogram of source text lengths. The figure shows the break-down of text length for each data source. All data sources are biased to shorter text lengths with an average of 96 words. Reddit has highest number of words on average, Twitter and News are almost identical in terms of average number of words.

The data was broken-down by word length to reveal under or over represented data indicating any bias between data sources. Figure 1 shows a significant bias towards shorter text in all data sources where News has the shortest text on average at 21.55, Twitter has 22.5 words and Reddit has the longest at 881.4 words. Between the social media platforms, Twitter was expected to have less words, as it has a limit of 280 characters for tweets while there is no such limit for Reddit. Twitter can be seen as the type of platform where users post short text about their ideas or thoughts. Synopsis of news articles are slightly longer on average whilst still being succinct, whereas Reddit has longer more detailed posts due to its' more forum-centred nature. Longer text length could play to Reddit's advantage in this thesis, as the sentiment analysis is more likely to be correct with a larger sample of words and therefore the machine learning algorithms can infer patterns [40].

3.1.5.2 **Important Words**

Word clouds were created by source to obtain insights into the datasets, as seen in Figure 2. Word clouds give greater importance to higher frequency words which are shown to be larger than less frequent words. They were used to identify new stop words that were specifically applicable to this thesis.

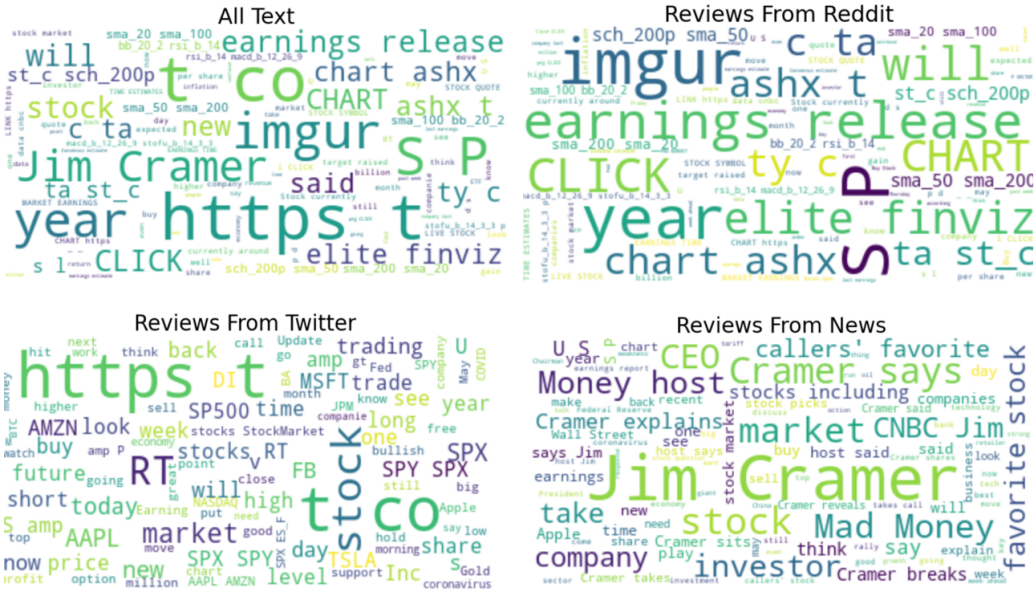

Figure 2: Word clouds of sources. The more frequent occurring words can be seen in the word clouds which are split by data source. This helps to identify low discriminatory words which can be added to the stop word list to improve machine learning prediction.

Words with low discriminatory power were identified, such as 'Jim Cramer', 'This', 'Inc', 'imgur' among others, where the process of handling them will be discussed in the Pre-processing section. Words such as 'buy', 'short', 'break' etc have a high frequency for all datasets and can be considered differentials to help identify the sentiment of text.

3.2 **Pre-processing**

3.2.1 **Textual Data**

Pre-processing is the removal of noise from data in order to improve its value. Once the pre-processing procedures have been completed, a machine learning model can only identify the most significant features from the dataset; else, the model would train on the noise rather than the core of the text, resulting in an overfitting model. Social media text consists almost entirely of unstructured data and so cannot be used as an input to a classifier without prior cleaning and manipulation.

Null values and duplicates were dropped at the first stage of cleaning the Reddit, Twitter and News datasets. Duplicates were removed to prevent the model from overfitting on same data, which would cause poor generalisation on new information. Certain special characters were removed, such as @, hashtags, RT and hyperlinks through the Python library re, to simplify the textual data for machine learning. For example, #S&P500 was replaced with S&P500. The final cleaning process was merging the text data by date for each dataset.

The data was pre-processed in the following steps:

1. Stop word removal entails dropping all non-discriminatory and insignificant terms from the text. Words such as ‘a’, ‘the’, ‘in’ etc are examples of non-sentimental words. The NLTK stopword list was used and extended based upon the low discriminatory words found in the previous EDA section. The following words were added to the list; 'one', 'make’, ‘get', 'go', 'like', 'This', 'today', 'imgur', 'in', 'see', 'also', 'would', 'think', 'come’, ‘say', 'look', 'could', 'back', 'to', 'for', 'RT', 'Jim', 'Cramer', 'says', 'The', 'even', 'group', 'say', 'look', 'could', 'back', 'CLICK', 'FOR', 'HERE'. This list was built in tandem with the work from [41] which demonstrated the positive effects of their own stop word list. These words appeared more frequently in all datasets, and it was found that their existence improved the performance of VADER analyser by up to 3%. Moreover, words such as ‘stocks’ and ‘analyst’ were not removed despite their frequent appearance because they provided context to the text.

2. Lemmatization is the process of restoring a word to its base form, which is known as a "lemma." For example, the term "playing" will be replaced with "play”. This step was applied to reduce the size of the features and preserve the discriminative ones. It was implemented using NLTK, WordNetLemmatizer, and was compared to stemming, which also returns a word to its root. Stemming was found to unnecessarily introduce more representations of one word. While preliminary testing results showed that lemmatisation and stemming resulted in the same scores, it was chosen because it is more consistent and achieves the correct form of the vocabulary [42].

Tokenisation, case normalisation, and expanding contractions were considered but were rejected due to initial testing results falling when applied.

3.2.2 **Numeric data**

The stock price dataset included close price and volume. To predict the daily stock trend, the problem was transformed into a supervised learning, classification problem with the output changing from price to stock direction. As first mentioned in the Introduction chapter the actual price of a stock is less meaningful than the direction of the stock. The work in [43] shows that trends for time series are more interesting to study than exact price outputs, as investors follow trends more closely for investment decisions. This view is furthered by [44], which states that trading methods based on classification models generate higher risk-adjusted returns than regression models.

A new column was created which was the daily stock difference where y = close price today - close price yesterday. These values were then converted into a binary output where a value of +1 represents bullish market behaviour (stock price increases) and a value of 0 represents bearish market behaviour (stock price decreases or stays the same) representing stock trend. In addition, to measure the success of the prediction, the target values needed to be converted into binary outputs which were the following day's trend value.

Stock price data is not available on weekends therefore textual data on these days were removed. It is rare for new market information to be released on weekends so the data would have little impact on stock trend. Finally, the data was scaled using StandardScalar which centres the mean around 0 and scales each variable to unit variance helping to normalise the data within a set range.

3.3 **Sentiment Analysis**

The aim of sentiment analysis is to transform textual data into quantitative data that corresponds to how the author felt about a particular subject. This can be used as an input into machine learning algorithms.

Valence Aware Dictionary for Sentiment Analysis (VADER) is a 'pre-built' lexicon and rule-based sentiment analysis library that can output the polarity depicted in a text passage. Each word in the passage is given a score ranging from -4.0 for very negative sentiment to +4.0 for very positive sentiment. Four scores are produced positive, neutral, negative, and compound. The compound score is calculated by averaging the sentiment ratings for each word in the sentence, and then normalising the result in the range between -1 and 1. For this thesis, only the compound score was used where a score greater or equal than 0 represents a positive sentiment and a score less than 0 represents negative sentiment. Neutral sentiments were not used to minimise bias to the centre as much of the text would have no value for prediction. The compound score was appended to the data along with +1 or -1 depending on if the sentiment was positive or negative. The dictionary was self-developed and implemented by the VADER Python library.

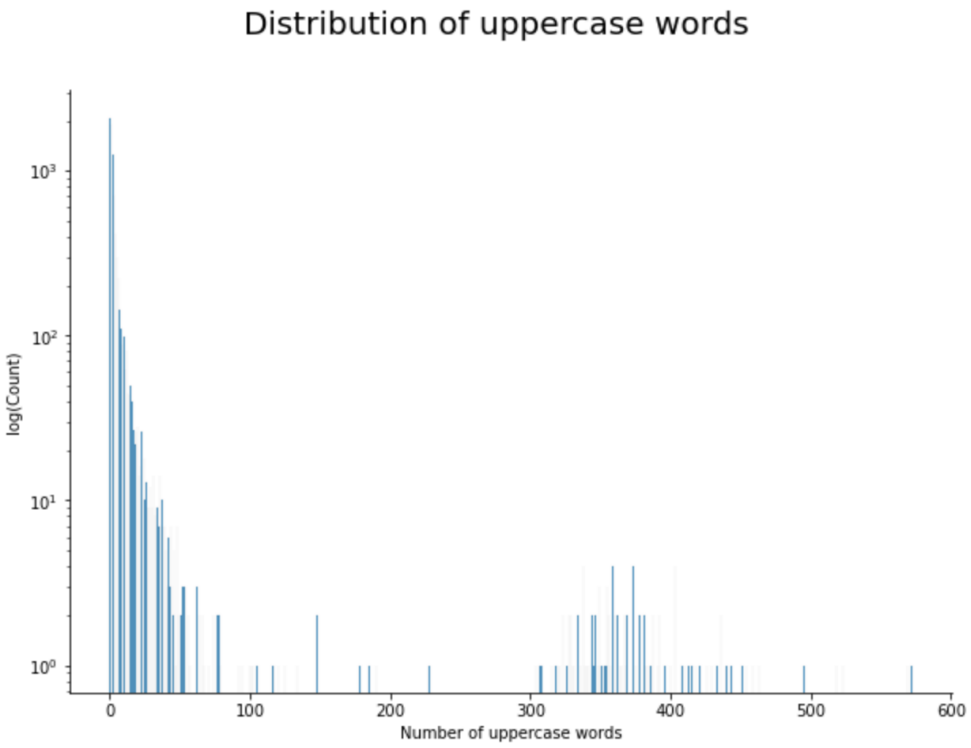

Figure 3: Histogram of uppercase words from all data sources. *The graph shows that uppercase words are prevalent throughout all data sources with an average of 9.1.*

Once the text is analysed VADER weighs the sentiment based on 5 criteria. One is punctuation where exclamation marks increase the magnitude of sentiment and careful care was taken not to remove these in the cleaning process. Similar to exclamation marks, capitalisation of words increases the intensity of the sentiment and as shown by Figure 3, upper casing was prolific throughout the text data. Thirdly, conjunctions such as 'but' cause a change in polarity with more emphasis placed on the words after than those prior. Degree modifiers such as 'extremely' and 'marginally' add weight to the sentiment, increasing or decreasing the polarity, respectively. Finally, tri-grams are used by VADER due to negation which flips the polarity of a sentence [19]. This is useful when the word 'not' is used, e.g. 'The S&P500 is not a good investment today'. Figure 4 shows a breakdown of the criteria and how polarity is calculated.

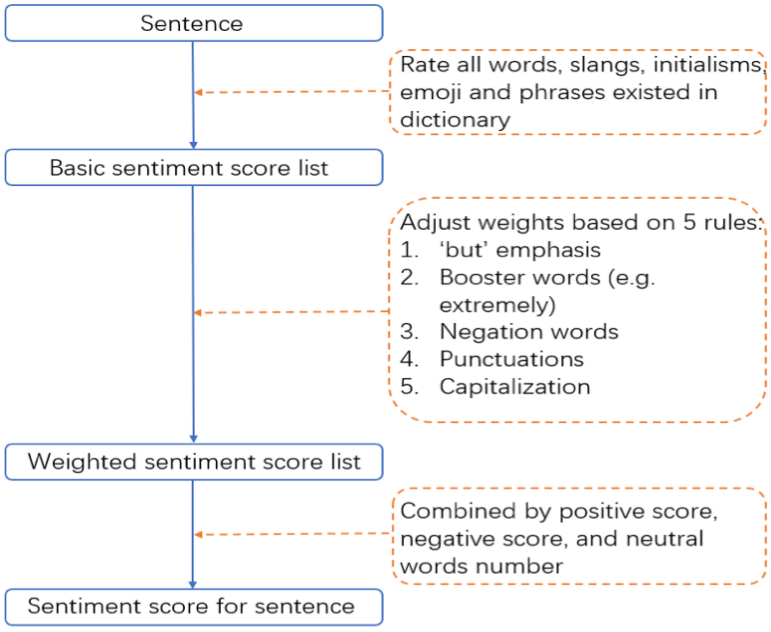

Figure 4: Process of VADER sentiment package [19]. The figure shows how the VADER algorithm determines polarity.

VADER was chosen as it has been used well in numerous studies similar to this thesis [15] and [45]. Social media text is plagued by emojis and slang because of the word count limit shown by [45]. Not only is this factored in with VADER, but profanity is also handled well [19]. Additional benefits of VADER include not suffering from a speed-performance trade off and requiring no training data in order to learn.

TextBlob was considered as an alternative to VADER based upon the work in [19]. TextBlob is an NLP library based on Python and is used for sentiment analysis. A manually labelled Twitter dataset containing 5791 posts from Kaggle [46] was used to compare sentiment accuracy from VADER and Texblob. VADER achieved a higher accuracy score by 0.91% and overall was determined to be the better sentiment analyser for this thesis.

A finance lexicon, NTUSD-Fin [47], was updated to VADER to make it more finance specific, catering more towards this thesis' subject area. Furthermore, previous research in [16], shows adding additional words to VADER improves sentiment accuracy. Finance specific words were chosen based upon previous literature such as 'bullish', 'bearish', 'crushes', 'misses' and 'moon'. However, once implemented on the labelled Twitter data, sentiment accuracy fell, compared to the base VADER analyser and therefore the aforementioned words and lexicon were removed.

3.4 **Stock Prediction**

After sentiment analysis, machine learning models were built to predict the trend of the S&P 500. The three machine learning models are LR, SVM and XGB and were implemented using sklearn and xgboost libraries. The models were tuned using GridSearch which is an exhaustive search used to find the optimal parameters for the models to achieve the best performance.

The datasets were divided into a 75:25 split between training and testing. The training data was split using a rolling origin approach similar to the work in [8] to cross validate the data. A rolling origin approach uses prior training data to predict the following validation data then adds the validation data to the training data, and so on until training is complete. This style of approach is needed for a time series analysis because the sequential order of values must be adhered to, to create a robust model [48]. The training data is split into 5 smaller subsets with equivalent evaluation data.

Most of the previous literature uses past stock time series data to forecast future stock prices [25]. This study's final input features are close price, daily sentiment, compound VADER score, trend direction and volume. The output is a binary value of 0 or 1 predicting the next day's trend direction. Numerical features and technical indicators are not considered in this work as the focus is on how sentiment affects predictive performance of machine learning algorithms.

3.4.1 **Models**

3.4.1.1 **Logistic Regression**

Logistic regression is employed as a performance baseline model since it is powerful yet efficient for prediction. LR is a statistical approach which has been shown to work well on similar studies; paper [49] shows LR achieved an accuracy of 70% with input features of daily stock price, news sentiment and Twitter sentiment. LR is useful as it directly shows how the presence or absence of a variable affects the outcome, in this case sentiment on stock trend prediction [50]. The algorithm works well on complex nonlinear data, is quick to implement and acts as a good baseline for more complex algorithms. Compared to SVM, Logistic Regression is simpler and more easily implemented [51].

However, LR is sensitive to outliers and is prone to overfitting on high dimensional datasets. The algorithm struggles to capture complex relationships which is not ideal for predicting the stock market as it is characterised by noise and uncertainty [49].

Logistic regression is an extension of linear regression for classification problems. LR is used to estimate the probability that a data point belongs to a certain class. In this case, if the estimated probability is greater than 50% then the model predicts that the point belongs to the class 1 and if less than 50% the instance belongs to 0. The model computes the weighted sum of the input features with a bias term [52]. The outcome of logistic regression is a function that describes how the probability of the event (0 or 1) varies with the predictors by applying the sigmoid function.

3.4.1.2 **Support Vector Machine**

Previous research has shown SVM to be a popular choice for text classification with very good predictive accuracy ability [22]. The study of [22] showcases the high predictive accuracy of SVM with a resulting accuracy of 73% on the VN30 index, with news sentiment. This is significantly above the 50% threshold and partly explains why many researchers have used this machine learning model. SVM was used instead of neural networks because it can find the optimal global solution instead of only the local optimum due to the algorithm solving a linearly constrained quadratic programming problem [53]. Additionally, SVM works well on classification problems with high dimensions and is resistant to overfitting, unlike LR.

However, this algorithm is sensitive to parameter selection which is not always easy for large datasets. The model can be difficult to fine tune which is detrimental in stock market prediction as it is a highly sensitive market.

SVM is a machine learning algorithm whose objective is to find a maximum margin hyperplane that maximally separates two classes of data. Furthermore, SVM plots the data points in an n-dimensional space (where n is the number of features) before plotting a hyperplane which best divides the two classification classes.

3.4.1.3 **Extreme Gradient Boosting**

The last model used for stock trend prediction was Extreme Gradient Boosting
(XGB) Classifier due to the algorithm winning many competitions for supervising learning problems on Kaggle [54]. Previous research [8] uses XGB to predict the daily trend of the OMXS30 index with sentiment

analysis, accomplishing an accuracy of 73.71%. Since the work shown in [8] is a binary classification problem and the data uses is akin to our own, a similar performance is expected. The advantages of XGB are that it is less prone to overfitting because the model includes a regular term in the target function. Moreover, the model produces good performance with good execution speed due to its system optimisations and algorithm enhancements. This includes parallel optimisation where the order of loops within the model are interchangeable and using a regular term on the target function to prevent overfitting [55].

On the other hand, the model needs to be tuned correctly to achieve a good performance which is difficult because there are many hyperparameters to tune. The model can be difficult to interpret.

XGB is an ensemble algorithm where it groups many weak classifiers to generate a stronger classifier. A decision tree is grown by splitting features, which generates a new tree that corrects the previous tree's mistakes, making the model iterative. Several trees are aggregated to predict, in this case, the stock trend [55].

3.5 **Performance Criteria**

The aims of this thesis are answered by computing several performance metrics for each machine learning model on each dataset. Using sklearn accuracy, Matthew's Correlation Coefficient, Receiver Operator Characteristic curve and Error Rate are found to describe the performance of the classification models.

The main terminology used throughout this chapter are as follows [56]:

- True Positives (TP) - the actual value is true and model predicted true, for this thesis, the model correctly predicted up movements
- True Negatives (TN) - the actual value is false and model predicted false, for this thesis, the model correctly predicted down movements
- False Positives (FP) - the actual value is true and model predicted false, for this thesis, the model incorrectly predicted down movements as up
- False Negatives (FN) - the actual value is false and model predicted true, for this thesis, the model incorrectly predicted up movements as down.

3.5.1 **Accuracy**

Accuracy is the number of correct classifications divided by the total number of predictions and works well when the model is balanced. The equation can be seen in (1).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)\ [8]$$

Accuracy was chosen as the baseline measurement for the hypothesis since it is the most widely used performance metric in previous literature shown in [21], [24] and [27].

3.5.2 **Matthews Correlation Coefficient**

This leads to the next performance metric, the Matthews Correlation Coefficient (MCC). An MCC metric shows how correlated predictions are to the true values, where -1 is an inverse prediction, +1 is perfect prediction and 0 represents a random prediction.

More details on how MCC is calculated can be located in [57]. The equation can be seen in (2).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (2)\ [57]$$

The hypothesis is evaluated with the MCC metric because it takes into considers account all four values (FP, FN, TP and TN), avoiding bias as all classes are equal. MCC is a very useful indicator as shown from research in [58], which finds that MCC scores are more truthful and reliable than accuracy. Accuracy uses TP and TN and does not consider FP and FN meaning it is asymmetric by nature. [57] and [59] use MCC well in similar studies.

3.5.3 **The Receiver Operator Characteristic Curve**

The Receiver Operator Characteristic (ROC) curve is an evaluation metric used to measure performance. ROC is a useful measure of performance for determining the hypothesis since it does not depend on class distribution and so is a better choice when comparing models.

The graph plots the false positive rate (FPR) on the x-axis against the true positive rate (TPR) on the y-axis, separating the signal from the noise. The equations can be seen in (3) and (4).

$$TPR = \frac{TP}{TP + FN} \quad (3)\ [62] \qquad FPR = \frac{FP}{FP + TN} \quad (4)\ [62]$$

The steeper the curve, the more TPR is maximised and FPR is minimised. More details of how to calculate these parameters can be seen in [8]. Curves which are closer to the top-left corner indicate a better performance whereas curves closer to the baseline of a random classifier (the diagonal from bottom left to top right) are deemed worse performers.

The area beneath an ROC curve is known as Area Under Curve (AUC), which measures how well a classifier can separate classes. The higher the AUC, the better the classifier is at distinguishing between positive and negative classes where a value of 1 means the classifier is perfectly able to predict and a value of 0 means the classifier cannot predict at all. An AUC value of 0.5 means the model predicts no better or worse than random chance.

3.5.4 **Error Rate**

The Error Rate (ERR) is calculated from the confusion matrices as the total of incorrect predictions divided by the sum of the entire total dataset. The best error rate is 0 (0%), whereas the worst is 1 (100%). The values were calculated using equation (5).

$$\mathrm{ERR} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{P} + \mathrm{N}} \quad (5)\ [61]$$

Only the error rate associated with the XGB algorithm is considered in detail here due to it being the best performing classifier in previous literature [8]. This motivates why it was chosen to determine the hypothesis in this work.

Chapter 4

Results

This chapter illustrates the results achieved from the performance criteria and is structured into accuracy, Matthews Correlation Coefficient, Receiver Operator Characteristic and Error Rate. Key findings are highlighted and discussed in subsequent chapters.

4.1 Accuracy

Figures 5-7 bar charts show the accuracy of each algorithm with and without sentiment for all three datasets. The average accuracy with sentiment for Financial News, Twitter, and Reddit, across all three models, are 55.2%, 63% and 62.2%, respectively. The accuracy obtained without sentiment is lower for Twitter, 51.5%, and Reddit, 50%, but slightly higher for Financial News, 55.5%. Without sentiment, prediction for each dataset is around 50% which is equivalent to random accuracy. There is a substantial increase in accuracy with the addition of sentiment to the Twitter and Reddit datasets.

With the addition of sentiment, the maximum accuracy mean difference was the Reddit dataset at 12.2%. It is interesting to note the baseline data source, Financial News, has a negative accuracy mean difference with sentiment, while prediction accuracy has increased with sentiment for the other two data sources.

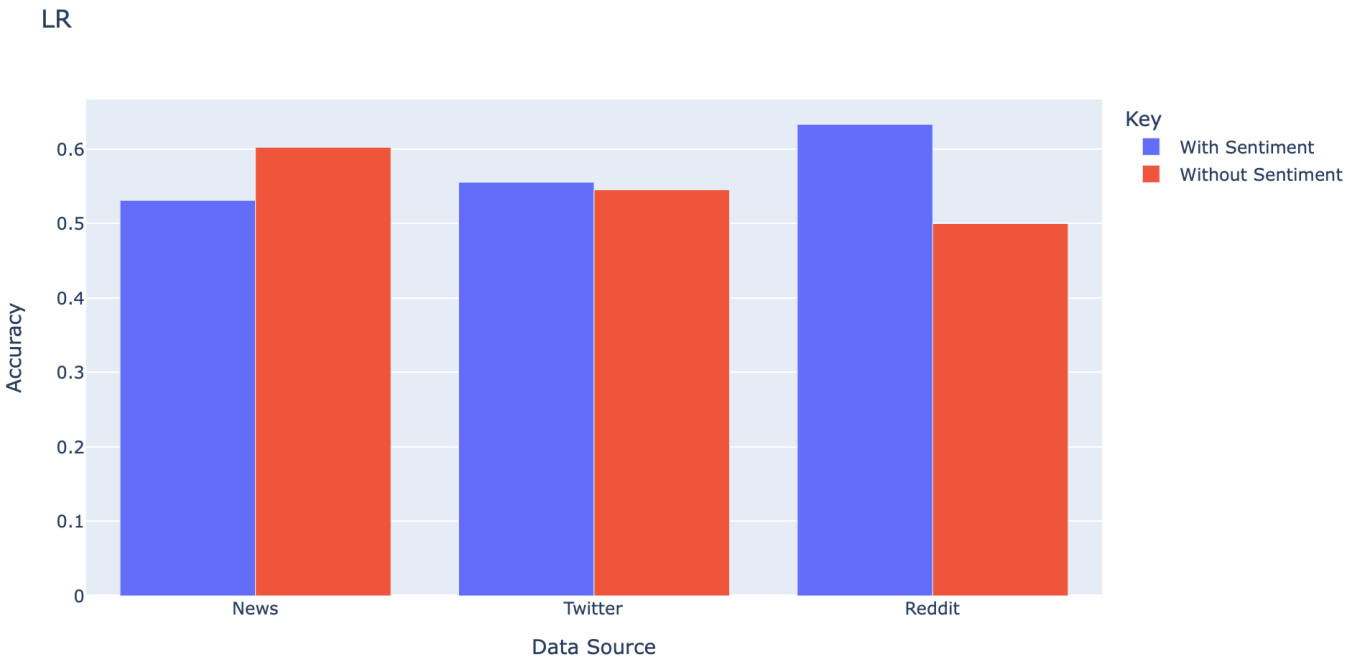

*Figure 5: Bar Chart to show accuracy with and without sentiment for **Logistic Regression**. Accuracy scores are on the y-axis and Data source on the x-axis. Scores without sentiment are highlighted in pink and with sentiment in purple. **The highest score can be seen by Reddit with sentiment at 0.633.***

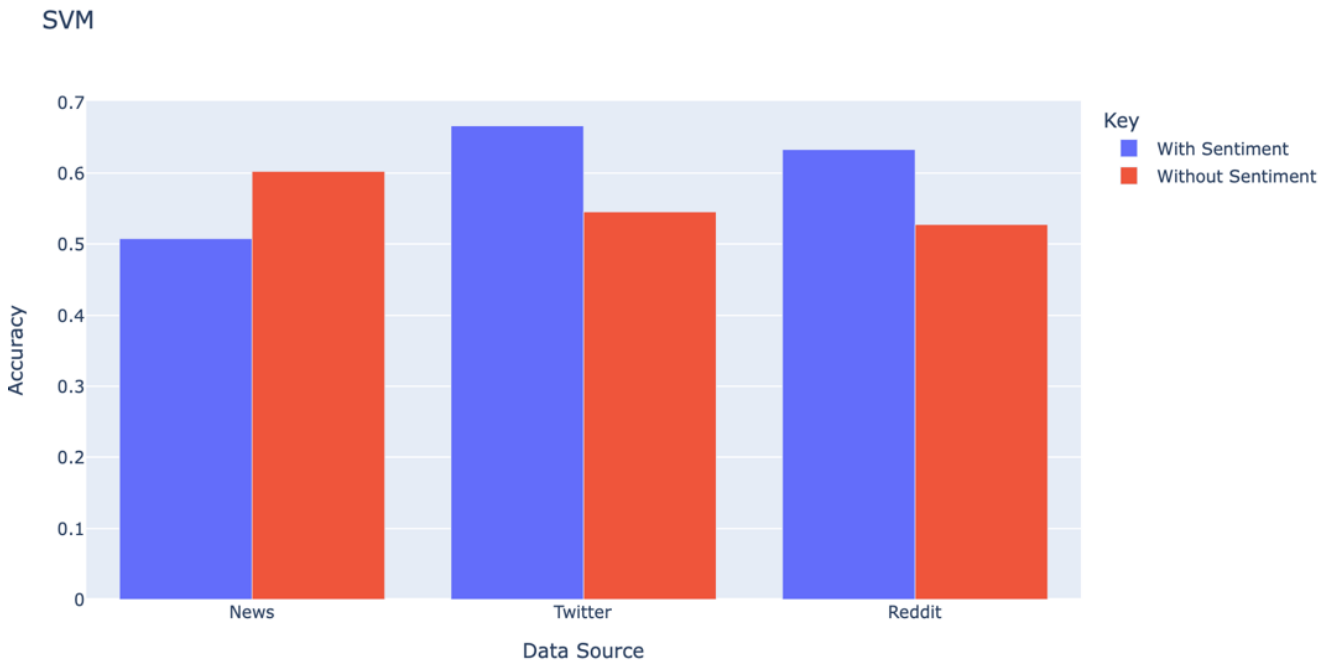

*Figure 6: Bar Chart to show accuracy with and without sentiment for **Support Vector Machine**. Accuracy scores are on the y-axis and Data source on the x-axis. Scores without sentiment are highlighted in pink and with sentiment in purple. **The highest score can be seen from Twitter with sentiment at 0.667.***

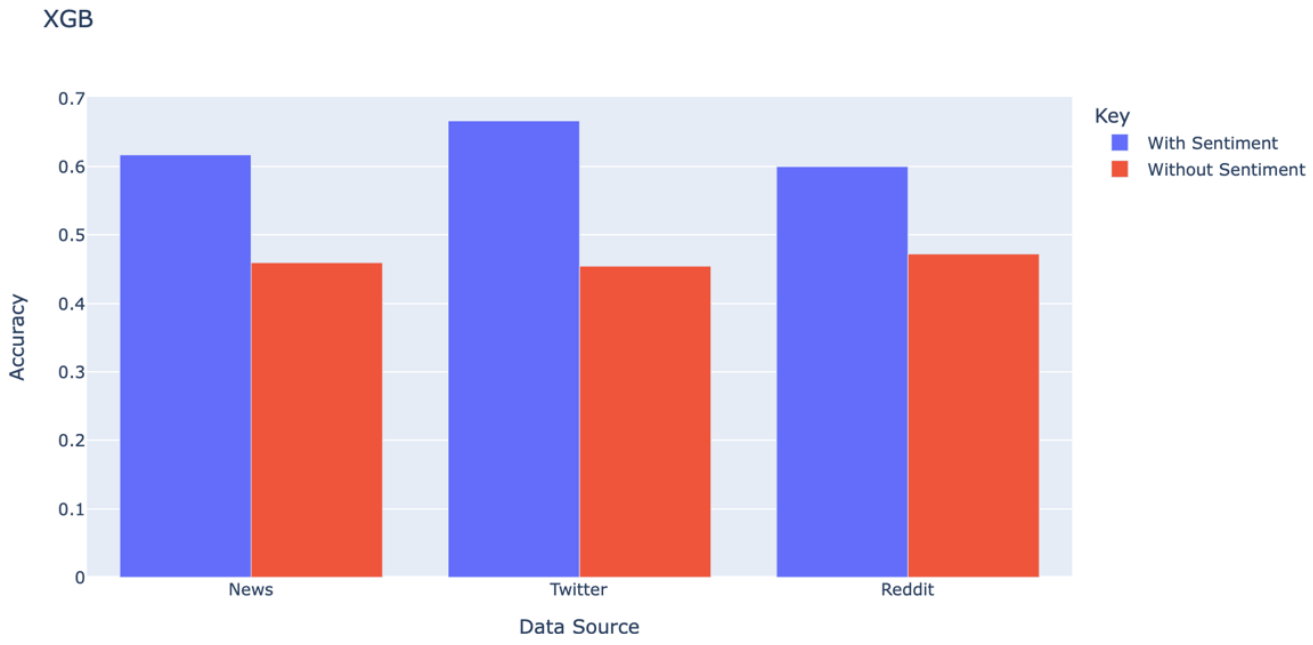

*Figure 7: Bar Chart to show accuracy with and without sentiment for **Extreme Gradient Boosting**. Accuracy scores are on the y-axis and Data source on the x-axis. Scores without sentiment are highlighted in pink and with sentiment in purple. **XGB shows the biggest increase with sentiment compared to the other two models.***

4.2 **Matthews Correlation Coefficient**

Tables 1-3 summarise accuracy and MCC performance criteria.

Financial News

Model	Accuracy with sentiment feature	Accuracy without sentiment feature	Accuracy difference, %	MCC with sentiment	MCC without sentiment	MCC difference, %
LR	0.531	0.602	-7.1%	0.124	0.065	5.9%
SVM	0.508	0.602	-9.4%	0.002	0.023	-2.1%
XGBoost	0.617	0.46	15.7%	0.24	-0.092	33.2%
Average	0.552	0.555	-0.267%	0.122	-0.001	12.3%

Table 1: Table of results showing accuracy and MCC for the News dataset. News has a slight negative accuracy difference with and without sentiment but a positive MCC difference. News sentiment has an inconclusive impact on prediction.

Twitter

Model	Accuracy with sentiment feature	Accuracy without sentiment feature	Accuracy difference, %	MCC with sentiment	MCC without sentiment	MCC difference, %
LR	0.56	0.545	1.5%	0	0	0
SVM	0.667	0.545	12.2%	0	0	0
XGBoost	0.667	0.455	21.2%	0.189	-0.028	21.7%
Average	0.63	0.515	11.5%	0.063	-0.009	7.2%

Table 2: Table of results showing accuracy and MCC for the Twitter dataset. Twitter shows a higher positive difference with accuracy than MCC. Twitter sentiment has a positive impact on prediction.

Reddit

Model	Accuracy with sentiment feature	Accuracy without sentiment feature	Accuracy difference, %	MCC with sentiment	MCC without sentiment	MCC difference, %
LR	0.633	0.5	13.3%	0.167	-0.203	37%
SVM	0.633	0.528	10.5%	0	-0.034	3.4%
XGBoost	0.6	0.472	12.8%	0.208	-0.019	22.7%
Average	0.622	0.5	12.2%	0.125	-0.085	21%

Table 3: Table of results showing accuracy and MCC for the Reddit dataset. Reddit shows a higher positive difference with MCC than accuracy. Reddit sentiment has a positive impact on prediction.

Contradictory with accuracy, MCC reports a positive mean difference for all three data sources with sentiment. Additionally, Reddit achieves the highest average difference at 21% and Twitter the lowest at 7.2%. The baseline data source, News, shows an increase in prediction with sentiment using the MCC metric at a 12.3% mean increase. Without sentiment, all three data sources are centred slightly below 0 indicating random prediction.

4.3 The Receiver Operator Characteristic Curve

The ROC curves and AUC values are seen in Figures 8-25.

Logistic Regression

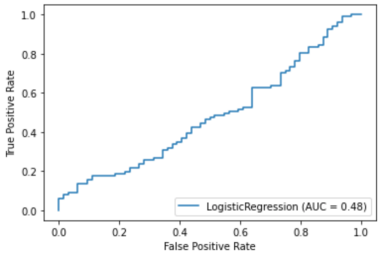

Figure 8: ROC curve of LR without News sentiment

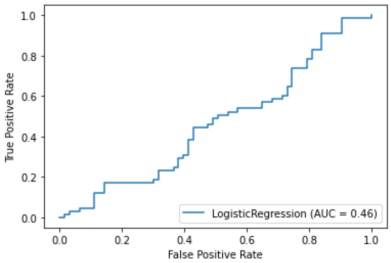

Figure 9: ROC curve of LR with News sentiment

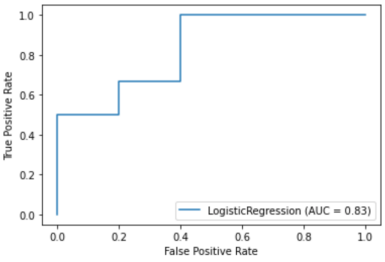

Figure 10: ROC curve of LR without Twitter sentiment

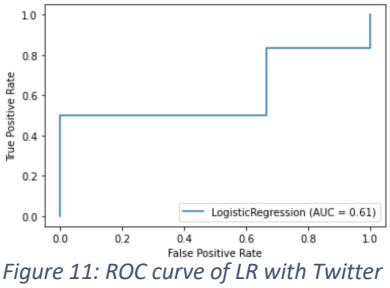

Figure 11: ROC curve of LR with Twitter sentiment

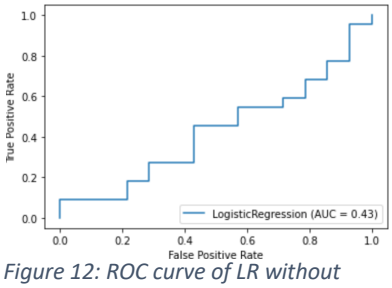

Figure 12: ROC curve of LR without Reddit sentiment

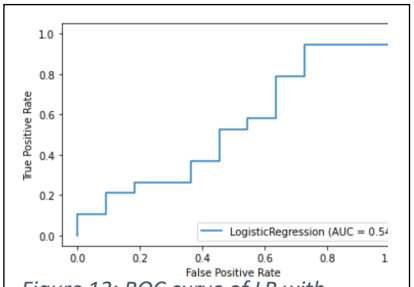

Figure 13: ROC curve of LR with Reddit sentiment

Support Vector Machine

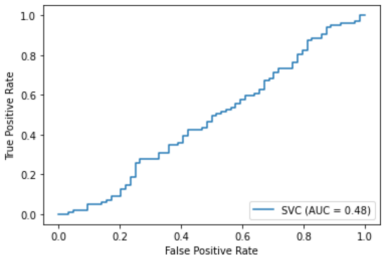

Figure 14: ROC curve of SVM without News sentiment

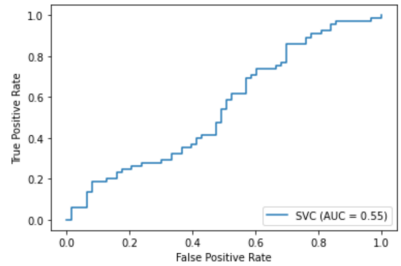

Figure 15: ROC curve of SVM with News sentiment

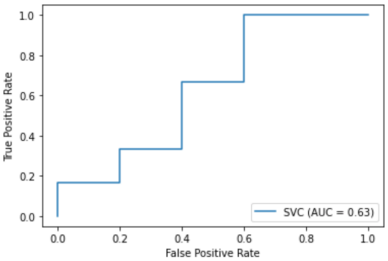

Figure 16: ROC curve of SVM without Twitter sentiment

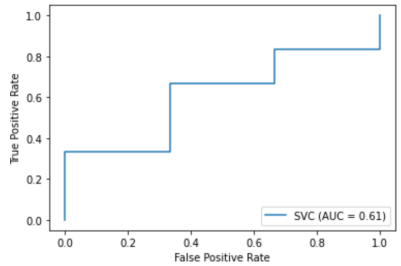

Figure 17: ROC curve of SVM with Twitter sentiment

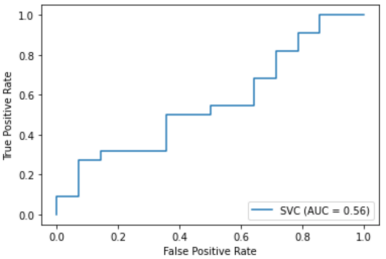

Figure 18: ROC curve of SVM without Reddit sentiment

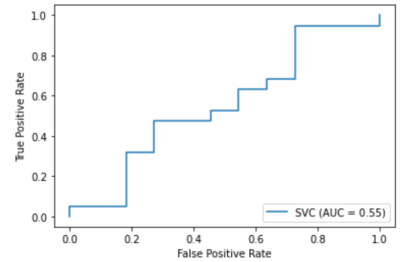

Figure 19: ROC curve of SVM with Reddit sentiment

Extreme Gradient Boosting

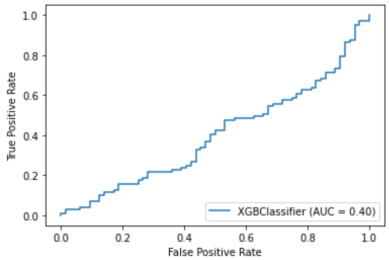

Figure 20: ROC curve of XGB without News sentiment

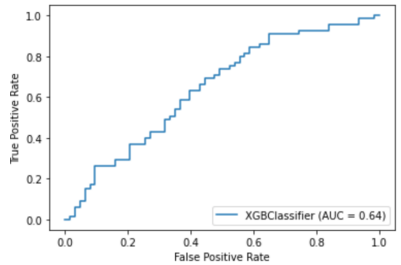

Figure 21: ROC curve of XGB with News sentiment

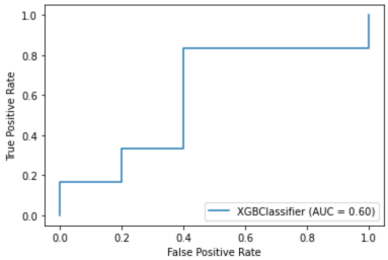

Figure 22: ROC curve of XGB without Twitter sentiment

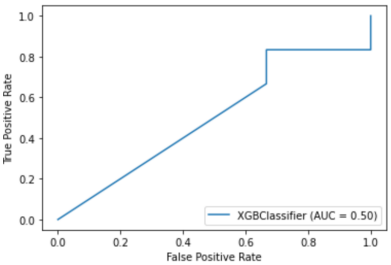

Figure 23: ROC curve of XGB with Twitter sentiment

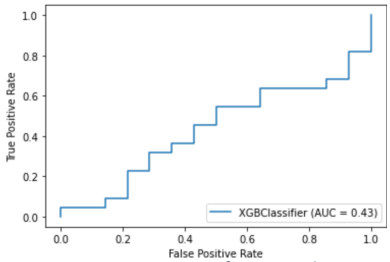

Figure 24: ROC curve of XGB without Reddit sentiment

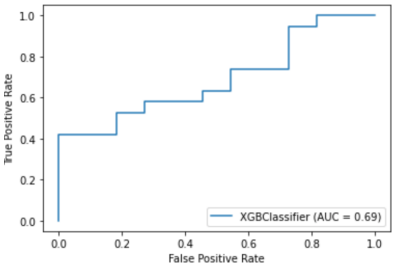

Figure 25: ROC curve of XGB with Reddit sentiment

The AUC values are seen from Figures 13-24. For two out of the three algorithm models used, AUC values increased for News and Reddit with a negligible difference in the third. Reddit is credited with the highest mean difference with sentiment at 12%, News achieves the next highest at 9.7% while Twitter results in a poor mean difference of -11.4%. Interestingly, AUC values fall for Twitter when sentiment is applied in all three models even if its mean AUC value is marginally higher than the News data source, 0.573 > 0.55.

The best AUC value computed with sentiment is by Reddit at 0.69 with XGB and the worst AUC value found with sentiment is by News at 0.46 with LR. As shown from the ROC curves XGB performs the best with sentiment while LR performs the worst. With all 3 data sources when sentiment is applied the mean AUC values are greater than 0.5 showing there is a high chance that the classifiers will distinguish the positive class values from the negative class values and therefore will be able to predict better than random chance.

4.4 **Error Rate**

Table 4 shows the error rate (ERR) for each data source with and without sentiment.

Data Source	ERR with sentiment feature	ERR without sentiment feature	ERR difference, %
News	0.38	0.54	16%
Twitter	0.33	0.54	21%
Reddit	0.4	0.53	13%
Average	0.37	0.54	16.7%

Table 4: Table of results showing ERR for the XGB algorithm. According to ERR difference for XGB, Twitter has the biggest increase in correct classifications when sentiment is applied and Reddit the worst. This metric may be skewed due to Twitter's smaller time period.

With the addition of sentiment, News misclassified 38% of up and down movements, 33% for Twitter and 40% for Reddit. The ERR shows an improved classification of 16.7% on average when sentiment is applied. In general, across all algorithms, the data sources are better at predicting up movements than down.

Chapter 5

Discussion

This chapter will discuss the aims proposed at the start of the thesis and the main results.

The chapter is broken down into discussing the first aim and how the results found compare to the literature, the second aim and how the results found compare to the literature and Reddit vs Twitter.

5.1 To Investigate if Sentiment Analysis Can Improve Stock Trend Prediction

The first aim explores if sentiment analysis can improve stock trend prediction. As seen by Tables 4-6, the MCC results demonstrate that sentiment positively affects stock prediction for all data sources. Prior to adding sentiment, the MCC values are centred slightly below 0 indicating random prediction. By adding sentiment, values have increased to 0.122 for News, 0.063 for Twitter and 0.125 for Reddit which are beyond the random prediction first obtained.

Additionally, these results are consistent with those achieved by the error rate in Table 4. They show with VADER sentiment applied the ERR shows an improved classification of 16.7% on average. Reddit improves correct classifications by 13%, Twitter by 21% and News by 16% where without sentiment the ERR was just above 50%.

Without sentiment, accuracy prediction for each dataset is around 50% which is equivalent to random chance. The average accuracy with sentiment for News, Twitter, and Reddit, across all three models, are 55.2%, 63% and 62.2%, respectively. The Reddit and Twitter datasets show increased performance of mean accuracy difference at 12.2% and 11.5%, respectively. These results indicate the positive significance of sentiment as an additional variable to stock trend prediction of the S&P 500.

The results obtained for AUC values with sentiment for two out of the three algorithm models used, show increased scores for News and Reddit with a negligible difference in the third. Reddit is credited with the highest mean difference with sentiment at 12%, News achieves the next highest at 9.7%. Not only do these findings illustrate that stock market prediction may not be a random walk as suggested by the EMH hypothesis but provides reasonable evidence that sentiment can improve stock prediction.

However, the baseline classifier, News, fell by -0.267% by adding sentiment in terms of mean accuracy difference. The fact that News fell by a small margin and the overall accuracy of 55.2% is still above the 50% random chance, with the addition of sentiment, suggests sentiment may not be the factor that caused a decrease in accuracy and this area needs to be explored further in the future. While Twitter results in a negative AUC mean difference of -11.4% its mean AUC value is marginally higher than the News data source, 0.573 > 0.55 and so the poor score could be attributed to the lower number of instances compared to the other data sources.

Generally, these results highlight that sentiment has a favourable impact on stock trend prediction because sentiment analysis reflects the public's psychology, helping machine learning algorithms infer stock market patterns. The results from the performance criteria accept the hypothesis as implementing sentiment analysis prediction of the S&P 500 is higher than random chance.

5.2 Comparison with Results in the Literature

The results in this study agree with the work found in [20], [21] and [25] that suggest sentiment can improve stock prediction. Compatible work uses polarity and sentiment alongside closing price to predict a stock's trend. This work showcases results for the entire S&P 500 which algins with existing research using single companies.

However, the results are not as optimistic as those in [22] and [26] where accuracy was 72%-73% with SVM and 69% with LR. This may be because in this work text data is extracted throughout the whole day while stock price data is only considered at market close. This means daily price direction may not coincide with the text data on that day. Sentiment may change during the day based upon financial reports, macroeconomic factors, public announcements which obscure the polarity for the sentiment analyser. To rectify this, additional sources of information could be used and a cut off time for textual data could be employed to coincide with the market close.

5.3 How do Reddit and Twitter Compare to the Standard Benchmark, Financial News, as a Source of Sentiment Analysis

Regarding the second aim, the results are less clear cut. According to results from accuracy and AUC values both Reddit and Twitter come out ahead against News. Reddit obtains scores of 62.2% for accuracy and 59.3% for AUC. Twitter obtains scores of 63% for accuracy and 57.3% for AUC. News obtains scores of 55.2% for accuracy and 55% for AUC. These results suggest both Twitter and Reddit are better data sources for sentiment analysis in this area of research as they exceed the values of the benchmark data source. Despite the social media scores not being greatly superior, as paper [62] states stock prediction is a challenging task where even small improvements lead to large potential profits for investors.

An explanation for this observation could be due to using VADER as the sentiment analyser which is predominantly a social media sentiment classifier as stated on the official Github [63]. The algorithm is better at predicting social media as it was built by inputting more informal language, emojis and sentiment intensity tools such as '!!'. These attributes are rarely found in a news article where the language is more formal and written to engage a reader much like a story. News articles may already be priced into the market as they are released after events whereas tweets and posts on social media more heavily speculate about future events which have not been priced in yet and therefore logically would be better indicators of future stock trends.

However, results from MCC suggest News is not the worst predictor as it attains an MCC score of 0.122 which is better than Twitter's score of 0.063. News may be a better predictor than social media platforms because the content is more insightful and of a richer quality. In order to truly determine which data source is the better predictor additional experimentation is needed with bigger and more balanced data. The work helps to fill in the research gap of current literature by comparing all three data sources.

5.4 Comparison with Results in the Literature

The results achieved for Reddit and Twitter are similar to those in [24] and [25], where accuracy ranged from 61.2%-65.6%. Moreover, work in [24] highlights that sentiment improves accuracy by 10% for social media data which is concurrent with this thesis, 11.5%-12.2%. This may be due to the similar specific pre-processing conducted in this thesis which improved the quality of the text data. This work achieves a better increase in accuracy with and without sentiment, for Twitter at 11.5% compared to [25] which obtained an increase of 0.6%-3.3%. The authors also only used positive and negative sentiment but used machine learning algorithms to determine sentiment unlike this study.

The results in this thesis find a negative correlation between News sentiment and stock trend agreeing with [18]. However, the work in [20], [21] show a positive increase in prediction when News sentiment is applied. The discrepancy between the results in this thesis and previous literature may occur because as stated by the author in [20] accuracy depends on the news media used as they have their own characteristics, suggesting the quality of CNBC articles may be poor compared to other news sources such as the BBC.

5.5 **Reddit vs Twitter**

The second aim can be expanded to determine which social media platform is the better predictor in this work. According to the mean scores for the MCC metric and AUC values, Reddit would be the better social media for sentiment analysis with scores of 0.125 and 59.3% compared to Twitter's scores of 0.063 and 57.3%. While Twitter's mean accuracy is marginally better than Reddit's at 63% to 62.2%, accuracy is seen as more biased and less reliable than the MCC metric as explained in the Methodology chapter.

There are several factors to why Reddit may be a better predictor than Twitter. Firstly, the content on Reddit is filtered by subreddit where most users only interact with one subreddit which increases the value of the information as active users are more knowledgeable about their respective topic [64]. This effect is furthered as Reddit posts contain less spam and jargon due to moderators which keep posts on topic. Both these reasons increase the significance of the textual data on Reddit making it more impactful for sentiment analysis. Furthermore, the two platforms are different in nature. Reddit aggregates news and topics for discussion while Twitter is a networking service where users can like and tweet about their interests. This is shown by the average time users spend on their respective platforms where Reddit's users spend 16 minutes and Twitter's users spend 3.39 minutes [65]. This may be due to the fact Twitter's users are capped to 280 characters, so posts are designed to be short and to the point. On the other hand, Reddit posts can be longer in length where users can analyse and discuss content making it more useful for sentiment analysis. This point was first alluded to in the explanatory data analysis section where more textual information could result in a more accurate polarity from the VADER analyser. This in turn increases the accuracy of machine learning algorithms as they can better determine stock trends.

Chapter 6

Limitations

One important limitation for this thesis is the effect of the different time periods as the three data sources did not completely correlate with each other in terms of time. Twitter had the smallest time period while News had the longest. This hinders the reliability of the results when comparing them due to outside variables such as financial reports being released or macroeconomic changes affecting the data sources. The comparison on which data source is the best is not entirely fair due to this discrepancy and datasets covering the same time period should be used in future work.

Furthering this point, the datasets are imbalanced where the number of up movements outweigh the number of down movements. While in the last 10 years on average the stock market has risen, the imbalanced datasets used could be why the models were better able to predict up movements degrading the quality of the results. Balanced datasets make evaluations fairer as there is no bias, improving prediction quality. To combat these points, datasets covering the 2008 financial crisis period would be beneficial as they contain many down movements making the research more valuable [2].

A further limitation for this study is the use of GridSearch for finding hyperparameters. Only a small selection of hyperparameters were chosen for each model due to time constraints, meaning the ideal choice was likely not found, potentially harming prediction [8].

Chapter 7

Conclusion

The purpose of this thesis is to investigate the classification performance of LR, SVM and XGB in predicting the daily trend movement of the S&P 500, with and without the addition of sentiment analysis from Reddit, Twitter and Financial News as a feature. Different pre-processing techniques were carried out to reduce noise and improve classification prediction and a pre-built lexicon, VADER, was used to determine the polarity of data sources to use as an input in the machine learning models.

Overall, this work provides evidence that sentiment is a viable and useful variable for stock prediction. All models for the Reddit dataset, two models for the Twitter dataset and one for the Financial News dataset achieves scores above 60% accuracy which is better than the 50% random walk. This finding is concurrent with the other performance metrics. As stated by NikolaNEWS, if a model can reach 60% on predicting stock movement directions, it can deliver good returns for investors [66]. The results show the hypothesis can be accepted since implementing sentiment analysis can predict the S&P 500 trend with higher than random chance. From the majority of results, Reddit and Twitter have shown to have an advantage over Financial News as a source of textual data giving investors additional data sources to explore as they fine-tune their own models for prediction. Furthermore, while work comparing Reddit vs Twitter is not conclusive, Reddit seems to have an edge in this study due to its unrestrictive word length. From a model aspect, XGBoost performs better compared to LR and SVM. In most of the performance criteria, XGBoost displayed the best predictions with each data source.

The results achieved are largely in line with previous research. Nevertheless, because different datasets, methods and techniques were utilized, a direct comparison of the results with those in previous studies is not feasible [8]. This work enhances the field of stock prediction, specifically social media's relationship with stocks. It gives evidence to disprove the EMH's hypothesis that the stock market can only be predicted with up to 50% accuracy.

7.1 Future Work

With respect to future work, adding neutral sentiment could be worthwhile to reduce noise and misclassification. While large amounts of textual data may be labelled neutral, making it less useful for the models; prediction quality could improve as slightly positive or negative sentiments may give false readings without neutral sentiment.

Numerous recent studies only exploit features of textual data and disregard the effect of stock price data. Future work could use numeric features such as technical indicators, volatility, and expected revenues of companies. To obtain the best predictions, numerical data could be combined with textual data that has been very successful in past research with scores of up to 86% in [8] and [23].

This work focuses exclusively on a basket of stocks, S&P 500. Adding single stocks or another index would help to reinforce the validity of the findings.

Finally, to enhance this research for investors, a trading system could be developed to compare with the popular investment strategy, buy and hold. This would increase the meaningfulness of the work as it would be more comparable to real life situations.

Chapter 8

Project Management

This chapter summarises how the thesis was managed.

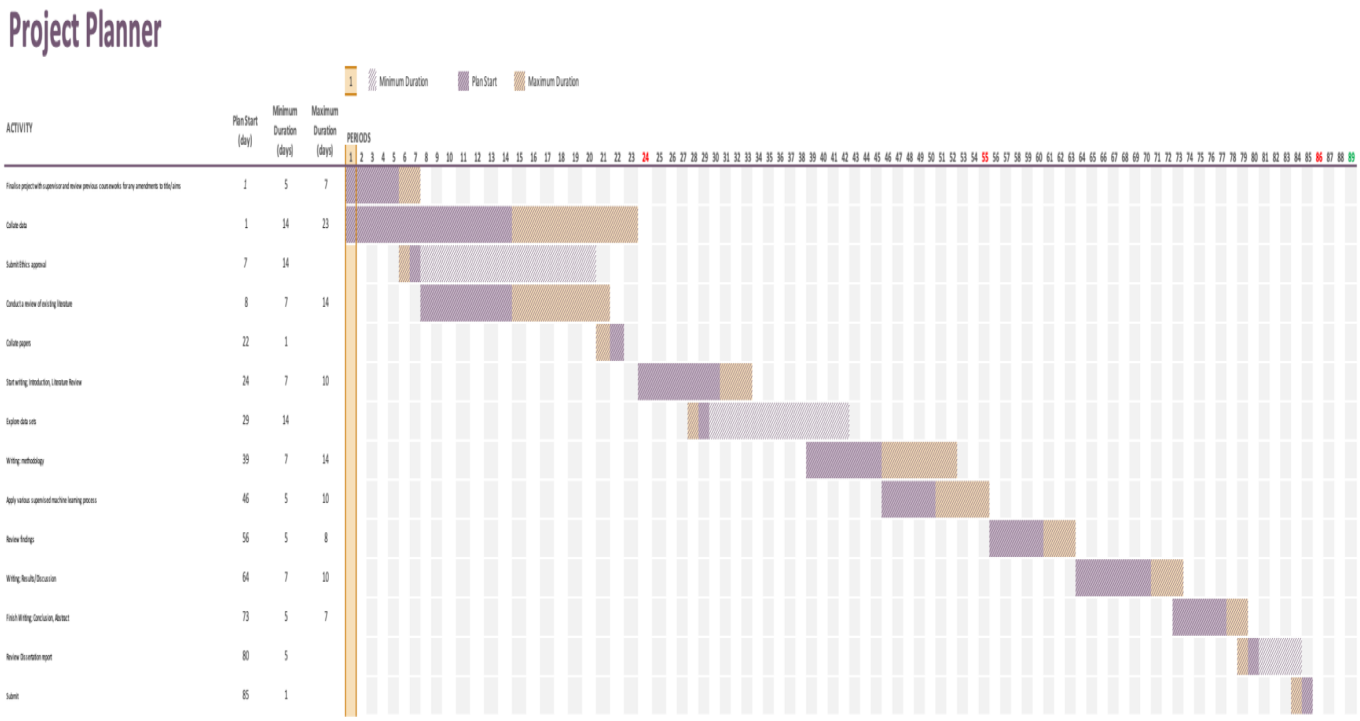

Figure 26: The Gnatt chart shows the initial project plan for this thesis. The plan was modified due to COVID illness which was accounted for in the 'float' and extension. (A bigger version can be seen in the appendix)

The main activities shown in Figure 26 were split into smaller activities to provide more detail of the plan:

Collate data: The data was collected for Twitter, Reddit, Financial Times, and stock prices. This involved looking for datasets available online as extracting the data manually was deemed ineffective.

Conduct a review of existing literature: The most recent literature of relevance, historically significant papers, research literature on different methodologies, sentiment analysers and evaluation metrics were accessed and summarised.

Collate papers: Finalised the methodology, and what stock(s) to use.

Explore data sets: A preliminary explanatory data analysis on each data set was conducted. The next stage involved applying pre-processing techniques, cleaning the datasets, and utilising VADER on the data. This section contained coding on Python to apply the chosen techniques and find information about the datasets.

Apply various supervised machine learning process: This work concentrated on implementing three machine learning models to determine prediction of stock trends and applying the performance criteria using various libraries in Python.

Review findings: The code was checked, results analysed, and findings were interpreted. This section of work also focused on parameter tuning.

Review Dissertation report: The first draft was reviewed and amend based upon a self-evaluation checklist. The report was created and formatted with an initial inspection from the supervisor. The supervisor comments were incorporated along with feedback from the demonstrations.

Challenges: The initial Gnatt chart in Figure 26 was modified due to COVID illness on the 37th day. This meant work was delayed by two weeks. Unforeseen circumstances were accounted for by the 'float' where certain activities mentioned above required less time to complete such as collating the data only took 13 days instead of 23. The 'float' number of days was crucial to include upon initial planning as it allowed for a contingency plan for problems such as the illness. Additionally, a two extension was granted so the total number of days for the project was 103. The ethics approval form was re-submitted due to the new deadline.

The order of activities was altered to be more cohesive. Applying machine learning processes was completed before writing the Methodology so the coding work was done together.

Tools: Certain tools helped to not only manage time more effectively but also to track progress. These included weekly meetings with my supervisor and using the Microsoft Planner tool available online. The Planner tool was set up in the first week with my supervisor. The tool helped to organise the work into 'buckets' and then into tasks. The work was colour coded where red was used for work that need finishing right away. Deadlines for tasks were applied and ticked off on completion aiding progress tracking. Weekly supervisor meetings helped to stay on task and discuss questions based upon the previous week's work.

All the objectives for this thesis were completed before the deadline with time to review and check.

References

[1]Bo Qian and Khaled Rasheed, Stock market prediction with multiple classifiers Applied Intelligence, vol. 26, pp. 2533, February 2007

[2]Lubitz, M., 2017. Who drives the market? Sentiment analysis of financial news posted on Reddit and Financial Times. University of Freiburg: http://ad-publications. informatik. uni-freiburg. de/theses/Bachelor_ Michael_Lubitz_2018. Pdf.

[3] K. C. Butler and S. J. Malaikah, “Efficiency and inefficiency in thinly traded stock markets: Kuwait and saudi arabia”, Journal of banking & finance, vol. 16, no. 1, pp. 197–210, 1992.

[4] B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers", Applied Intelligence, vol. 26, pp. 25-33, 2007.][M. G. Kavussanos and E. Dockery, “A multivariate test for stock market efficiency: The case of ase”, Applied financial economics, vol. 11, no. 5, pp. 573– 579, 2001.

[5] Yildirim, H., 2017. Behavioral finance or efficient market hypothesis?. International Journal of Academic Value Studies, 3, pp.151-158.

[6] Blasco, N., Corredor, P. and Ferreruela, S., 2012. Market sentiment: a key factor of investors’ imitative behaviour. Accounting & Finance, 52(3), pp.663-689.

[7] Asadi S, Hadavandi E, Mehmanpazir F, Nakhostin MM. Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction. Knowledge-Based Systems. 2012;35:245-258

[8] Elena, P., 2021. Predicting the Movement Direction of OMXS30 Stock Index Using XGBoost and Sentiment Analysis.

[9] Chan, E. (2013). Algorithmic trading : Winning strategies and their rationale. Somerset: John Wiley & Sons, Incorporated.

[10] Guo, X., Lai, T. L., Shek, H., & Wong, S. P. (2016). Quantitative trading : Algorithms, analytics, data, models, optimization. Boca Raton: CRC Press LLC.

[11] Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., Sakurai, A.: Combining technical analysis with sentiment analysis for stock price prediction. In: 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, Sydney, NSW, pp. 800–807 (2011)] and can improve the predictive ability of machine learning models

[12] X. Li, P. Wu and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong", Information Processing & Management, vol. 57, no. 5, p. 102212, 2020

[13] W. Khan, M. Ghazanfar, M. Azam, A. Karami, K. Alyoubi and A. Alfakeeh, "Stock market prediction using machine learning classifiers and social media, news", Journal of Ambient Intelligence and Humanized Computing, 2020.

[14] Yu, L., Wu, J., Chang, P., & Chu, H. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news.

[15] Sohangir, S., Petty, N. and Wang, D., 2018, January. Financial sentiment lexicon analysis. In 2018 IEEE 12th international conference on semantic computing (ICSC) (pp. 286-289). IEEE.

[16] Agarwal, A., 2020, September. Sentiment analysis of financial news. In 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 312-315). IEEE.

[17] Li, Xiaodong and Haoran Xie and Li Chen and Jianping Wang and Xiaotie Deng. 2014. "News impact on stock price return via sentiment analysis." Knowledge-Based Systems (Elsevier) (69): 14-23. https://doi.org/10.1016/j.knosys.2014.04.022.

[18] Apcentral.collegeboard.org. 2014. [online] Available at: <https://apcentral.collegeboard.org/pdf/ap19-apc-research-sample-b.pdf?course=ap-research>.

[19] Gui Jr, H., 2019. Stock Prediction Based on Social Media Data via Sentiment Analysis: a Study on Reddit (Master's thesis).

[20] Kim, Y., Jeong, S.R. and Ghani, I., 2014. Text opinion mining to analyze news for stock market prediction. Int. J. Advance. Soft Comput. Appl, 6(1), pp.2074-8523.

[21] Rahman, A.S.A., Abdul-Rahman, S. and Mutalib, S., 2017, November. Mining textual terms for stock market prediction analysis using financial news. In International Conference on Soft Computing in Data Science (pp. 293-305). Springer, Singapore.

[22] Dang, M. and Duong, D., 2016, September. Improvement methods for stock market prediction using financial news articles. In 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS) (pp. 125-129). IEEE.

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

[24] Bouktif, S., Fiaz, A. and Awad, M., 2019, October. Stock market movement prediction using disparate text features with machine learning. In 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS) (pp. 1-6). IEEE.

[25] Gupta, R. and Chen, M., 2020, August. Sentiment Analysis for Stock Price Prediction. In 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 213-218). IEEE.

[26] Pagolu, V.S., Reddy, K.N., Panda, G. and Majhi, B., 2016, October. Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.

[27] Oberlo.com. n.d. 10 Reddit Statistics You Should Know in 2021 [Infographic]. [online] Available at: <https://www.oberlo.com/blog/reddit-statistics#:~:text=The%20very%20first%20thing%20you%20need%20to%20know,percent%20year-over-year%20increase%20from%20October%202019’s%2036%20million>.

[28] Buntain, C. and Golbeck, J., 2014, April. Identifying social roles in reddit using network structure. In Proceedings of the 23rd international conference on world wide web (pp. 615-620).

[29] Kaggle.com. n.d. Reddit - Finance Posts (r/wallstreetbets, r/gm...). [online] Available at: <https://www.kaggle.com/leukipp/reddit-finance-data>.

[30] En.wikipedia.org. n.d. Twitter - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Twitter#Usage>.

[31] Oberlo.com. n.d. 10 TwitterStatistics You Should Know in 2021 [Infographic]. [online] Available at: https://www.oberlo.co.uk/blog/twitter-statistics
[32] Kordonis, J., Symeonidis, S. and Arampatzis, A., 2016, November. Stock price forecasting via sentiment analysis on Twitter. In Proceedings of the 20th Pan-Hellenic Conference on Informatics (pp. 1-6).
[33] Salač, A., 2019. Forecasting of the cryptocurrency market through social media sentiment analysis (Bachelor's thesis, University of Twente).
[34] IEEE DataPort. n.d. Stock Market Tweets Data. [online] Available at: <https://ieee-dataport.org/open-access/stock-market-tweets-data>.
[35] CNBC Digital Has Record Year in 2020
[online] Available at: <https://www.cnbc.com/2021/01/19/cnbc-digital-has-record-year-in-2020.html>.
[36] Kaggle.com. n.d. Financial News Headlines Data. [online] Available at: <https://www.kaggle.com/notlucasp/financial-news-headlines>.Kaggle.com. n.d. Financial News Headlines Data. [online] Available at: <https://www.kaggle.com/notlucasp/financial-news-headlines>.
[37] Uk.finance.yahoo.com. n.d. Yahoo Finance. [online] Available at: <https://uk.finance.yahoo.com>.
[38]En.wikipedia.org. n.d. S&P 500 - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/S%26P_500>.
[39] Xu, S.Y. and Berkely, C.U., 2014. Stock price forecasting using information from Yahoo finance and Google trend. UC Brekley.
[40] Priya, S., Sequeira, R., Chandra, J. and Dandapat, S.K., 2019. Where should one get news updates: Twitter or Reddit. Online Social Networks and Media, 9, pp.17-29.
[41] Kalyani, J., Bharathi, P. and Jyothi, P., 2016. Stock trend prediction using news sentiment analysis. arXiv preprint arXiv:1607.01958.
[42] Secure.ecs.soton.ac.uk. n.d. ECS - ECS Intranet Login. [online] Available at: <https://secure.ecs.soton.ac.uk/notes/ >.
[43] Chan, M., 2002. Advances in Knowledge Discovery and Data Mining.
[44] D. Enke and S. Thawornwong, "The use of data mining and neural networks for forecasting stock market returns", Expert Systems with Applications, vol. 29, no. 4, pp. 927- 940, 2005.]
[45] Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. Paper presented at the Eighth International AAAI Conference on Weblogs and Social Media,
[46] Kaggle.com. n.d. Stock-Market Sentiment Dataset. [online] Available at: <https://www.kaggle.com/yash612/stockmarket-sentiment-dataset>.
[47] Chen, C.C., Huang, H.H. and Chen, H.H., 2018, May. NTUSD-Fin: a market sentiment dictionary for financial social media data applications. In Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018).
[48] Medium. n.d. CROSS-VALIDATION IN TIME SERIES MODEL.. [online] Available at: <https://medium.com/@pradip.samuel/cross-validation-in-time-series-model-b07fbba65db7>.
[49] Attigeri, G. V., MM, M. P., Pai, R. M., and Nayak, A. (2015). Stock market prediction: A big data approach. In TENCON 2015-2015 IEEE Region 10 Conference, pages 1–5. IEEE.
[50] Dutta, A., Bandopadhyay, G. and Sengupta, S., 2012. Prediction of stock performance in the Indian stock market using logistic regression. International Journal of Business and Information, 7(1), p.105.
[51] Shi, Y., Zheng, Y., Guo, K. and Ren, X., 2021. Stock movement prediction with sentiment analysis based on deep learning networks. Concurrency and Computation: Practice and Experience, 33(6), p.e6076.
[52] Géron, A., 2019. Hands-on Machine Learning with Beijing Boston Farnham Sebastopol Tokyo Scikit-Learn, Keras, and TensorFlow. 2nd ed.
[53] Hu, Z., Zhu, J. and Tse, K., 2013, November. Stocks market prediction using support vector machine. In 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering (Vol. 2, pp. 115-118). IEEE.
[54] Zvonarev, A. and Bilyi, A., 2019. A comparison of machine learning methods of sentiment analysis based on Russian language Twitter data. In Proceedings of the 11th Majorov International Conference on Software Engineering and Computer Systems (MICSECS) (pp. 1-7).
[55] Yang, Y., Wu, Y., Wang, P. and Jiali, X., 2021. Stock Price Prediction Based on XGBoost and LightGBM. In E3S Web of Conferences (Vol. 275, p. 01040). EDP Sciences.
[56] Medium. n.d. Classification Report: Precision, Recall, F1-Score, Accuracy. [online] Available at: <https://medium.com/@kennymiyasato/classification-report-precision-recall-f1-score-accuracy-16a245a437a5>.
[57] Xu, Y. and Cohen, S.B., 2018, July. Stock movement prediction from tweets and historical prices. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1970-1979).
[58] Chicco, D. and Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), pp.1-13.
[59] Boyi Xie, Rebecca J Passonneau, Leon Wu, and Germa´n G Creamer. 2013. Semantic frames to pre- dict stock price movement. In Proceedings of the 51st Annual Meeting of the Association for Com- putational Linguistics. Sofia, Bulgaria, volume 1, pages 873–883.
[60] Analytics Vidhya. n.d. AUC-ROC Curve in Machine Learning Clearly Explained - Analytics Vidhya. [online] Available at: <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning>.
[61] Classifier evaluation with imbalanced datasets. n.d. Basic evaluation measures from the confusion matrix. [online] Available at: <https://classeval.wordpress.com/introduction/basic-evaluation-measures/>.
[62] Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In Proceedings of the 53rd Annual Meeting of the Association for Compu- tational Linguistics and the 7th International Joint Conference on Natural Language Processing. Bei- jing, China, volume 1, pages 1354–1364.
[63] GitHub. n.d. GitHub - cjhutto/vaderSentiment: VADER Sentiment Analysis. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.. [online] Available at: <https://github.com/cjhutto/vaderSentiment>.

[64] Buntain, C. and Golbeck, J., 2014, April. Identifying social roles in reddit using network structure. In Proceedings of the 23rd international conference on world wide web (pp. 615-620).
[65] FXStreet. n.d. Reddit vs Twitter : Are they playing or helping investors?. [online] Available at: <https://www.fxstreet.com/analysis/reddit-vs-twitter-are-they-playing-or-helping-investors-202107122058>.
[66] R. Adusumilli, “Predicting stock prices using a keras lstm model,” NikolaNews

Appendix

Below are the files related to the code in order to achieve the project:

1. Code included in Jupyter files:
    1. Explanatory Data Analysis of three datasets – EDA.ipynb file
    2. To test different sentiment analysers and pre-processing techniques for best performance – Diss_Tester_Dataset.ipynb file
    3. Pre-processing and sentiment analysis of all three datasets – Sentiment_Analysis.ipynb file
    4. Merging text and numeric data for News – Work 2 (News).ipynb file
    5. Merging text and numeric data for Twitter – Work 2 (Twitter).ipynb file
    6. Merging text and numeric data for Reddit – Work 2 (Reddit).ipynb file
    7. Performance criteria for all three data sources for Logistic Regression – Work 4 (Logistic Regression).ipynb file
    8. Performance criteria for all three data sources for SVM – Work 4 (SVM).ipynb file
    9. Performance criteria for all three data sources for XGB – Work 4 (XGB).ipynb file
2. Original text datasets:
    1. News.csv
    2. Tweets3.csv
    3. Reddit.csv
    4. Stock_data.csv – Tester dataset
3. Original numeric datasets:
    1. SP_News.csv
    2. SP500_Tweets.csv
    3. SP500_Reddit.csv
4. Related csv files for the above ipynb files:
    1. final_polarity_data.csv
    2. sp_news_cleaned_data.csv
    3. news_sent_data.csv
    4. sp_tweets_cleaned_data
    5. tweets_sent_data
    6. sp_reddit_cleaned_data
    7. reddit_sent_data.csv

Figure 26 Gnatt Chart: