SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA AND ITS EFFECT ON STOCK MARKET TRENDS

ABIMBOLA STEPHEN FALODU

Candidate No: 713189

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Business Analytics at Aston University, Birmingham

On this day, 5th September 2022

Supervisor: Dr Shahin Ashkiani

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to the almighty God for his ever-enduring mercies upon my life and for preserving me till this moment to obtain this degree. I would also thank my loving mother and sisters for their prayers, encouragement and support in every step of the way on this journey. I would also love to express my profound gratitude to my supervisor, Dr Shahin Ashkiani, for his guidance, assistance and support during this project. Finally, my sincere appreciation must go to Dr Viktor Pekar, who provided support at a critical stage of this dissertation.

Table of Contents

ABSTRACT

Stock market forecasting is a classic phenomenon that has historically piqued the interest of investors looking to secure a return on their investment and researchers across domains aiming to comprehend the market and devise more sturdy prediction techniques. Stock markets are influenced by various factors such as News, government policies, social media and microblogs. Advancements in machine learning and natural language processing mean we can leverage on the enormous pool of user-generated textual content on the Internet as a precious source of information to reflect investor sentiment and predict stock price trends potentially. This study conducts a sentiment analysis on social media data and analyses the impact of sentiment in forecasting stock trends. We selected five of the top twenty companies trading on the London stock exchange across various industries; Astra Zeneca, BP, GSK, HSBC and Vodafone. A state-of-the-art pre-trained roBERTa-base model was used for sentiment classification to investigate whether sentiment variables from Twitter contain any predictive power for forecasting stock trends. We utilise the Support Vector Classifier (SVC) with RBF kernel and Random Forest Classifier (RFC) for comparison. Our results show that only BP is influenced by social media sentiment for the selected equities. We recorded an accuracy of 74.5% when incorporating the sentiment variable compared to 69.6% without it, while other equities showed improved performance without the sentiment variable. Using the RFC, we assessed the feature importance. We identified the exponentially weighted moving average as the most relevant variable in classifying the stock trend, while the sentiment and trend variables were the least important. Hence implying that sentiments derived from twitter do not contain any predictive power for classifying the trends of selected equities. The SVC showed consistent results across both cases yielding the highest accuracy of 82.4%, precision of 82.3%, recall of 83% and F1 measure of 82.5% for GSK equities therefore highly recommended for stock trend predictions.

INTRODUCTION

"Sentiment analysis (SA) is the algorithmic classification of users' evaluations of a brand (positive, negative, or neutral) in posts and comments". ("Sentiment analysis - Oxford Reference") Advancement in data generation, transmission and storage capabilities coupled with the social media boom means for an unprecedented time in recorded human history, there exists a gargantuan pool of opinionated data. This data pool fuelled research interest in SA and its applications across industries. One area that has gained increased application for SA is in finance where stakeholders have sought methods to improve traditional prediction techniques.

Efficient Market Hypotheses proposed by Fama, 1970 opines that it should be impossible to predict stock prices or returns as markets will reflect all available information and react accordingly. However, market commodities are traded by humans who make decisions based on mood and psychology (Audrino, Sigrist and Ballinari, 2020 cited Daniel, Hirshleifer, and Teoh, 2002; Tseng, 2006; Johnson and Taversky, 1983). Prediction of stocks and other securities however remains a daunting task with different researchers reporting varying degrees of success trying to incorporate sentiment data into stock predictions. (Nguyen, Shirai and Velcin, 2015 cited Antweiler and Frank, 2004; Tumarkin and Whitelaw, 2001; Bollen, Mao and Zeng, 2011)

The goal of this dissertation is to carry out aspect-based SA on tweets regarding specific companies to ascertain whether the social mood of the population regarding these companies and their equities is an important variable that affects their stock trends. Variety of techniques have been utilised in SA tasks ranging from lexicon rule-based approach to more advanced supervised and unsupervised machine learning approach.

In this dissertation, we utilise a hybrid model that uses NLP transformers architecture to perform sentiment analysis and assign labels. As a proof of concept, we utilise the labelled data done via transformers to carry out SA using the Long short-term memory (LSTM) network famed for its ability to capture long term dependencies and showed consistent results in recent research hence we adopt the same approach. This approach was then combined with classic Support Vector Machine with radial based function (RBF) kernel to predict the future stock trend. The aim of the experiments conducted is to access the impact of the social mood/sentiment on the Equity market of the selected companies.

The major contributions of this dissertation are summarised as follows:

- To investigate if social mood or sentiment about a company and its traded securities has any impact/significance in the price movement and trends.
- We propose a hybrid approach for stock trend prediction using a combination of sentiment analysis (previous day average sentiment) and a support vector machine with RBF kernel to classify the future stock trends.
- We carry out our sentiment analysis using Timelms, a roBERTa-base model based on NLP transformers architecture and pre-trained up to date with ~124m tweets.
- We conducted a case study using equities from five of the top twenty traded companies on the London stock exchange (LSE) market, aggregating sentiment about these companies from Twitter to predict the commodities' future trend to measure our efficacy proposed approach and test the hypotheses.

The remainder of the dissertation is organised as follows: Chapter 2 reviews existing literature, theory and historical background of SA, Stock prediction, techniques and highlights related works on SA of Twitter Data and stock prediction utilising some of the proposed approaches. Section 3 highlights the experiment and analysis carried out by application of python packages. For illustration, section 4 presents the results of the analysis and assess the performance of the proposed method and model. Finally, section 5 concludes the dissertation with remarks, a discussion of results and recommendation for future work.

LITERATURE REVIEW

Sentiment analysis

"Sentiment analysis (SA) or opinion mining (OM) is the computational study of people's opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu, 2020). Historically, the focus of sentiment analysis was opinion polarity i.e., analysing and detecting a person's exact opinion (positive, neutral, or negative) towards something which might explain why SA and OM have been used interchangeably (Mäntylä, Graziotin and Kuutila, 2018). Ideally, SA and OM are fungible as they both express a mutual meaning. However, some researchers have stated that both have slightly distinct ideas as they originate from different communities. On the one hand, SA finds the sentiment in a text and analyses it, while OM extracts and examines opinions about an entity. (Tsytsarau and Palpanas, 2012).

Interest in the opinion of others is a classic phenomenon; medieval Greek leaders were invested in their subjects' opinions to detect internal dissent (Mäntylä, Graziotin and Kuutila, 2018 cited Thorley, 2004). Despite the long, storied history of linguistics and Natural Language Processing (NLP), there existed miniature research on people's opinions and sentiments before the 21st century. However, the consolidation of the world wide web and the advent of social media, blogging and discussion forums in the middle of the century has resulted in an exponential amount of user-generated content which means that for the first time in recorded human history, there is a huge volume of opinionated data recorded in digital forms (Zhang, Wang, and Liu, 2018). This data explosion, alongside a wide range of applications across domains from computer science to marketing, finance and political science, has made sentiment analysis one of the fastest growing areas of research and underlined its importance to business and society (Zhang et al., 2018; Mäntylä, et al., 2018).

Sentiment analysis can be appraised as a classification process where each category represents a sentiment. (Prabowo and Thelwall, 2009; Medhat, Hassan and Korashy, 2014). A generic approach to a typical SA process is illustrated in figure 1.

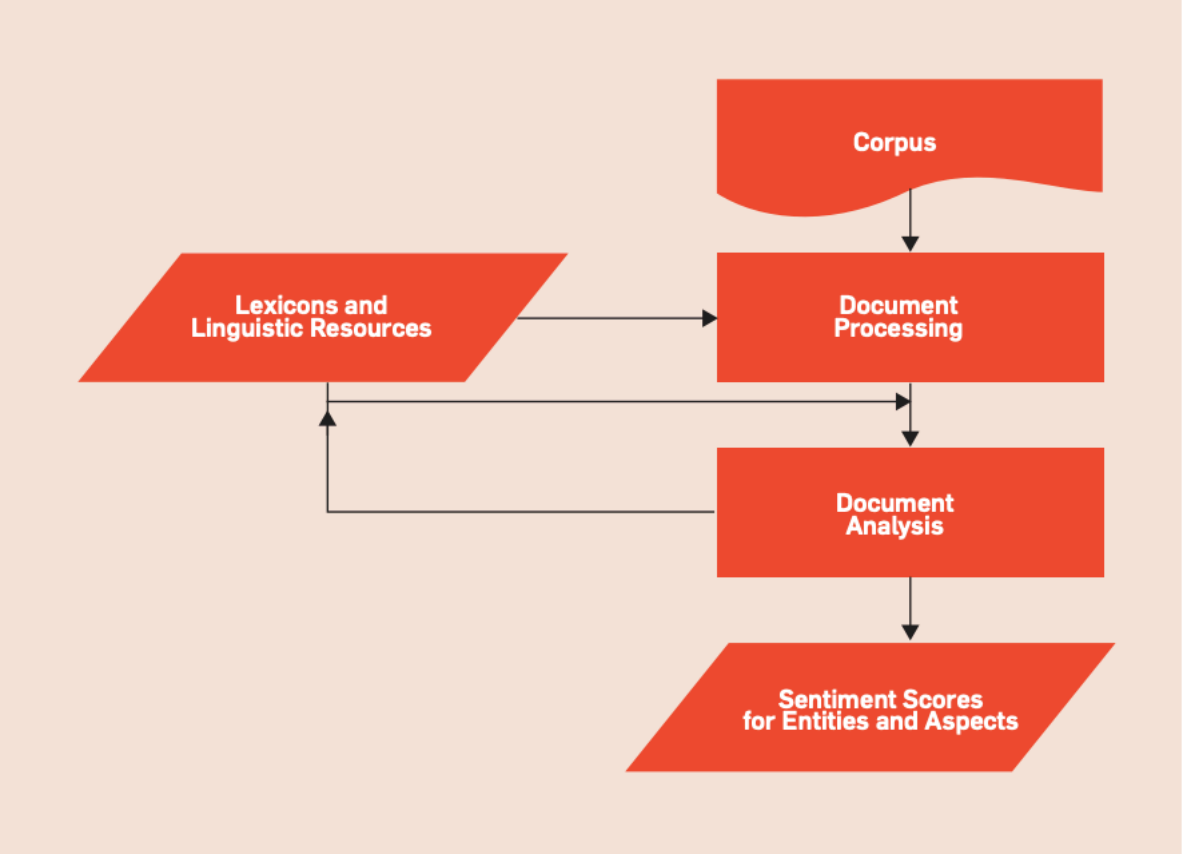


Figure 1: Architecture of a generic SA system. (Feldman, R., 2013; 84)

Medhat, Hassan and Korashy, 2014 also illustrated a typical SA task using an example of a product review shown in figure 2 below:

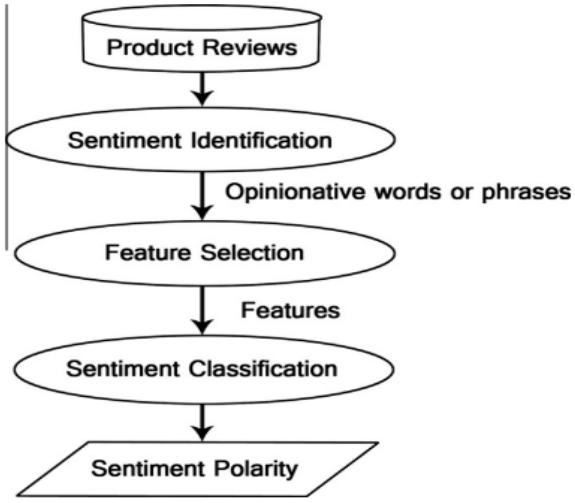


Figure 2: Sentiment analysis process on product reviews. (Medhat, W. et al., 2014; 1094)

In figure 1 above, the system receives as input a corpus of documents in any format (Word, PDF, HTML, XML, etc.), content of the corpus is transformed to text and pre-processed with a range of linguistic methods in preparation for analysis. The document analysis module is the

main component of the system where linguistic resources are utilised to create sentiment annotations with the pre-processed document, attachment of these annotations may be to whole documents, individual sentences or specific aspects of entities depending on the level of sentiment analysis. The output of the system are the sentiment annotations which may be rendered to the user with a range of visualisation tools. (Feldman, 2013).
Based on granularity, classification in SA is broadly grouped into 3 main levels which are document-level, sentence-level, and aspect-level sentiment analysis.

Document-level sentiment analysis

The quest at this level is to classify the sentiment expressed by the entirety of the document as either positive or negative (Liu, 2012 cited Pang, Lee and Vaithyanathan, 2002; Turney, 2002). Majority of existing research assumes that an opinionated document d (e.g., a product review) expresses opinions on a single object o and these opinions are from a single holder h. (Liu, 2012; Liu, 2010; Feldman, 2013), this is also commonly referred to as document level sentiment classification as the whole document is considered a basic information unit hence inapplicable to documents evaluating or comparing multiple entities (Liu, 2012; Medhat, Hassan and Korashy, 2014). Two main approaches exist for this level of SA: supervised and unsupervised learnings, the approaches are perfectly captured in Medhat, Hassan and Korashy, 2014 shown below.

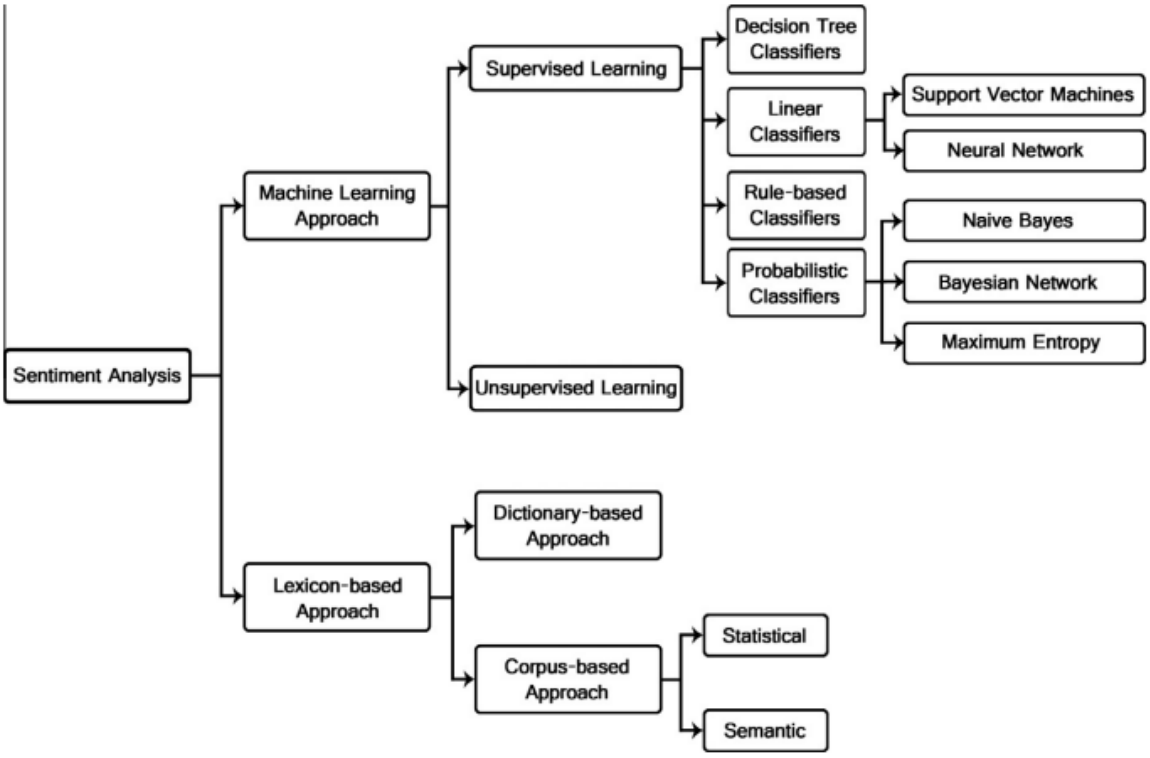


Figure 3: Sentiment classification techniques (Medhat, W. et al., 2014; 1095)

Sentence-level sentiment analysis

At this level of SA, the intent is to classify the prevailing sentiment expressed in a sentence into positive, neutral, or negative classes. Existing literature equate this level of SA to subjectivity classification where it is determined whether the sentence is subjective or objective, a second subtask is then performed to determine if the sentiment expressed is positive or negative. (Feldman, 2013; Liu, 2012; Medhat, Hassan and Korashy, 2014). Majority of SA performed at this level utilise supervised learning approaches however the potential complexities of compound sentences that may contain multiple opinions, subjective and objective clauses mean that efforts at this level is insufficient. This highlights the need for another level of SA capable of capturing the finer-grained details in the content of the corpus (Liu, 2012; Medhat, Hassan and Korashy, 2014; Feldman, 2013 all cited Wilson, et al., 2005).

Aspect-level sentiment analysis

Feldman, 2013 perfectly captures the essence of this level of SA in his definition "Aspect-based sentiment analysis is the research problem that focuses on the recognition of all sentiment expressions within a given document and the aspects to which they refer." For majority of practical situations, the document usually contains discussion(s) about multiple entities with multiple attributes combined with different opinions about individual attributes. This exasperated the need for a level of SA that would observe the finer details in the corpus. (Liu, 2012; Feldman, 2013). This level of SA also commonly referred to as feature-based SA focuses mainly on discerning sentiments on entities and their attributes; it is mostly useful when analysing product reviews where users give a range of feedback on different attributes of an entity.

Social media, data generation and sentiment analysis.

Due to its nature, where users post messages in real-time, which may express their thoughts on numerous issues, conversations about current trends and events, complaints, and exhibit feelings about regularly used items/products, social media has become a source of a variety of information. As stated previously from Zhang, Wang, and Liu, 2018, the consolidation of the world wide web and the advent of social media, blogging and discussion forums in the middle of the 21st century has resulted in an exponential amount of user-generated content which means that for the first time in recorded human history, there is a huge volume of opinionated data recorded in digital forms. Data generated from social media apps e.g., Twitter, Facebook

etc are mostly semi-structured data as they are a mixture of structured and unstructured data also there is a sheer amount of data points that could be collected on a social media post. A tweet for instance could be a record with features time of the tweet, user_id, tweet location, text (actual tweet content), reply to the user, number of retweets, number of likes etc. The actual tweet content for example, is unstructured data in textual format while the time of tweet (created_at) is a structured piece of data that is clear, precise, and easily mapped to a column "time" in a table. (May, 2020). Data collection from social media apps/sites is usually carried out by API calls or queries into their database (paid access) or web scraping. However the latter can be a tedious collection technique when attempting to collect a colossal volume. Analysis of this data is of utmost value to businesses, investors, and governments as they could provide valuable insights that potentially alter society or be useful to make predictions.

Related work on SA of twitter data

Barbosa and Feng, 2010 was one of the early efforts for SA of Twitter data. They trained a classifier by introducing polarity predictions from three websites as noise in the data, then proceeded to tune and test with 1000 manually labelled tweets respectively. They suggested combining information like the preceding polarity of words and part of speech (POS) tagging of words with syntactic elements of tweets like retweet, hashtags, links, and exclamation points. However, they make no indication as to what method was used to collect and sample the test data hence a high likelihood of bias. This approach was extended by Agarwal et al., 2011 by utilizing real-valued prior polarity in conjunction with a combination of prior polarity and POS. Their findings demonstrate that the features that combine the prior polarity of words with their POS are the ones that improve classifier performance the most. At the same time, the syntactic elements of tweets are only marginally significant..

Agarwal et al., 2011 highlight the early and recent results of work by various researchers on sentiment analysis of Twitter data and the levels that SA had been handled. n highlighting these results, they observed the performance of SVM over other classifiers while showing different researchers' contrasting observations of the feature space characteristics. In their own work, they tried to eliminate bias by reporting results on manually annotated data as they had pointed out that previous researchers performed analysis on data collected via search queries hence biased. For the feature space, they presented features that significantly outperformed the previous work's unigram baseline and explored a different data representation method.

Medhat, Hassan and Korashy, 2014 on the one hand, also note the performance of SVM and Naïve Bayes over other classifiers hence their frequent usage for SA tasks. However, on the

other hand, they opined that the usage of social media and microblogs as a data source for SA requires further deep analysis and accentuates the importance of considering text context and user preferences in SA tasks.

Stenqvist & Lönnö, 2017 attempted to predict bitcoin (BTC) price fluctuations by critically analysing BTC-related tweets for sentiment fluctuations that could indicate future price movement. They utilised a lexicon rule-based approach called Valence Aware Dictionary and Sentiment Reasoner (VADER) and polarity classification. They found that Aggregating tweet sentiments over a 30 min period with four shifts forward, and a sentiment change threshold of 2.2%, yielded a 79% accuracy although there was a static threshold for the model, limitation in domain specific lexicon of VADER not accounting for financial terms and the data used was limited to a timespan of 1 month which ultimately resulted in no indication of success in their method.

In a substantial effort to improve the performance of SA of tweets containing fuzzy sentiments, Phan et al., 2020 proposed a novel approach (feature ensemble method) that incorporated lexical, word-type, semantic, position and sentiment polarity of words with a Convolutional Neural Network (CNN) to classify sentiments of tweets into five sets from negative to strong positive. The proposed model of feature ensemble and CNN yielded better performance in SA of fuzzy tweets in terms of F1 score over the baseline; this was a strong point of this piece of work as it highlighted the effectiveness of feature ensemble models over traditional methods however SA was performed on fuzzy tweets only and no considerations were made for the impact of slangs, sarcasm, and domain specific lingo.

Stock market prediction

A daily task negotiated globally is the decision to buy or sell shares in the stock market. This is a fascinating research task across disciplines as the question about whether the stock market can be predicted accurately is one that have preoccupied the minds of investors, economists, and market stakeholders for decades. Shah, Isah and Zulkernine, 2019 likened the stock market to behaving like a voting machine in the short term where sentiments can drive market fluctuations which in turn results in disconnect between the price and the true value of a company's shares and like a weighing machine in the longer term as a company's fundamentals ultimately cause the value and market price of its shares to converge.

In a rapidly evolving industrial world, accurate forecasting of the market is crucial for portfolio management hence advancement in stock market prediction techniques is of utmost significance to investors and various stakeholders more importantly because of the

complexities in analysing the stock market trends due to intrinsic noise in the data and large volatility in relation to market trends. These complexities stem from the fact that both non-economic and economic elements are considered while analysing the behaviour of stock trends, also stock prices adapt to several market events, quarterly earnings releases, and shifting consumer trends. Traders rely on different technical indicators that depend on equities gathered daily by them. Even though these indicators are used to analyse stock returns, it is difficult to predict daily, and weekly market movements are considered a significant challenge for increasing production widely. (Gandhmal and Kumar, 2019 cited Yeh, Huang and Lee, 2011; Ticknor, 2013).

Nguyen, Shirai and Velcin, 2015 highlight the two distinct trading philosophies for stock market prediction which are fundamental and technical analysis. Fundamental analysis relies on the economic forces of demand and supply that cause the stock price to either fluctuate or remain constant, whereas technical analysis concentrates on the study of market actions and behaviour patterns using charts. Hu et al., 2015 based fundamental analysis on three essential aspects which are (a) macroeconomic analysis such as Gross Domestic Product (GDP) and Consumer Price Index (CPI) which analyses the effect of the macroeconomic landscape on the subsequent profit of a company, (b) industry analysis which estimates the value of the company based on industry status and prospect, such as analysis of billings of industry upper stream entities and (c) company analysis which analyses the current operation and financial status of a company to evaluate its internal value mostly by critical examination of its financial reports. Additionally, they classified the areas of technical analysis into sentiment, flow-of-funds, raw data, trend, momentum, volume, cycle, and volatility. Sentiment is a representation of how different market participants behave and analysis of these indicators is grounded on the hypothesis that different types of investors exhibit different behaviours at the main market turning points. The flow of funds is a technical indicator used to investigate different investors' financial situations to pre-evaluate their propensity for purchasing and selling stocks. Based on this information, techniques like short squeeze can then be implemented. Stock price series and price patterns like K-line diagrams and bar charts are examples of raw data. The former is commonly used for time series analysis or trend judgment combined with other indicators while the latter generally suggests price patterns can reflect the changes in market sentiment which affect short movements of stocks. Trend and momentum meanwhile are price-based indicators, trend is used to track stock price trends, while momentum assesses the speed of price change and determines whether a trend reversal in stock price is going to take place. The volume serves as a basis for forecasting changes in stock prices and is a measure of the gusto of buyers and

sellers for investing. The cycle meanwhile is based on the classic theory that stock prices fluctuate on a cyclical basis over a period of more than ten years, with shorter cycles every few days or weeks. Finally, volatility is frequently used to assess risk, determine the degree of support and resistance, and explore the price fluctuation range of stocks. (Hu et al., 2015 cited Bodie et al., 2005; Yahoo Finance, 2013; Fidelity Mutual Fund, 2013; CFA Institue,2013; Colby, 2002; Goldman Sachs, 2013; NASDAQ, 2013; Pring, 2002; Wikipedia, 2013; MTA knowledge base, 2014).

Stock market price prediction remains a difficult phenomenon and many ideas that have been developed over the years that either attempt to explain the nature of stock markets or whether it is possible to beat the markets. Recent developments in stock analysis and prediction are best characterised into four (4) broad categories which are statistical, pattern recognition and sentiment analysis. Most of these categories are included in the larger category of technical analysis, however some machine learning techniques also incorporate the larger categories of technical analysis with fundamental analysis methods to forecast stock markets. (Shah, Isah and Zulkernine, 2019). A taxonomy of these approaches is shown in the figure 4 below.

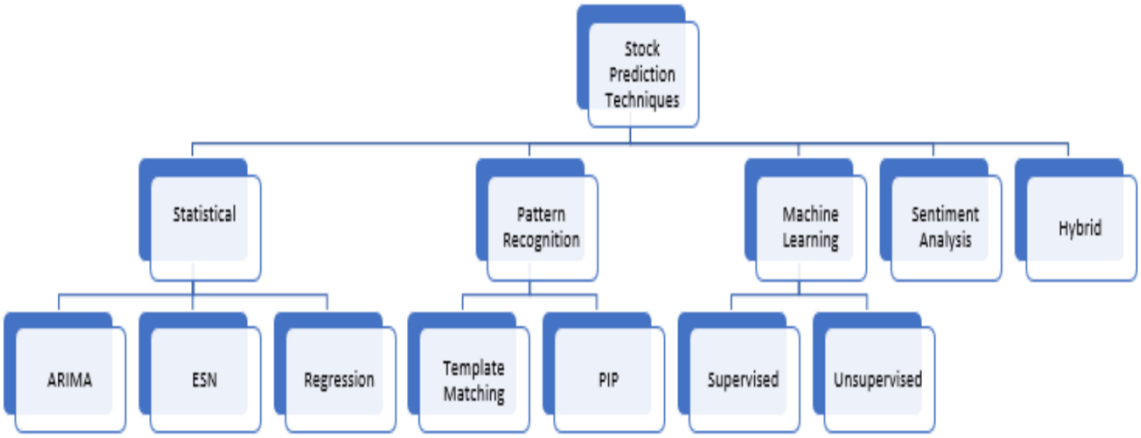


Figure 4: Taxonomy of stock prediction techniques (Shah, Isah and Zulkernine, 2019; 3)

Based on the taxonomy shown in the figure 4 above, we can notice a 5th category which is the hybrid approach; this approach combines multiple different approaches for an improved performance. E.g., a hybrid approach of statistical and machine learning or like the scope of the current work that will utilize a hybrid approach of sentiment analysis and machine learning.

Related work on methodology

Nguyen, Shirai and Velcin, 2015 was a notable piece of work on using SA to predict stock price movement as it was the first to show the effectiveness of incorporating SA on a large-scale test data. They formulated a model with a novel approach that incorporated sentiment of specific topics of the considered companies using the SVM with linear kernel as the classification model and were able to achieve results that showed a 2% increase over historical methods. Limitations of their effort however include the fact that the proposed model could only predict the direction of price movement (up or down) and not the degree of movement, also they had specified beforehand the number of topics and sentiment for both methods used.

Ruan, Durresi and Alfantoukh, 2018 examined the correlation between user reputation built by trust networks and abnormal stock returns using SentiStrength, an existing SA tool to perform tweet-level SA for each tweet and Pearson correlation coefficient to model the relationship between Twitter Sentiment Valence (TSV) and stock returns. They were able to establish the relationship between user reputation, TSV and stock returns as they found out that TSV reflect abnormal stock returns better when user reputation is accounted for. The data they used though was very limited in terms of time span as they collected data for a 7-month period, also SentiStrength used for SA was not particularly designed for financial text analysis and the algorithm of choice could have been better to model a non-linear relationship between variables.

Derakhshan and Beigy, 2019 extended Nguyen, Shirai and Velcin's work by examining the limitations of model-based opinion mining and introduced an alternative. They utilized the same Latent Dirichlet Allocation technique and extended it by adding part-of-speech (POS) tags to separate words (LDA-POS), they also extended and used the same dataset used by Nguyen by adding comments on 18 stocks for an additional year. As opposed to using the SVM used by Nguyen, they used a two-layered Neural Network with one hidden layer as the classification model. The result of this yielded an improved accuracy on existing work using the same dataset although the topic selection was also limited to 50 and the granularity of stock price prediction was still coarse (up or down) like Nguyen's work.

Ren et al., 2019 explored investor sentiment from news data using SA and day of the week effect into an SVM algorithm to forecast stock movement direction. They collected data for SA primarily on stock forums and discovered that combining sentiment features with stock market data potentially improves model performance and adding a stop loss order strategy

with this approach can reduce investor risks. This conforms to existing literature however the time interval for the data collected was limited.

Audrino, Sigrist and Ballinari, 2020 analysed the impact of sentiment and attention variables on stock market volatility using a state-of-the-art sentiment classification technique with a penalized regression framework. They utilised a novel dataset that combined daily economic and sentiment variables over a 5-year period and discovered that attention and sentiment variables can improve volatility forecasts significantly when controlling for a large set of economic and financial variables. However, they noted that the magnitude of improvements is relatively small from economic POV, and the effects are short-term.

Khan et al., 2020 assessed the role of external factors such as financial news and social media in stock prediction using several ML classification and regression algorithms such as Random Forest (RF), K-Nearest Neighbour (KNN), SVM etc. They used the Stanford NLP package for SA generating sentiment scores of 0-4 for tweets ranking from more negative to more positive. To improve their model performance and quality of prediction, they carried out spam tweet reduction using a dataset of 380 spam and ham tweets, performed feature selection (FS) and principal component analysis (PCA). Ultimately, they found out that social media has more effect on stock prediction by day 9 while financial news shows effects on day 8 and 9. They also discovered that some stock stocks (New York and Red Hat) are more difficult to predict than others and that some stocks (New York and IBM) are more influenced by social media while others (London and Microsoft) were more influenced by financial news. In terms of the best performing algorithm, the RF classifier showed the most consistency and the highest accuracy of 83.22% was achieved using its ensemble. However, their technique lacked any method to determine stock related keywords, and this may have affected their results. It must be stated that results from this work was consistent with Audrino et al., 2020 that combining sentiment from social media and financial news can increase overall accuracy of many classifiers after day 3.

Nti, Adekoya and Weyori, 2020 was another more recent effort that tried to build on some of the ideas from Nguyen, Shirai and Velcin. They critically examined the association between public sentiments found in web news, SM, forums, and predictability of future stock price movement. They were particularly focused on increasing model performance by improving data quality before being fed into the model, they amalgamated relevant related data from various sources and their model; an Artificial Neural Network (ANN) was able to achieve a stunning accuracy of (70.66 – 77.12%) using the combined dataset. However, the data that was collected for this work was small due to limited social media data in the region under

consideration hence making it hard to rely on the prediction obtained for larger and more robust datasets.

A review of existing literature shows a variety of SA techniques being used however we note the immense popularity of the SVM classifier over others, also small layered neural networks with few hidden networks are popular as well. However, two recent pieces of work show some interesting observations. Wong, 2021 examined two different models used in evaluating sentiment scores of twitter data and its predictive performance for bitcoin price movement. They compared the long short-term memory (LSTM) ANN model with the Naïve Bayes and find that the LSTM model yields better results. While this effort was able to highlight the unique advantages of both approaches considered, limited data was used to test the model and the performance of both approaches were below par for the specified purpose especially for synthetic data.

The second recent piece by Critien et al., 2022 tries to build on modern knowledge regarding bitcoin prediction and predict both price direction and magnitude. They utilised a lexicon rule-based approach (VADER) with LSTM for SA and CNN as the classification algorithm. Ultimately, they proved competitive results can be achieved with Bi-LSTM with a 1day lag period using 7 diff lagged features. However, limitations in data utilised and possible data shrinkage leading to fewer records could have tampered with their results.

To summarise, many research efforts have addressed predicting stock market using a variety of SA and supervised ML hybrid techniques. However, there has been polarity on whether sentiment variables are significant in stock prediction. This motivates us to adopt the ideas of Khan et al., 2020 to predict the stock future trend and tweaking it by:

- Adopting NLP Transformers architecture as a means to carry out initial SA and label the tweets.
- Initialising the model by creating features that are technical indicators incorporating moving averages.
- Utilising the Support Vector classifier with radial based function kernel to classify the future stock trend.
- We also initialise the model without the sentiment variable and compare the performance with the model incorporating sentiment variables to test our hypotheses using the selected companies.
- Finally, we repeat the stock trend classification task with Random Forest and compare results with those generated using the SVC with RBF kernel.

METHODOLOGY

In this section, I detail the steps performed in my proposed framework to analyse the sentiment present in the tweets and subsequently use these sentiments to predict the stock direction of related stocks. The proposed model will utilise a hybrid approach of supervised and unsupervised machine learning techniques to carry out sentiment analysis on the tweets then use a supervised approach for stock prediction. A pipeline of steps involved in our proposed framework is shown below:

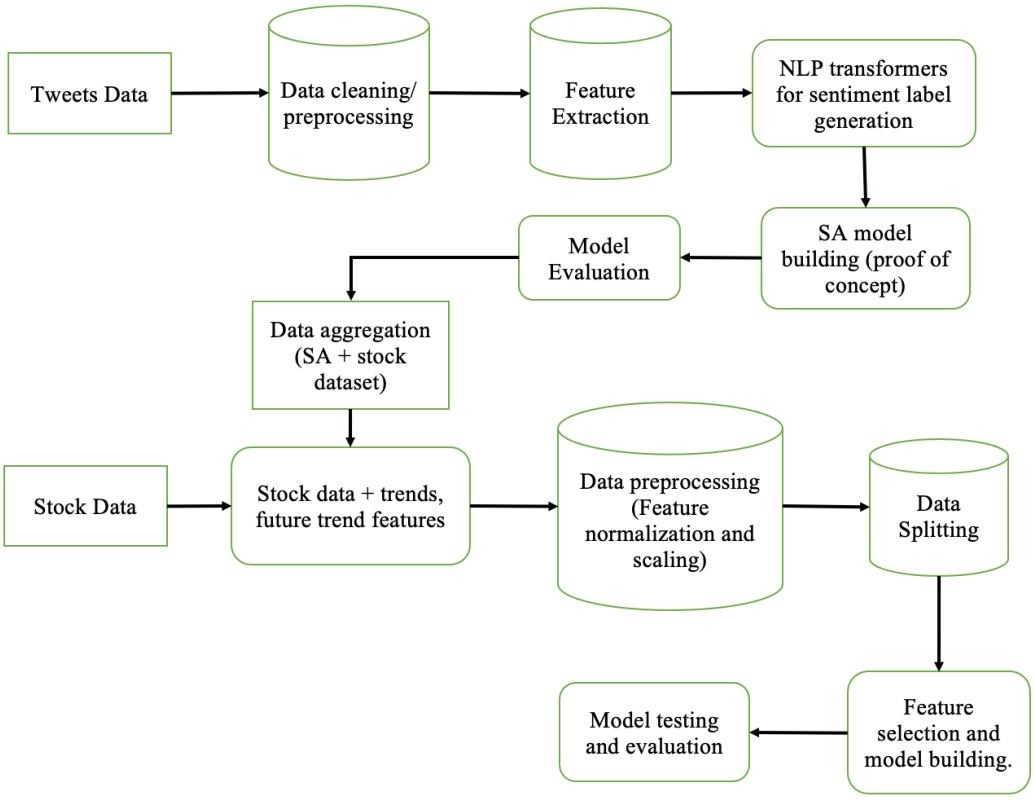


Figure 5: Flow chart of steps in the proposed framework for stock prediction using twitter data.

SENTIMENT ANALYSIS

The proposed model for sentiment analysis will utilize a combination of techniques. To train the model, the dataset needs to be annotated with a sentiment target label which will be done by utilising NLP transformers. The target labels generated by NLP transformers was utilised in the final dataset due to the dependability, recency and robustness of the model built on over 100m tweets of training data. The essence of using this approach is to avoid having to manually label the dataset, we then use the target (sentiment) labels generated by transformers to train a model using Long Short Term Memory technique, a supervised learning approach based on Recurrent Neural Network architecture as a proof of concept.

Social Media Data collection

We consider five different UK companies: Astra Zeneca, HSBC, BP, Vodafone, and GlaxoSmithKline (GSK) all listed on the London Stock Exchange (LSE). The choice of these companies is motivated by the fact that they are among the top companies by market capitalisation on the LSE and sentiments might have different influence on the stock trend depending on the type of stock under consideration. For this reason, we select companies from different industries (e.g., Astra Zeneca and GSK are pharmaceutical companies with Astra Zeneca growing in popularity in the aftermath of the pandemic, HSBC is a financial institution, BP is an energy company and Vodafone is a telecommunications company. Tweets in English language mentioning these companies or their cashtags along with the words stock or financial market are streamed from the Twitter API using the Tweepy streaming client. The timeframe for our analysis was between January 2021 and April 2022. We found a total of 14,030 English tweets that contain any of the companies with the words (stock or financial) market for the 16-month period.

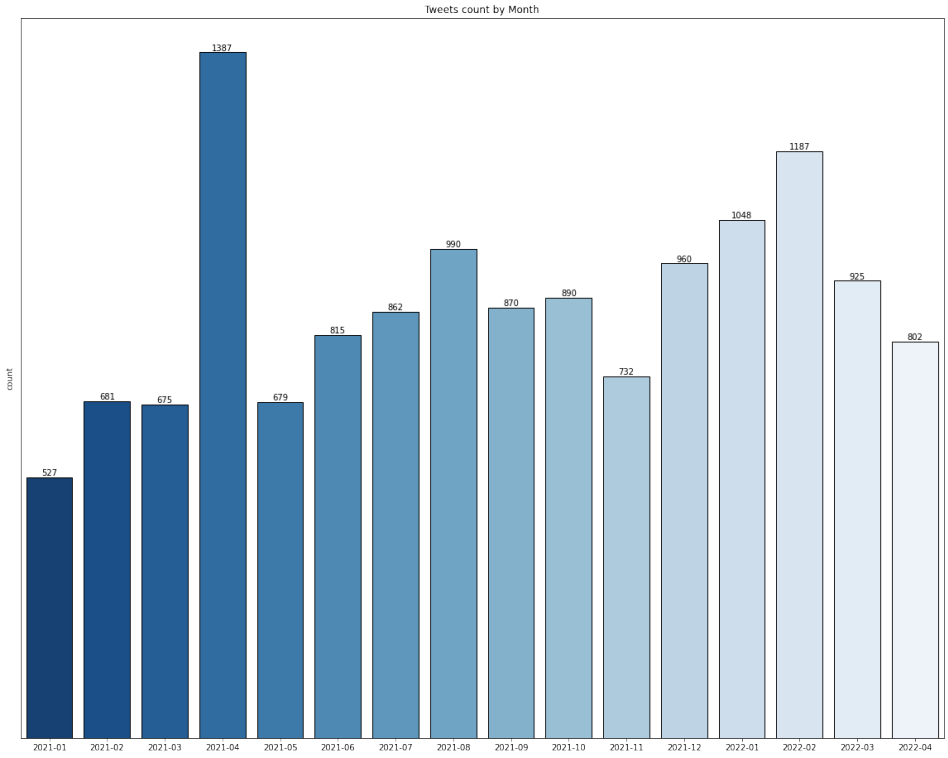


Figure 6: Number of monthly tweets about the selected companies.

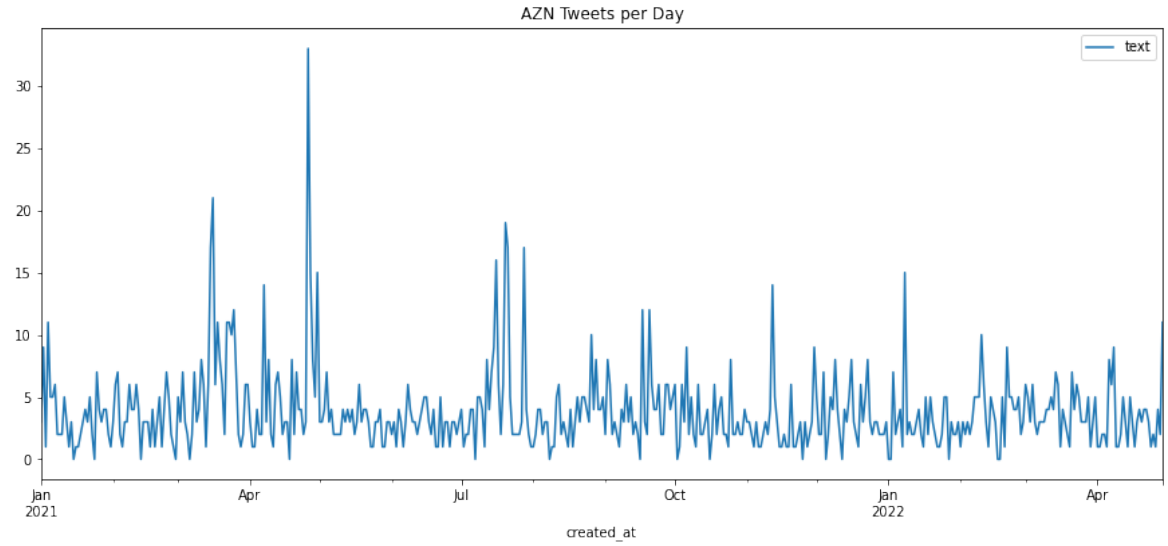


Figure 7: AstraZeneca tweets per day.

The bulk of tweets about the related companies was in April 2021 with 1387 related tweets with the least being in January 2021 with 527 tweets. BP was the most talked about company with 4475 related tweets with Astra Zeneca having the least with 1849 related tweets.

Data cleaning/pre-processing

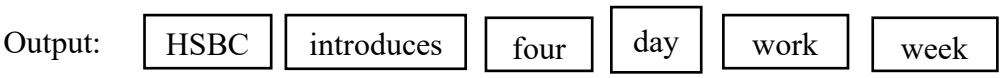
The streamed tweets are in their raw format as retrieved from the Twitter database which means some of them would contain unwanted characters, slangs, links, hashtags and emojis. These would need to be removed as they are not contributing to the goal of sentiment detection in the tweets by performing the following steps:

1. Tokenisation
2. Removal of duplicate tweets, HTML and other tags, username tag (@), cashtags and URLs.
3. Stop words removal.
4. Stemming

Word Tokenisation

For a given set of character sequence and a defined corpus unit, the task of tokenisation is described as the process of converting character sequence or streams into smaller units, known as tokens, while also removing punctuation characters. Tokens are often commonly referred to as terms or words, however, a token is an instance of a series of characters in a corpus that are grouped together as a meaningful semantic unit for processing. (Manning, Raghavan and Schütze, 2009). A simple example of tokenisation is described thus:

Input: “HSBC introduces four-day work week”



Tokenisation was carried out by using the keras library word tokenizer.

Stop words removal

Stop words are extremely common words that are usually of negligible semantic weighting in helping select documents that matches the need in various use cases. (Manning, Raghavan and Schütze, 2009). Lexical processing of index terms involves the elimination of stop words. Although stop words play a crucial grammatical role in language, such as in the formation of phrases, they do not contribute to the overall semantic content of a document in a keyword-based representation. Such words are commonly used in corpus content regardless of topic or domain and thus have no topical specificity. A typical example of stop words is articles and prepositions. Removing stop words reduces the number of index terms. A particular drawback of stop word elimination is that it can sometimes result in the removal of some practical terms, for instance, the stop word A in Vitamin A hence making it impossible to search a corpus correctly. Siddiqui and Tiwary (2008, pp 258-259). In this task, stop words are removed using the Natural Language Toolkit (NLTK) package. NLTK has inbuilt a stop list for different languages and since we're dealing with English tweets, the English stop list is downloaded and used for stopwords removal. NLTK currently has by default 179 stop words in the English stop list. A list of these words is shown below:

i	me	my	myself	we	our	ours
ourselves	you	you're	you've	you'll	you'd	your
yours	yourself	yourselves	he	him	his	himself
she	she's	her	hers	herself	it	it's
its	itself	they	them	their	theirs	themselves
what	which	who	whom	this	that	that'll
these	those	am	is	are	was	were
be	been	being	have	has	had	having
do	does	did	doing	a	an	the
and	but	if	or	because	as	until
while	of	at	by	for	with	about
against	between	into	through	during	before	after
above	below	to	from	up	down	in
out	on	off	over	under	again	further
then	once	here	there	when	where	why
how	all	any	both	each	few	more
most	other	some	such	no	nor	not
only	own	same	so	than	too	very
s	t	can	will	just	don	don't
should	should've	now	d	ll	m	o
re	ve	y	ain	aren	aren't	couldn
couldn't	didn	didn't	doesn	doesn't	hadn	hadn't
hasn	hasn't	haven	haven't	isn	isn't	ma
mightn	mightn't	mustn	mustn't	needn	needn't	shan
shan't	shouldn	shouldn't	wasn	wasn't	weren	weren't

Table 1: NLTK English stop words list.

WordCloud

After removing stopwords, we can visualise the most frequent words in the corpus in a word cloud as shown below. We can see that the prevalent words in the dataset are highly relevant to the companies and topic under consideration

Figure 8: WordCloud representation of the most frequent words in the corpus.

Stemming

For a given corpus, different forms of a word are utilised for grammatical reasons. For instance, organize, organizes, organized and organizing. Also, there are families of related words with similar meanings, such as democracy, democratic, and democratization, words usually differ in their flavour. Stemming is typically a primitive heuristic procedure that cuts off the ends of words and frequently involves removing derivational affixes. This is done to reduce inflectional and occasionally derivationally related variants of a word to a common base form. Manning, Raghavan and Schütze (2009, pp32). Stemming just like lemmatization is a normalization process that aims to normalize variants of a word albeit in a crude manner

by removing affixes from the words to reduce them to their stems. Siddiqui and Tiwary (2008, pp259) e.g., the words am, are, and is, are all reduced to the same word stem, be. However, unlike stemming, Lemmatization uses vocabulary and morphological analysis of words to ensure that the resulting base form of a word is a known word in a dictionary although a downside to this is that the additional checking process makes the lemmatization slower than the stemming. Bird, Klein, and Loper (2009, pp 107-108). For this task, we perform stemming by using one of the most widely used algorithms developed by Martin Porter in 1980.

Porter's Stemmer

Multiple sources describe the Porter stemmer as empirically adequate, a good choice for text indexing and great for supporting search using an alternative form of words. (Bird, Klein, and Loper, 2009; Manning, Raghavan and Schütze, 2009; Siddiqui and Tiwary, 2008).
Porter's algorithm consists of five successive word reduction steps and condition/action pairs. Actions take the form of rewrite rules while conditions may be on stems, suffix, or rules. Stem conditions take either of the following form:

i. $m = 0, 1, \text{or } 2.$
ii. Stem contains or ends with (pattern)

where *m*, the measure, is the number of vowel-consonant (VC) sequence. For example, for the word 'sea', m = 0 as the number of vowel-consonant sequence is 0, whereas for the word "astronaut", m = 3 ('as', 'on', and 'ut'). The patterns take the form:

*<*x*>	stem ends with a given letter x
υ	stem contains a vowel
*d	stem ends in a double consonant
*o	stem ends in a C-V-C sequence, where the concluding consonant (C) is not w, x, or y.

Suffix condition takes the form 'current_suffix = = pattern' and rule conditions take the form 'rule was used'. Action rules are of the form 'old suffix → new_suffix'.
The five steps involved in Porter stemming algorithm as detailed by Siddiqui and Tiwary (2008, pp392-395) are described in the appendix section.

Word Embeddings

Word embedding is a process which entails the transformation of words in a vocabulary to vectors of continuous real numbers which may encode linguistic regularities and patterns. (e.g., the word "HAT"! (…, 0.15, …, 0.23, …, 0.41, …)). The method often entails embedding from a high-dimensional sparse vector space (for example, a vector space for one-hot encoding, where each word has a dimension) to a lower-dimensional dense vector space. The embedding vector's many dimensions each correspond to a word's latent feature. (Zhang, Wang, and Liu, 2017)

Feature Extraction

Feature Extraction (FE) is perhaps the most challenging task in SA (Siqueira and Barros, 2011). However, most reviewed literature did not focus on feature extraction but instead emphasised the sentiment classification task. For a typical SA task, various feature sets are extracted to investigate the contribution of each type of feature. Agarwal and Mittal, 2016 details the four initial types of essential features that are extracted, namely.

- Unigrams (F1 feature set)
- Bigrams (F2 feature set)
- Bi-tagged (F3 feature set) and
- Dependency parsing tree-based features (F4 feature set).

In this project, feature extraction was carried out by creating word embeddings using one of the most popular word encoders, Word2Vec which has been shown to perform significantly better than classic well-known models such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). (Mikolov, et al., 2013 cited Mikolov, Yih and Zweig, 2013; Zhila, et al., 2013).

Word2Vec

In 2013, a Google research team led by Tomas Mikolov introduced the breakthrough model for word representation called Word2Vec by publishing two papers (Mikolov, et al., 2013a; Mikolov, et al., 2013b) with model architectures for computing continuous vector representations of words by using the unsupervised approach: Continuous Bag-of-Words (CBOW) and Continuous Skip-gram Model, using either hierarchical SoftMax or negative sampling. The word2vec architecture is shown in figure 9 below:

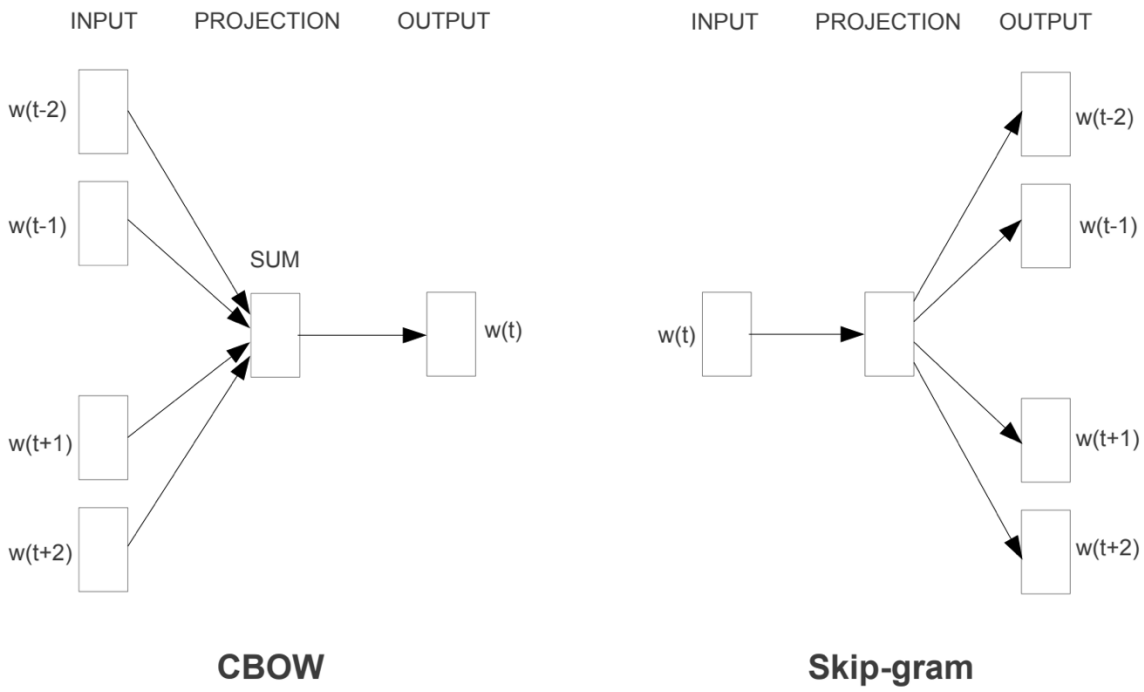


Figure 9: Word2Vec CBOW and Skip-gram architecture (Mikolov, et al., 2013; 5).

The models are trained using stochastic gradient descent, and backpropagation and the training complexity is proportional to:

$$O = E \times T \times Q \quad \text{----------} \quad (1)$$

where T is the total number of words in the training set, E is the number of training epochs, and Q is further defined for each model architecture. A common choice is E = 3 − 50 and T up to one billion. The CBOW architecture is like a feedforward Neural Net Language Model (NNLM), where the projection layer is shared across all words and the non-linear hidden layer is removed; hence, all words get projected into the same position by their vectors being averaged. However, CBOW, unlike the standard bag-of-words model, uses a continuous distributed representation of context to predict the current word. Training complexity is subsequently defined by:

$$Q = N \times D + D \times \log_2(V) \quad \text{----------} \quad (2)$$

The Skip-Gram architecture while similar to CBOW functions slightly differently as it seeks to maximise the classification of a word based on another word in the same sentence rather than relying on context to predict the current word. More precisely, a log-linear classifier with a continuous projection layer takes the current word as input and predicts words within a specific range before and after the word. (Mikolov, et al., 2013a)

During training, the Skip-gram model intends to find word representations that are potentially useful in predicting surrounding words in a sentence or document. Conventionally, given a sequence of training words *w1, w2, w3, . . . , wT*, the model seeks to maximize the average log probability.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \quad \text{----------} \quad (3)$$

where c is the size of training context. The basic Skip-gram formulation defines $p(w_{t+j}|w_t)$ using the SoftMax function:

$$p(w_o|w_{\mathrm{I}}) = \frac{\exp(\upsilon'_{w_o}{}^{\intercal}\upsilon_{w_{\mathrm{I}}})}{\sum_{w=1}^{W}\exp(\upsilon'_{w}{}^{\intercal}\upsilon_{w_{\mathrm{I}}})} \quad \text{----------} \quad (4)$$

However, due to computational cost, the skip-gram formulation modifies the SoftMax to a more computationally efficient approximation, the Hierarchical SoftMax, which defines $p(w_o|w_{\mathrm{I}})$ as follows:

$$p(w_o|w_{\mathrm{I}}) = \prod_{j=1}^{L(w)-1}\sigma([n(w,j+1) = ch(n(w,j))].\upsilon'_{n(w,j)}{}^{\intercal}\upsilon_{w_{\mathrm{I}}}) \quad \text{--- (5)}$$

where $\sigma(x) = 1/(1+\exp(-\mathrm{x}))$. (Mikolov, et al., 2013b).

Dataset Labelling

The tweets data collected from the twitter database do not contain any sentiment target labels. Due to the size of the dataset and the amount of human effort that will be required to label it, initial labelling was done by using a pretrained SA model based on transformers architecture.

Timelms

A pretrained roBERTa-base model, Timelms proposed by Loureiro et al, 2022 based on transformers architecture was used to carry out SA and assign sentiment labels to all tweets in the dataset. The capacity of the pre-trained word and sentence embeddings to maintain the semantics and syntax of the individual words in phrases contributes to their successful performance on NLP tasks. For context such as social media where the topic of discourse is constantly changing, diachronic specialization is therefore imperative which is a huge value proposition in selecting this model. As a case in point, a model built before the Covid-19 pandemic would be oblivious to terms like covid, lockdown and mask used frequently in recent communication. The model was trained with 124m tweets and comparison with other benchmarks is shown below:

Model	Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL(TE)	Referen
BERTweet	33.4	79.3	**56.4**	**82.1**	79.5	73.4	71.2	**67.9**	BERTwe
TimeLMs-2021	**34.0**	**80.2**	55.1	64.5	**82.2**	**73.7**	**72.9**	66.2	TimeLM
RoBERTa-Retrained	31.4	78.5	52.3	61.7	80.5	72.8	69.3	65.2	TweetE\
RoBERTa-Base	30.9	76.1	46.6	59.7	79.5	71.3	68	61.3	TweetE\
RoBERTa-Twitter	29.3	72.0	49.9	65.4	77.1	69.1	66.7	61.4	TweetE\
FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1	TweetE\
LSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5	TweetE\
SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5	TweetE\

Figure 10: Model comparison with benchmarks from TweetEval (Barbieri et al., 2020)

From the figure above, the selected model has the second highest overall accuracy score for all tasks and the highest for sentiment classification which is what we require from the model.

NLP Transformers architecture

A transformer represents an architecture that utilises two models: encoder and decoder to transform one sequence into another. Vaswani et al., 2017 proposed a novel transformer architecture solely based on multi-headed self-attention mechanisms shown below:

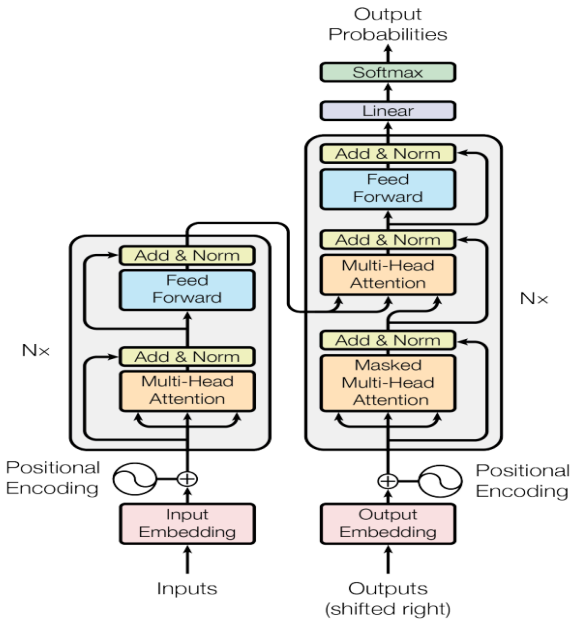


Figure 11: Transformers model architecture (Vaswani et al., 2017; 3)

The self-attention mechanism chosen for the model architecture is due to three main factors: parallelization, computational complexity and the ability to absorb long range dependencies between words in a sentence sequence, all crucial in creating context-based embeddings. The

feed forward and multi-head attention layers are the main building blocks (inputs) for the module as shown in the preceding figure. They make use of the scaled dot-product and multi-head attention function shown thus:

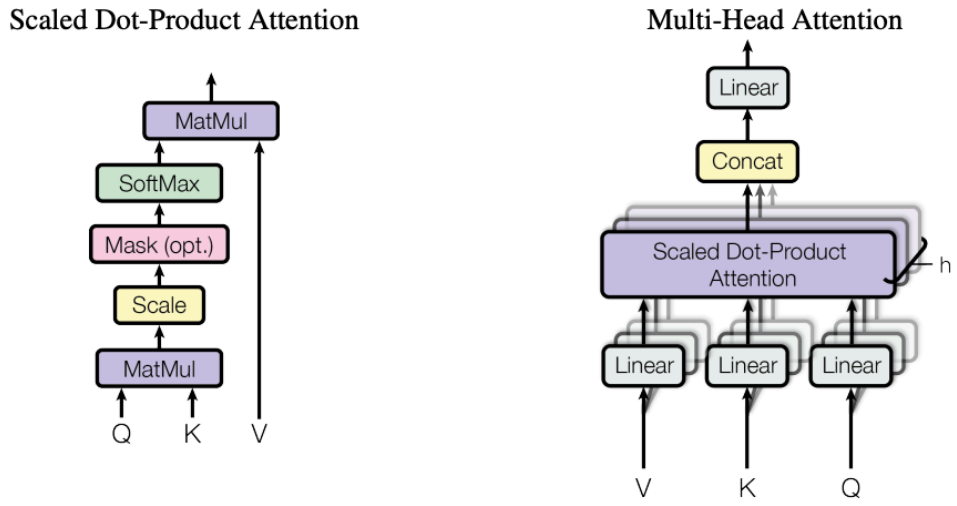


Figure 12: Scaled dot product (left) and multi-head (right) attentions. (Vaswani et al., 2017; 4)

The scaled dot product attention mechanism is described by the equation below:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \text{ -------- } (6)$$

Where Q represents the matrix of a set of queries packed together (vector representation of words in a sequence), K and V are the respective keys and values of all word vector representation and d_k is the dimensionality of the keys and query. The SoftMax function applied is used to obtain weights on all the values.

The multi-head attention mechanism embodies several attention layers running in parallel and empowers the model to incorporate data from different representation subspaces at different positions. The multi-head attention is specified by the equation below:

$$MultiHead(Q, K, V) = concat(head_1, \dots, head_h)W^O \text{ ------- } (7)$$

Where each head is evaluated by:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \text{ ---------- } (8)$$

The variables W_i^Q, W_i^K, W_i^V *and* W^O are parameter matrices that serve as projections used by the model in the learning process.

Long Short-Term Memory

After the pre-processing has been done and we have now successfully assigned sentiment labelled to the tweets, we proceed to build the sentiment analysis model. A lot of recent works have emphasized the efficiency and great results obtained by models setup with a LSTM architecture (Xing, Cambria, and Welsch, 2017; Shah, Isah and Zulkernine, 2018; Nabipour et al., 2020; Wong, 2021; Critien, Gatt and Ellul, 2022) hence we have adopted the same approach. We have initialised a sequential model which LSTM is a good fit as its commonly applied with an embedding layer using a vocabulary size equal to the total amount of word tokens in all the tweets with 100 memory units. Developed in 1997 by Hochreiter and Schmidhuber, the LSTM network is a particular Recurrent Neural Network (RNN) adept at studying long-term dependencies. RNNs are usually a chain of repeating modules, as shown in figure 13.

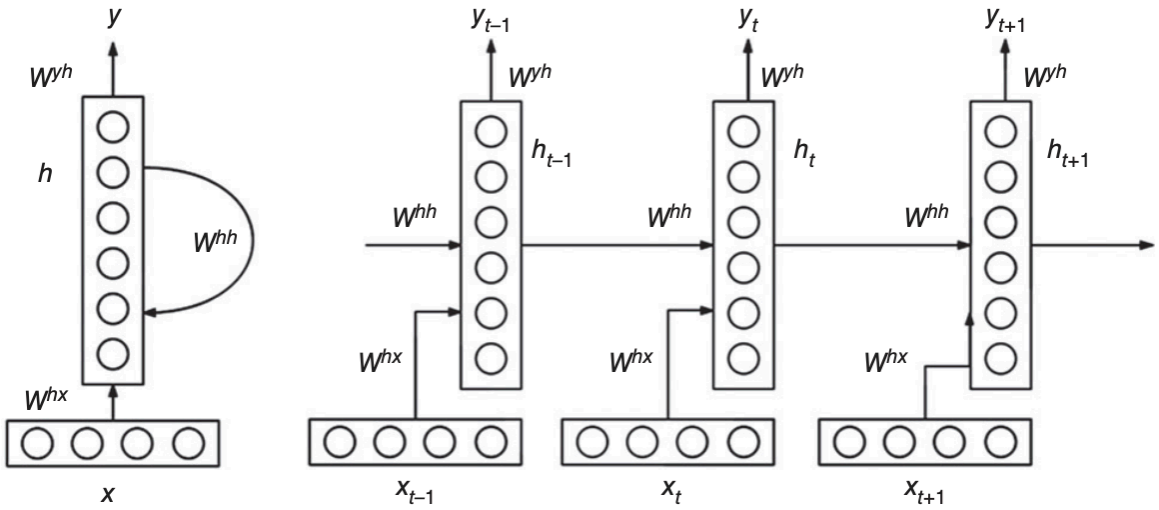


Figure 13: Recurrent Neural Network architecture (Zhang, Wang, and Liu, 2017; 6)

In standard RNNs, this repeating module typically has a simple structure. However, the repeating module for LSTM is more complicated. As opposed to having a single neural network layer, four layers interact specially with two states: hidden and cell states. Figure 14 shows an example of LSTM.

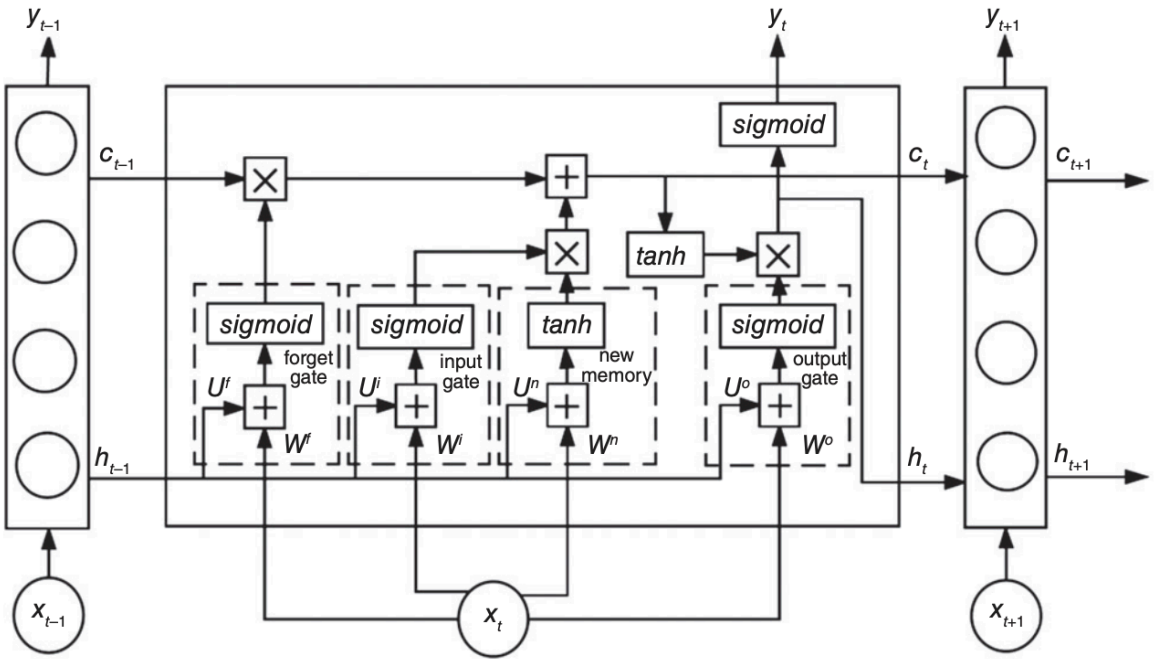


Figure 14: LSTM network (Zhang, Wang, and Liu, 2017; 8)

At time step t, LSTM first decides what information to dump from the cell state. A sigmoid function/layer σ, called the "forget gate" is responsible for this decision-making process, the function takes h_{t-1} (output from the previously hidden layer) and x_t (current input), and outputs a number in [0, 1], where 1 means "completely keep" and 0 means "completely dump" in Equation (10).

$$f_t = \sigma(w^f x_t + u^f h_{t-1}) \quad \text{----------} \quad (10)$$

Then LSTM determines the new data to be stored in the cell state. Two steps are involved here. First, a sigmoid function/layer, called the "input gate" as Equation (11), decides which values LSTM will update. Next, a tanh function/layer creates a vector of new candidate values $\widetilde{C_t}$, subsequently added to the cell state. LSTM combines these two and updates the state.

$$i_t = \sigma(w^i x_t + u^i h_{t-1}) \quad \text{----------} \quad (11)$$

$$\widetilde{C_t} = \tanh(w^n x_t + u^n h_{t-1}) \quad \text{----------} \quad (12)$$

The old cell state C_{t-1} is updated to a new one C_t as Equation (13). The forget gate f_t can control the gradient passes through it and allow for explicit "memory" deletion and update, which alleviates the vanishing or exploding gradient problems in standard RNN.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad \text{----------} \quad (13)$$

Finally, LSTM decides the output based on the cell state. First, LSTM runs a sigmoid layer to decide which parts of the cell state to output in Equation (14), this is the "output gate." Then,

the cell state goes through the tanh function and is multiplied by the output of the sigmoid gate, so that LSTM only outputs the parts it decides to as in (15).

$$\sigma_t = \sigma(w^o x_t + u^o h_{t-1}) \quad \text{----------} \quad (14)$$

$$h_t = o_t * \tanh(C_t) \quad \text{----------} \quad (15)$$

(Zhang, Wang, and Liu, 2017)

Embedding layer

The embedding layer learns a word embedding for all the words in the dataset. It takes as input the total word tokens from all tweets in the dataset and outputs a vector space in which words will be embedded. Word embeddings have been created using Word2Vec with a specified vector length of 300 to hold our word coordinates.

Final tweet sentiment

After model building, validation and tuning, a sentiment score is generated for tweet records in the testing set which is in turn used to assign target sentiment labels is as thus:

$$\text{Sentiment target label} = \begin{cases} Positive & if \quad s \geq 0.7 \\ Neutral & if \ 0.4 < s < 0.7 \\ Negative & if \quad s \leq 0.4 \end{cases}$$

where s is the generated sentiment score.

Average Daily Sentiment

In order to aggregate the sentiment and stock price dataset, tweets about each company were first aggregated on a daily level and an average daily sentiment score is generated by adding the sentiment scores of all tweets in a day and dividing by the number of tweets in a day. The generated average is hence used to assign the final sentiment about the company for that particular day as thus:

$$\text{Day Sentiment} = \begin{cases} Positive & if \quad \bar{s} > 1.0 \\ Neutral & if \quad \bar{s} = 1.0 \\ Negative & if \quad \bar{s} < 1.0 \end{cases}$$

Stock Market Prediction

Price data

Stock historical price data for the selected companies were collected from Yahoo Finance for the same timeframe of January 2021 to April 2022. For the selected timeframe, price data for

each transaction date is collected with features Open, High, Low, Close, Volume, and Adjusted Close totalling 335 trading days.

Feature	Description
Open	Opening stock price for specified date
High	Maximum trading stock price
Low	Minimum trading stock price
Close	Closing stock price
Adjusted close	Closing stock price adjusted for splits and dividends
Volume	Amount of stock traded during the day

Table 2: Stock data features.

Feature extraction from stock data

We can extract relevant features to predict the stock trend, as shown by Khan et al., 2020. The stock trend and future trend are extracted from existing features in the stock data. Values of these features are nominal and specified by positive, neutral, or negative. The value of the trend feature is calculated by subtracting the opening stock price from the closing price on a specific date and the selection criteria are given in the following equation:

$$Trend_d = \begin{cases} Positive \ \ if \ P_c - P_o > 0 \\ Neutral \ \ if \ P_c - P_o = 0 \\ Negative \ if \ P_c - P_o < 0 \end{cases}$$

Where $Trend_d$ defines the trend while $P_c \ and \ P_o$ denote the closing and opening stock price respectively for the day in view. The Future trend feature is the target variable to be predicted and is defined by the difference between a stock's current day closing price and closing price after n days. A positive difference connotes a positive trend after n days, a zero difference indicates a neutral future trend, and a negative difference implies that the stock future trend will shift downwards after n days. The following equation determines the future trend after n days:

$$Future\_trend_d = \begin{cases} Positive \ \ if \ P_{tc} - P_{nc} > 0 \\ Neutral \ \ if \ P_{tc} - P_{nc} = 0 \\ Negative \ if \ P_{tc} - P_{nc} < 0 \end{cases}$$

where P_{tc} denotes the stock closing price today and P_{nc} represents the closing price after n days. We have selected the value of $n = 5$, meaning that we will identify the stock's future trend for up to 5 days and thus examine the impact of the tweets on predictions five days into the future.

Technical Indicators

We create two new features (technical indicators) for the stock dataset that are representative of moving averages. Since we're predicting the stock trend five days into the future, we compute simple and exponential moving averages of the closing price for the selected stocks for a 5-day shift.

Simple Moving Average

The simple moving average of a stock is an arithmetic moving average computed by averaging prices over a time shift (Hayes, 2022). In this case for example, prices are averaged over a 5-day period by simply taking the arithmetic mean of price data over the days in consideration.

Exponential Moving Average

The exponential moving average of a stock, also referred to as the exponential weighted moving average is a kind of moving average that places huge significance and weights on more recent data points (Chen, 2022). The EMA is defined by:

$$EMA_{today} = \left(Value_{today} \times \left(\frac{smoothing}{1+Days}\right)\right) + EMA_{yesterday} \times \left(1 - \left(\frac{smoothing}{1+Days}\right)\right)$$

The charts (figure 15 – 19) show the trend in closing price, simple and exponential moving averages for the 5 stock tickers under consideration.

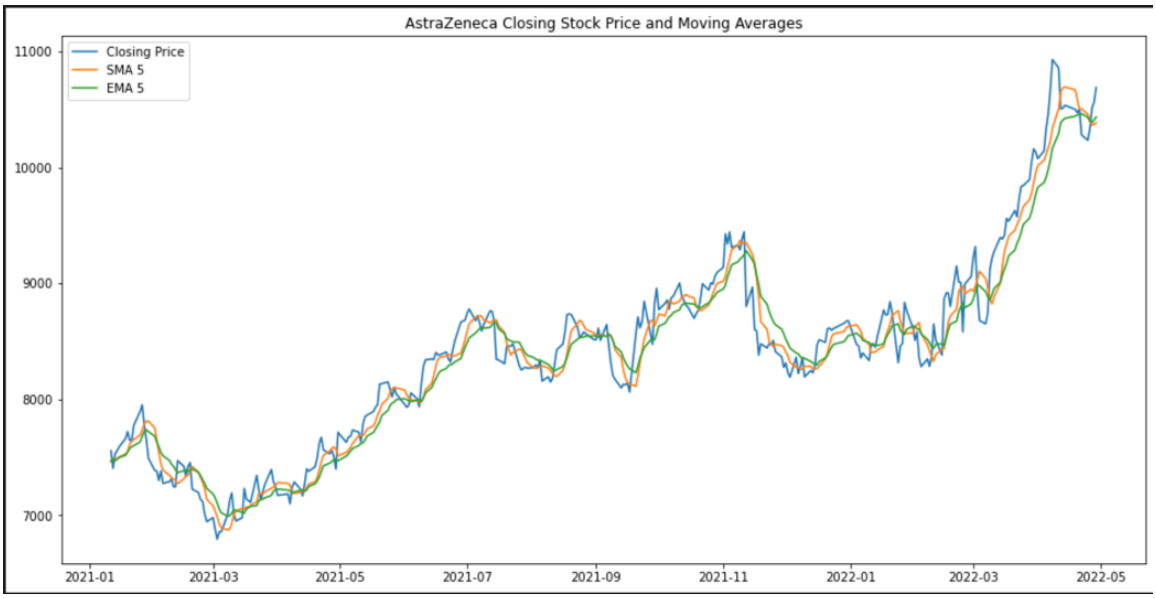


Figure 15: AstraZeneca closing price and moving averages.

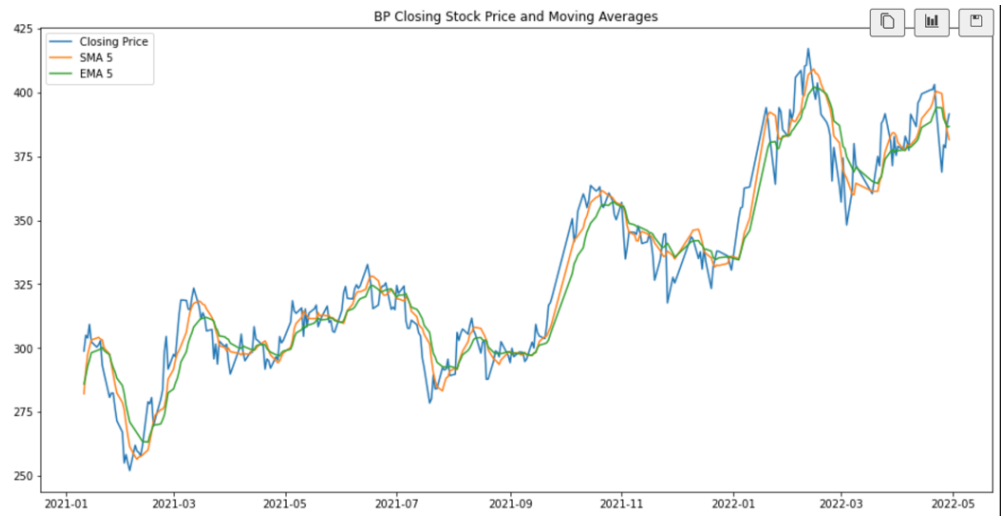


Figure 16: BP closing price and moving averages.

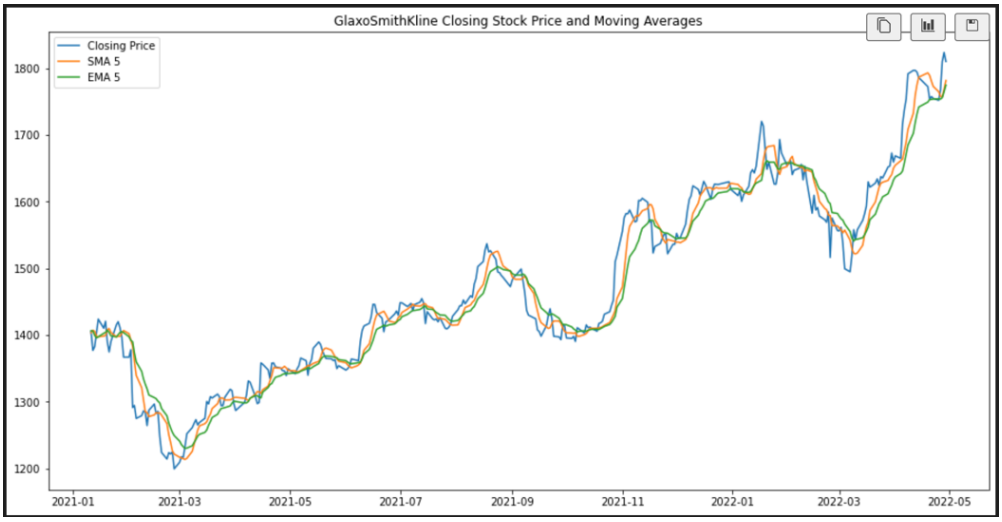


Figure 17: GlaxoSmithKline closing price and moving averages.

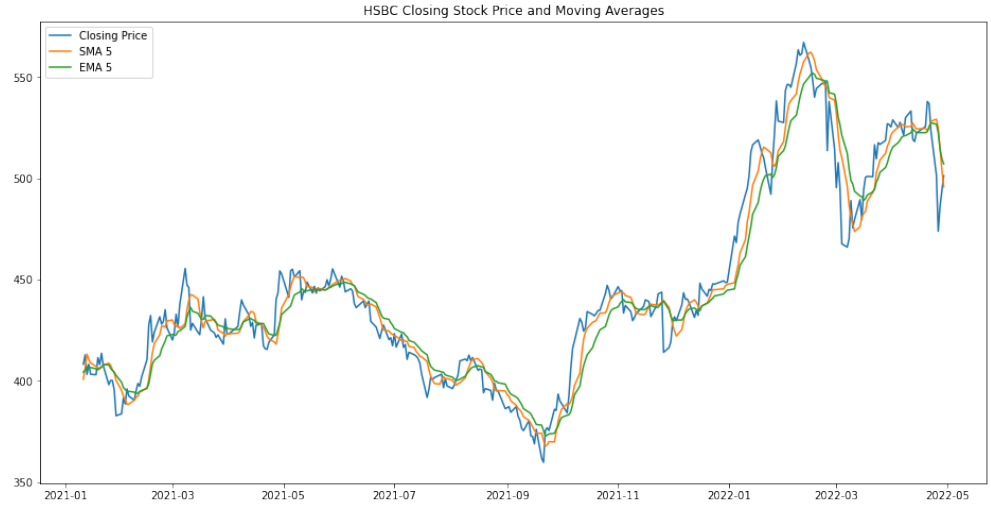


Figure 18: HSBC closing price and moving averages.

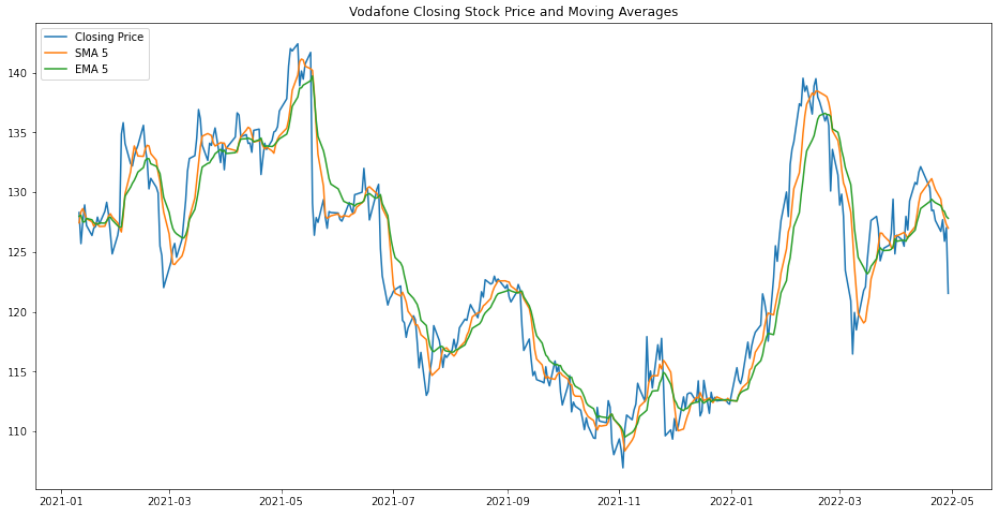


Figure 19: Vodafone closing price and moving averages.

Final dataset

The final data set for the task is created by incorporating the sentiment feature from the tweets as an average daily sentiment into the stock data file. Sentiments from the previous day are aggregated with current data to incorporate the futuristic, reactive nature of the market. A view of the final dataset for the stock market forecasting is shown below:

created_at	Open	High	Low	Close	Adj Close	Volume	sentiment	trend	future_trend	sma_5	ema_5
2021-01-11	411.000000	414.200012	404.600006	408.200012	375.869995	16564968	2.0	Negative	Positive	400.779999	404.266983
2021-01-12	410.049988	414.700012	410.049988	412.799988	380.105682	14890862	1.0	Positive	Positive	406.350000	405.252554
2021-01-13	404.299988	406.549988	401.549988	403.149994	371.219971	19141900	1.0	Negative	Negative	413.189996	406.997419
2021-01-14	404.549988	410.200012	404.049988	408.200012	375.869995	30898467	2.0	Positive	Negative	410.589996	406.161857
2021-01-15	406.049988	409.799988	400.000000	403.250000	371.312073	19699147	1.0	Negative	Negative	408.820001	406.583210
2021-01-18	401.149994	404.649994	399.750000	403.000000	371.081848	8564292	2.0	Positive	Negative	407.120001	405.920672
2021-01-19	414.200012	418.649994	408.149994	411.450012	378.862579	31443356	1.0	Negative	Negative	406.079999	405.358190
2021-01-20	411.200012	412.149994	404.250000	408.000000	375.685852	24043501	1.0	Negative	Positive	405.810004	406.501751
2021-01-21	417.200012	421.100006	412.250000	413.600006	380.842316	24604870	1.0	Negative	Positive	406.780005	406.777204
2021-01-22	407.950012	412.049988	405.600006	407.350006	375.087341	17135283	1.0	Negative	Positive	407.860004	408.010386

Figure 20: Final dataset inclusive of Technical indicators (Vodafone).

Data pre-processing

The final dataset does not require much pre-processing. The future trend is assigned as the target variable, and we can scale the other features using the *scikit-learn* StandardScaler class. However, as is usually the case with machine learning models, it is imperative to normalise or scale the features within a fixed range to avoid features with larger values unjustly interfering with the model or resulting in bias. The scaled dataset is shown below:

	Open	High	Low	Close	Adj Close	Volume	sentiment	sma_5	ema_5	encoded_trend	future_trend
0	-0.303165	-0.353871	-0.387562	-0.457590	-0.867542	-0.679846	1.521200	-0.799940	-0.660635	-1.213223	Positive
1	-0.346810	-0.331272	-0.147798	-0.249455	-0.668919	-0.841231	-0.143795	-0.539920	-0.611258	0.824251	Positive
2	-0.610977	-0.699632	-0.521743	-0.686088	-1.085594	-0.431429	-0.143795	-0.220614	-0.523841	-1.213223	Negative
3	-0.599492	-0.534660	-0.411759	-0.457590	-0.867542	0.701910	1.521200	-0.341988	-0.565703	0.824251	Negative
4	-0.530578	-0.552740	-0.589933	-0.681563	-1.081275	-0.377710	-0.143795	-0.424615	-0.544593	-1.213223	Negative
5	-0.755694	-0.785507	-0.600931	-0.692875	-1.092071	-1.451115	1.521200	-0.503975	-0.577786	0.824251	Negative
6	-0.156150	-0.152743	-0.231386	-0.310537	-0.727212	0.754438	-0.143795	-0.552524	-0.605966	-1.213223	Negative
7	-0.293976	-0.446526	-0.402960	-0.466640	-0.876177	0.041088	-0.143795	-0.565128	-0.548674	-1.213223	Positive
8	-0.018324	-0.042009	-0.051012	-0.213257	-0.634376	0.095204	-0.143795	-0.519847	-0.534874	-1.213223	Positive
9	-0.443288	-0.451046	-0.343569	-0.496050	-0.904243	-0.624868	-0.143795	-0.469430	-0.473092	-1.213223	Positive

Figure 21: Final Scaled dataset (Vodafone).

Time series Analysis

Data collected on securities like stocks is observed at many points in time hence forming a time series. McKinney (2018, p289). We can show the illustration for the daily opening and closing prices for the timeframe under consideration in the figures below:

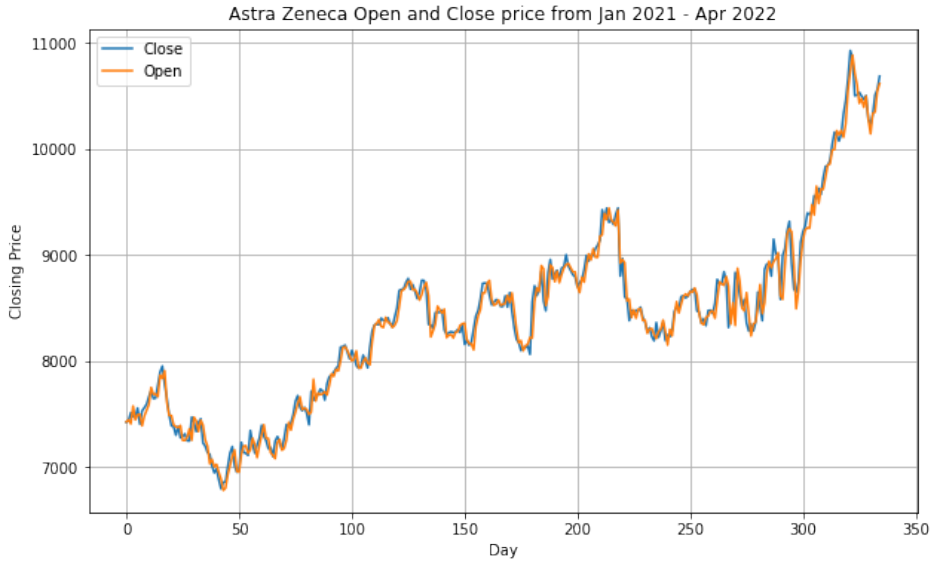


Figure 22: Astra Zeneca Equities Opening and Closing price from Jan 2021 to Apr 2022.

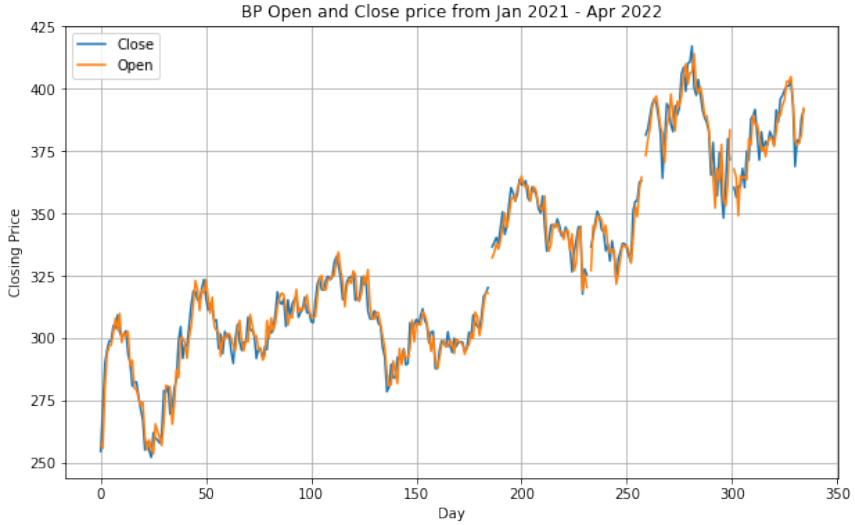


Figure 23: BP Equities Opening and Closing prices from Jan 2021 to Apr 2022.

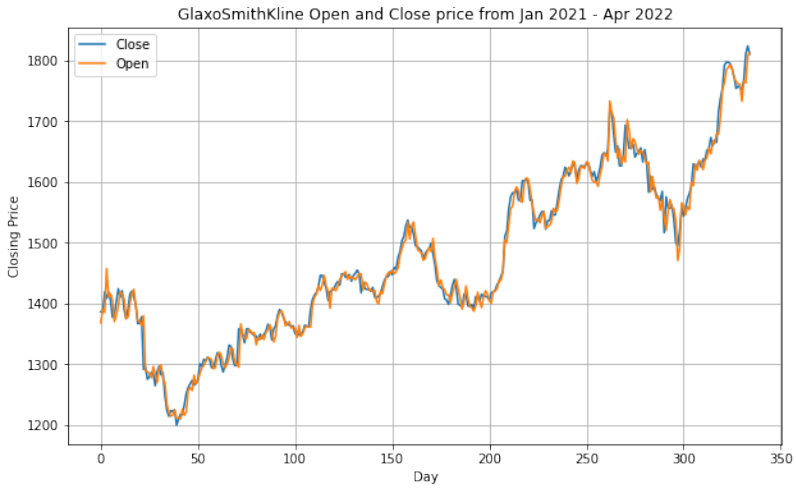


Figure 24: GSK Equities Opening and Closing prices from Jan 2021 to Apr 2022.

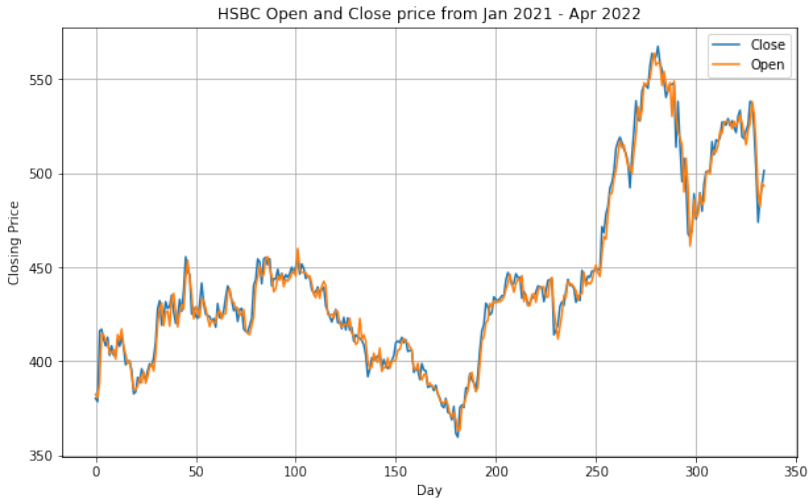


Figure 25: HSBC Opening and Closing prices from Jan 2021 to Apr 2022.

Figure 26: Vodafone Opening and Closing prices from Jan 2021 to Apr 2022.

The charts in figures 22-26 show the opening and closing prices for the selected equities, enabling us to examine the price trends. The charts, at first glance, do not show any noticeable trend or seasonality, implying non-stationarity in the data. However, we can confirm this by testing for stationarity. We can easily see the gradual increase in the stock

prices throughout the selected timeframe, with opening and closing prices for selected equities bar Vodafone higher on the last trading day than the first.

Stationarity Test

Although not required for the modelling purpose as we only intend to classify the direction which the stock price will move via the future trend of the closing price, we perform the augmented dickey fuller (ADF) test to check for stationarity in our dataset. In the ADF test, the null hypothesis is the presence of a unit root in a time-series data while the alternative hypothesis implies stationarity or trend-stationarity.

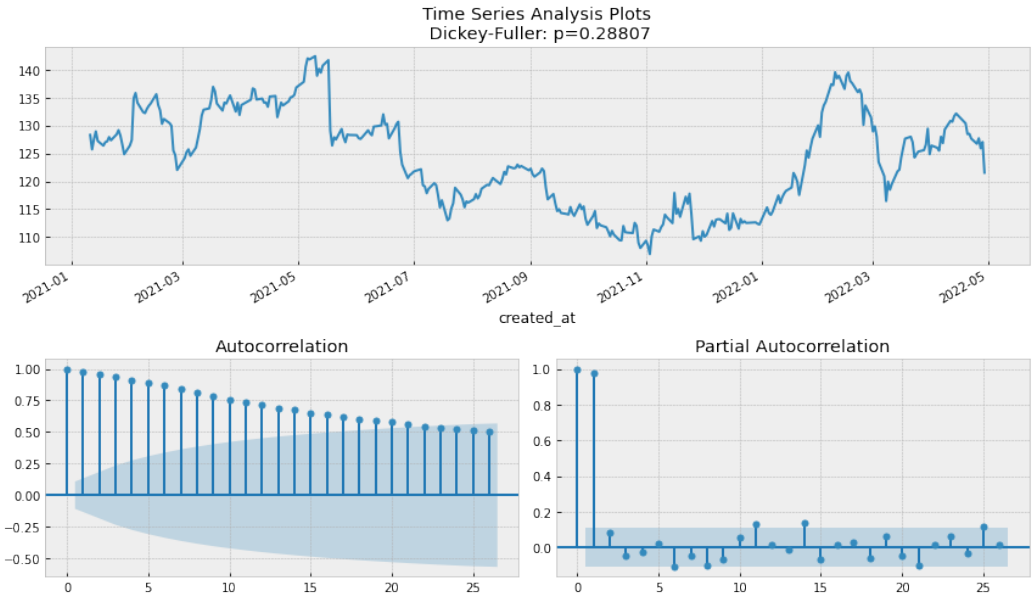


Figure 27: ADF test for Vodafone equities dataset.

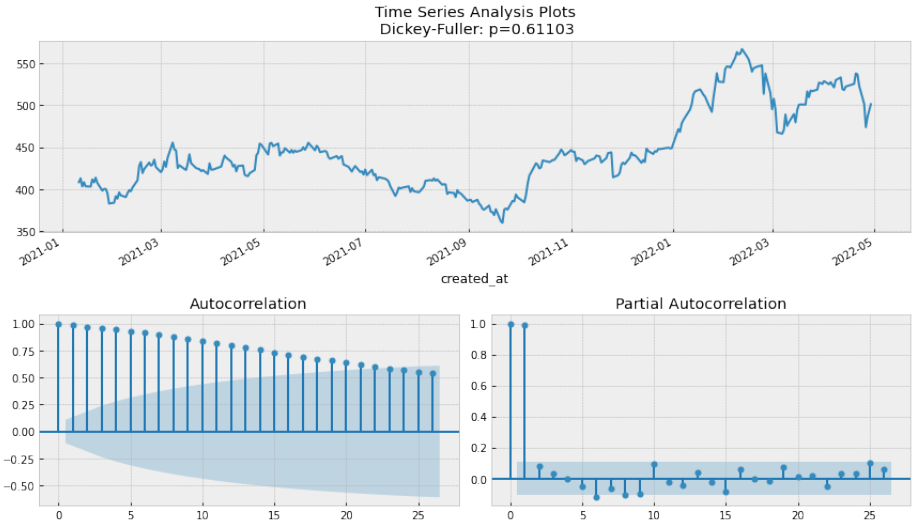


Figure 28: ADF test for HSBC equities dataset.

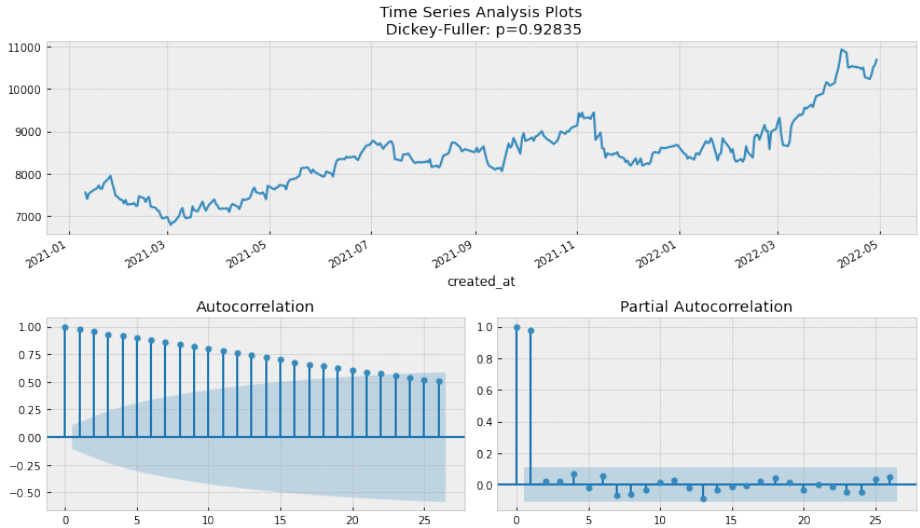


Figure 29: ADF test for AstraZeneca equities.

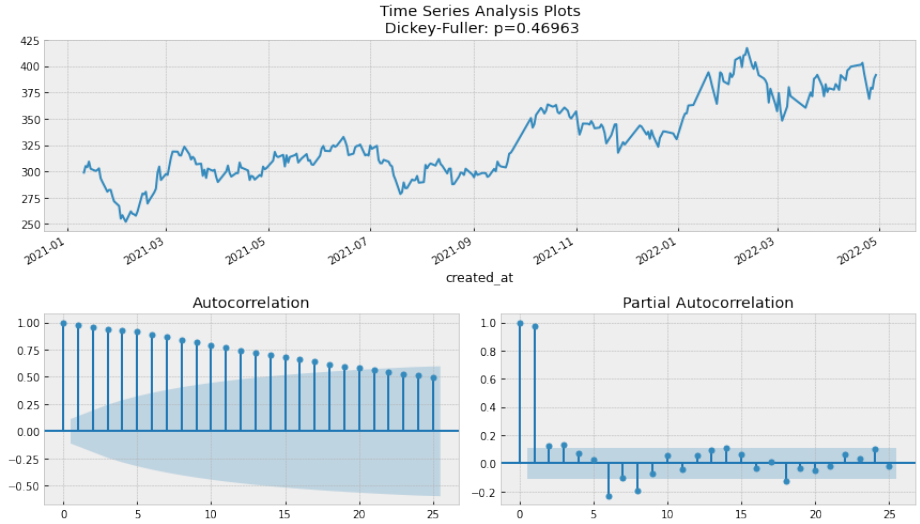


Figure 30: ADF test for BP equities.

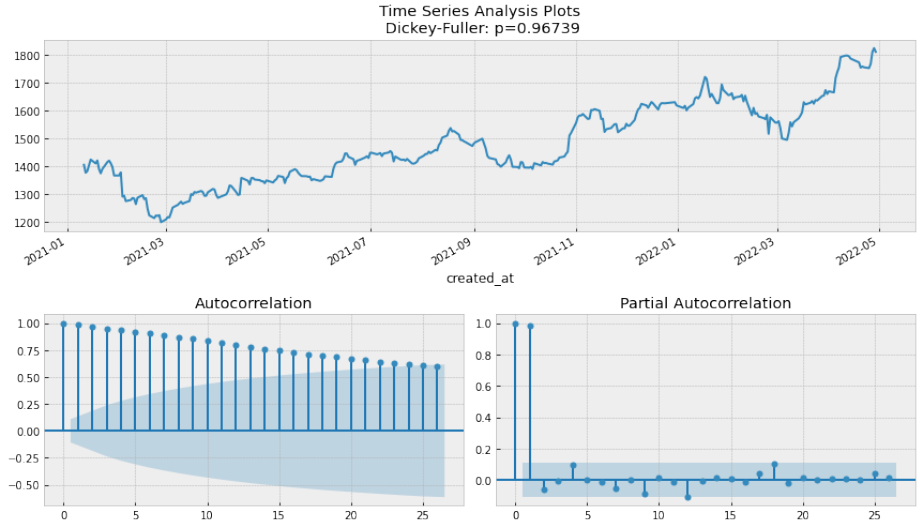


Figure 31: ADF test for GSK equities.

Stock daily volatility

We examined the behavior of the selected stocks in terms of their daily volatility. We measure the daily market volatility by utilizing the standard deviation which is the most common technique used by analysts and traders. (Boyte-White, 2022).

Stock trend prediction using SVM.

Since we have modelled the problem as a classification problem to predict the future trend of the stock, we will utilise one of the most robust and highly used classification algorithms which is also highly suitable for time series analysis, the support vector machine (SVM) (Okasha, 2014 cited Cao et al., 2005). SVM was developed by Vapnik and Cortes in 1995 and a model adopting the architecture can be thus represented:

$$\min_{\omega,b,\xi} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i$$

$$\text{s.t.} \quad y_i\left(\left(w.f(x_i)\right) + b\right) \geq 1 - \xi_i, i = 1$$

$$\xi_i \geq 0, i = 1, \dots l$$

where ξ_i is a tolerable training error, and C is a positive constant parameter that evaluates the trade-off between training errors and margin maximization. The optimization problem in the model is solved by transforming it to a dual problem whose solution set is the same. Finally, a decision function is constructed which can then be used for classification.

$$D(x) = sgn(\sum_{i=1}^{l} y_i\,\alpha_i * K\left(x_i, x_j\right) + b^*)$$

(Ren et al., 2019)

Evaluation Metrics

The metrics used in evaluating the model are introduced here. Accuracy has been selected as the primary metric for performance evaluation with precision, recall and f1 score as within class metrics.

Accuracy

The accuracy of a classifier is expressed simply as:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

If class distribution in a dataset is non-uniform, accuracy may not be a good metric for evaluating performance of the classifier. Therefore, a confusion matrix is introduced, and the model precision, recall, and F-measure are found.

Precision

Precision refers to how accurately a model classifies test samples and is calculated as follows.

$$Precision = \frac{TP}{TP + FP}$$

Where TP represents the true positive rate and FP is the false positive rate of the model.

Recall

Recall also referred to as sensitivity, measures how sensitive the classifier is to the maximum possible samples and is represented by the following equation.

$$Recall = \frac{TP}{TP + FN}$$

Where FN is the false negative rate of the model

F-measure

The F-measure is an aggregated measure calculated by taking the Harmonic mean of the precision and Recall of a classifier. The F-measure is expressed thus.

$$Recall = 2 \times \frac{Precision\ X\ Recall}{Precision + Recall}$$

Baseline

Since we are dealing with different companies compared to existing literature, we will select our baseline. The baseline (naïve) model will be a simplistic model that outputs the majority class as the predicted future trend. Baseline metrics for the selected companies are presented below:

Stock Ticker	Precision	Recall	F1-score
AstraZeneca	0.286	0.5	0.384
BP	0.267	0.5	0.348
GlaxoSmithKline	0.188	0.333	0.241
HSBC	0.266	0.5	0.347
Vodafone	0.257	0.5	0.339

Table 3: Baseline metrics for selected companies.

Model Validation

A range of techniques exist to validate predictive models. For the current model, k-fold cross validation was used to assess the predictive power of the model and its goodness of fit. The number of folds selected is 10 which is recommended by existing research as it has proven to be optimal in terms of variance and computation time. (Chou and Lin, 2013 cited Kohavi, 1995). In this method, ten mutually exclusive folds (or subsets) of the dataset are created,

each with roughly the same class distributions as the original dataset. Steps taken to extract the subsets are as follows.

1. Randomise the entire training set
2. Extract and remove 10% of original training set from randomised dataset (first fold)
3. Repeat steps 1-2 above eight times.
4. Assign the last 10% of the dataset as the last (10th) fold.

Nine subsets are used as training set for developing the model with the last fold acting as holdout each time. This procedure is looped ten times to obtain ten different performance estimates which is then averaged to calculate the overall CV accuracy. This method is imperative for overfitting control and selection of model best parameters. Cross Validation was carried out using the GridSearchCV class provided in the scikit-learn library.

RESULTS AND DISCUSSION

Sentiment Analysis

The final sentiment labels incorporated with the stock dataset were those generated by NLP transformers. 14,030 tweets were collected for the timeframe under consideration (January 2021 to April 2022). Sentiment labels for all tweets are presented below, 11,151 tweets were marked with the Neutral label with 1424 marked as Negative and 1455 marked positive.

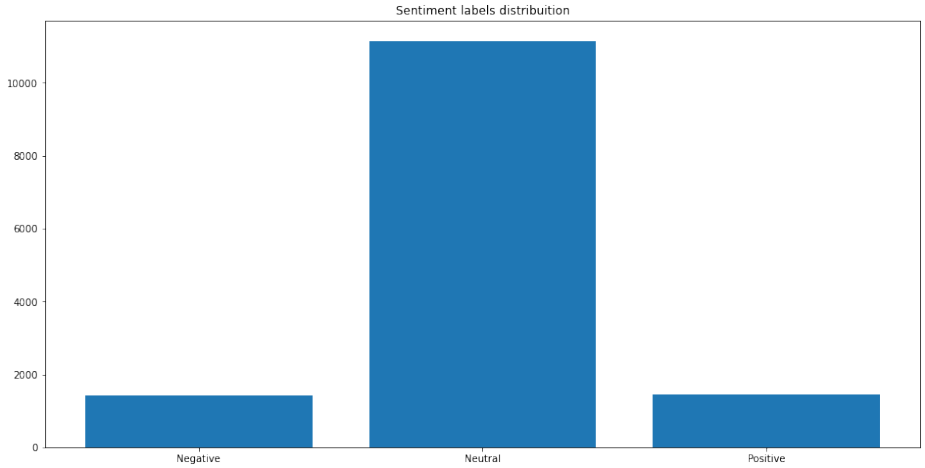


Figure 32: Sentiment labels generated by NLP transformers.

The tweets and sentiment labels were divided by companies and then aggregated on a daily level with irrelevant tweets that do not mention the companies discarded, results of the sentiment labels and the daily average as explained in section 3 is shown below:

Astra Zeneca

Total tweets mentioning Astra Zeneca - 1849

Label	Total number of tweets	Avg Daily count
Positive	161	169
Neutral	1581	197
Negative	107	114

Table 4: AZN tweet sentiment labels distribution.

BP

Total tweets mentioning BP - 4475

Label	Total number of tweets	Avg Daily count
Positive	504	169
Neutral	3570	197
Negative	401	114

Table 5: BP sentiment labels distribution

HSBC

Total tweets mentioning HSBC - 3186

Label	Total number of tweets	Avg Daily count
Positive	271	97
Neutral	2393	260
Negative	522	91

Table 6: HSBC sentiment labels distribution.

GSK

Total tweets mentioning GSK – 1867

Label	Total number of tweets	Avg Daily count
Positive	226	135
Neutral	1515	237
Negative	126	68

Table 7: GSK Sentiment labels distribution.

Vodafone

Total tweets mentioning Vodafone - 2584

Label	Total number of tweets	Avg Daily count
Positive	302	117
Neutral	2040	239
Negative	242	78

Table 8: Vodafone sentiment labels distribution.

From the tables above, it is apparent that the neutral sentiment is dominant in the original dataset and the aggregated daily averages.

Results of Stationarity test

Charts presenting the results of the ADF test are presented in figures 27-31. For all selected equities, the p-value for the test exceeds the significance level of 5% which means we cannot reject the null hypotheses of the presence of a unit root hence implying non-stationarity in the data and its statistical properties (Mean, Variance, Autocorrelation) do in-fact change over time. However due to the nature of the proposed model architecture we do not carry out any differencing and simply work with the data as it is.

Stock daily volatility

Stock	Average Daily volatility (%)
AZN	1.52
BP	2.19
GSK	1.16
HSBC	1.76
VODAFONE	1.52

Table 9: Daily volatility of selected equities.

Final Stock Prediction

The full table on results is shown below. Results are grouped by stock tickers for the selected companies.

AZN	SVM		RFC	
	With Sentiment Variable	Without sentiment variable	With Sentiment Variable	Without sentiment variable
Accuracy	0.790909091	0.818181818	0.809090909	0.672727273
Precision	0.517955646	0.535247432	0.535353535	0.446712018
Recall	0.535409035	0.545909646	0.522222222	0.382661783
F-score	0.523751239	0.540168188	0.526455026	0.366479925
BP	SVM		RFC	
	With Sentiment Variable	Without sentiment variable	With Sentiment Variable	Without sentiment variable
Accuracy	0.745098039	0.696078431	0.549019608	0.68627451
Precision	0.495956873	0.51125807	0.406819518	0.464472309
Recall	0.506709608	0.485507246	0.389694042	0.470746108
F-score	0.501065748	0.463366069	0.34855643	0.462061748
GSK	SVM		RFC	
	With Sentiment Variable	Without sentiment variable	With Sentiment Variable	Without sentiment variable
Accuracy	0.736363636	0.824826083	0.654545455	0.781818182
Precision	0.772941176	0.823030907	0.81372549	0.775815217
Recall	0.69701087	0.830163043	0.586956522	0.775815217

F-score	0.698459212	0.824826083	0.533690317	0.775815217
HSBC	SVM		RFC	
	With Sentiment Variable	Without sentiment variable	With Sentiment Variable	Without sentiment variable
Accuracy	0.672727273	0.745454545	0.627272727	0.763636364
Precision	0.720572057	0.737454304	0.568627451	0.75276012
Recall	0.5708731	0.752562743	0.519794981	0.767055497
F-score	0.534117647	0.738451087	0.461749612	0.755555556
VOD	Support Vector Machine		Random Forest Classifier	
	With Sentiment Variable	Without sentiment variable	With sentiment variable	Without Sentiment Variable
Accuracy	0.709090909	0.836363636	0.736363636	0.736363636
Precision	0.504273504	0.556587091	0.493578414	0.493962557
Recall	0.463050847	0.562485876	0.489152542	0.498305085
F-score	0.454054717	0.559410089	0.487522894	0.493133167

Table 10: Results showing metric scores for each stock ticker.

For this research analysis, we selected five companies (amongst the top twenty) listed on the London stock exchange and collected stock data from January 2021 to April 2022, culminating in 335 trading days. Average daily sentiments were incorporated into the dataset such that the sentiment from the previous day is reflected on the current trading day. Simple and Exponential moving averages were also calculated for a five-day period and incorporated as technical indicators into the final dataset which was then used in predicting the stock future trend as defined in section 3. After incorporating the moving averages and sentiment variable, we end up with data indicative of 330 trading days (spanning from Monday 11th January 2021 to Friday 29th April 2022) as the first five days are incapable of incorporating moving averages and were dropped as a result. In total, 10 models were built with the first five incorporating the sentiment variable and the remaining five incorporating stock data only without the sentiment variable. The primary metric used in evaluating the models is accuracy, with other metrics like Precision, Recall and F1 score used to further evaluate the models' quality. The data was partitioned into three equal parts with two-thirds used as the training set and the remaining one-thirds as the testing set. The training set is represents the trading days from Monday 11th January 2021 to Friday 19th November 2021 (220 trading days) with the testing set spanning the period from

Monday 22nd November 2021 to Friday 29th April 2022 (110 trading days).

We observe the accuracy scores on the testing set for each model in the table 8 above and for the choice algorithm (SVM with RBF kernel), accuracy scores ranged from 67.2% to 79.1% when incorporating the sentiment variable and from 69.6% to 83.6% when incorporating the stock data only. These accuracy scores hence show that for the selected companies, sentiment from social media has no predictive power on the equities future trend as four of the five models showed improved accuracy scores when incorporating stock data only. Vodafone equities showed the biggest jump in accuracy from 70.9% when incorporating the sentiment variable to 83.6% when incorporating stock variables only. However we note that BP equities showed improved accuracy when incorporating the sentiment variable which indicates some potential in the usefulness of sentiment from social media. Although, in showing this improved accuracy, the overall model quality is poor as shown by poor within class metrics of 51.1% precision, 48.6% recall and 46.3% f1 score. Among the companies studied, GlaxoSmithKline equities showed the most promising result with the model yielding 82.4% when incorporating stock data only and 73.6% accuracy when the sentiment variable is accounted for. Within class metrics are also very high indicated by 82.3% precision, 83% recall and 82.5% f1 score.

Classification report

The classification reports for the ten models are presented in the appendix section of this report (figures 32 to 51). We are able to further assess the models by looking at the quality of predictions generated for the future trend of the selected companies. Among selected companies, only GlaxoSmithKline had instances of a neutral future trend with other tickers experiencing either a positive or negative future trend. None of the chosen algorithms was able to correctly classify the neutral instances as we can see in figures 36-37 and 46-47 of the Appendix section. However, in the figures 46 and 50, we note that the promise shown by the chosen Support Vector Classifier algorithm in predicting majority of positive and negative samples correctly with 81% and 85% respective success for GSK equities and 85% and 84% success rates for Vodafone equities. Majority of the models are able to classify the positive samples correctly at a high rate, however, when incorporating the sentiment variable, majority of the models also largely misclassify the negative samples as shown by the high rate of false positives.

We can also use the classification report to compare the two algorithms further. In figures 32-41, we can note that the SVC has a higher success rate of classifying the negative future trend

correctly compared to Random Forest when we include the sentiment variable for all equities bar Vodafone. When removing the sentiment variable, the same pattern is noticeable with exception of BP and HSBC equities where Random Forest showed better success in classifying negative samples in the former (although with a lower true positive rate) and both algorithms had the same rate for the latter.

CONCLUSION AND LIMITATIONS

The concept of stock prediction is a classic phenomenon and remains a challenging task. Recently, the application of machine learning techniques in stock prediction has gained popularity. In this research, we sought to leverage this and determine if the social mood/sentiment about a company and its equities contain any predictive power towards its stock price trends. We proposed a hybrid approach that combined sentiment analysis and supervised learning to classify future stock trends. We utilise a state-of-the-art roBERTa base model based on NLP transformers architecture to perform sentiment classification, and using a predictive classification model, we analyse the predictive power of the sentiment variable when combined with stock data to forecast the future trend of the stocks. This study shows that the sentiment variable has no predictive power for the future trends of the stock when controlling for some technical indicators. The exponential weighted moving average was identified to have the most significant impact on the future trend of equities. The trend and sentiment variables were the least important in forecasting the equities' future trends.

In a nutshell, the goal of this research has been achieved as we have been about to conduct sentiment analysis on social media data and use it to test our hypotheses. Although evaluation metrics outperform the simplistic baseline we have utilised and are competitive compared to existing works, this study did not consider other factors affecting the stock price movement, such as News and Government policies. The Twitter dataset used for sentiment analysis consists of tweets from the general public who may or may not be investors in the selected companies, making it difficult to gauge the informativeness of sentiment obtained from this data. The application of Timelms for the sentiment classification task yielded excellent results due to the robustness and flexibility of the pre-trained model.

Selected stocks' behaviour was examined, and it was found that the stocks were not highly volatile during the selected timespan, as we would expect from companies with high market capitalisation. Despite this, only BP equities were influenced by sentiment from social media, with all remaining four equities showing worse performance when the sentiment from social media is included. Between the classifiers compared, the SVC gave consistent results in both cases and is highly recommended for stock trend classification.

However, the period considered in this study of 485 days from January 2021 to April 2022 meant that stock data was limited to 330 trading days and tweets data was limited to 14,030 tweets.

The total tweet volume generated showed that the selected companies are not among the most discussed companies on Twitter, making it difficult for Twitter sentiments to provide any relevant predictive power to determine the stock trend. Also, other social media platforms, blogs and message boards exist where investor sentiment can be more accurately captured for the selected companies. This study also did not take into consideration other factors such as post pandemic economic conditions, the start of the ongoing Russian invasion of Ukraine during the early months of the current year which would have contributed to market movement hence making it difficult to accurately model the stock trends with social media sentiment and stock data only.

RECOMMENDATION FOR FUTURE WORK

For future work, the selected timespan should be expanded to enable a high volume of data that can accurately highlight the complexities of stock price trends. Weighting can also be assigned to sentiment scores based on user reputation as some users are more influential than others and subsequently prevailing sentiment in their tweets. Also, analysing the predictive potential of sentiment variables at a higher frequency, such as intraday level, can be an intriguing expansion of this study. Empirical results from this study illustrate that combining sentiment features with stock market data can achieve a good performance. Combining this with more technical indicators, macroeconomic variables, and a stop-loss strategy can provide a springboard that can help investors mitigate against risks.

REFERENCES

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R.J., 2011, June. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* (pp. 30-38).
2. Agarwal, B. and Mittal, N., 2016. Prominent feature extraction for review analysis: an empirical study. *Journal of Experimental & Theoretical Artificial Intelligence*, *28*(3), pp.485-498.
3. Audrino, F., Sigrist, F. and Ballinari, D., 2020. The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, *36*(2), pp.334-357.
4. Barbosa, L. and Feng, J., 2010. 'Robust Sentiment Detection on Twitter from Biased and Noisy Data', 23rd International Conference on Computational Linguistics, Beijing, China, August 2010.
5. Bird, S., Klein, E. and Loper, E., 2009 *Natural Language Processing with Python.* 1st edn. USA: O'Reilly media Inc.
6. Boyte-White, C., 2022 *What Is the Best Measure of Stock Price Volatility?* Available at here. (Accessed 2nd September 2022)
7. Bustos, O. and Pomares-Quimbaya, A., 2020. Stock market movement forecast: A systematic review. *Expert Systems with Applications*, 156, p.113464.
8. Chen, J., 2022 *What is EMA? How to Use Exponential Moving Average with Formula.* Available at: https://www.investopedia.com/terms/s/sma.asp (Accessed: 3rd August 2022).
9. Chou, J.S. and Lin, C., 2013. Predicting disputes in public-private partnership projects: Classification and ensemble models. *Journal of Computing in Civil Engineering*, *27*(1), pp.51-60.
10. Critien, J.V., Gatt, A. and Ellul, J., 2022. Bitcoin price change and trend prediction through twitter sentiment and data volume. *Financial Innovation*, *8*(1), pp.1-20.
11. Derakhshan, A. and Beigy, H., 2019. Sentiment analysis on stock social media for stock price movement prediction. *Engineering Applications of Artificial Intelligence*, *85*, pp.569-578.
12. Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), pp.383-417.

13. Feldman, R., 2013. Techniques and applications for sentiment analysis. *Communications of the ACM,* 56(4), pp.82-89. Available at https://doi.org/10.1145/2436256.2436274. (Accessed 2022).
14. Fuller, W. A. (1976). 'Introduction to Statistical Time Series.' New York: John Wiley and Sons. ISBN 0-471-28715-6.
15. Gandhmal, D.P. and Kumar, K. (2019) 'Systematic analysis and review of stock market prediction techniques', *computer science review,* 34, p.100190, ISSN 1574-0137. Available at https://doi.org/10.1016/j.cosrev.2019.08.001 (accessed 18th July 2022).
16. Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*. MIT press.
17. Guresen, E., Kayakutlu, G. and Daim, T.U., 2011. Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, *38*(8), pp.10389-10397.
18. Hayes, A. (2021) *Simple Moving Average (SMA).* Available at: https://www.investopedia.com/terms/e/ema.asp (Accessed: 3rd August 2021).
19. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. *Neural computation*, *9*(8), pp.1735-1780.
20. Hu, Y., Liu, K., Zhang, X., Su, L., Ngai, E.W.T. and Liu, M., 2015. 'Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review'. *Applied Soft Computing*, *36*, pp.534-551. Available at https://doi.org/10.1016/j.asoc.2015.07.008 (accessed 15th July 2022).
21. Jiang, W., 2021. Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, *184*, p.115537. Available at https://doi.org/10.1016/j.eswa.2021.115537
22. Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K.H. and Alfakeeh, A.S., 2020. Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, pp.1-24.
23. Krishna, K. and Murty, M.N., 1999. Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), pp.433-439.
24. Likas, A., Vlassis, N. and Verbeek, J.J., 2003. The global k-means clustering algorithm. *Pattern recognition*, *36*(2), pp.451-461.
25. Liu, B., 2010. 'Sentiment analysis and subjectivity' in Indurkhya, N. and Damerau, F.J. (eds.) *Handbook of natural language processing*, *2*(2010), United States of America, Taylor & Francis, pp.627-666.

26. Liu, B., 2012. *Sentiment analysis and opinion mining* Morgan & Claypool.
27. Loureiro, D., Barbieri, F., Neves, L., Anke, L.E. and Camacho-Collados, J., 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.
28. Manning, D.C., Raghavan, P. and Schütze, H. (2009) *An Introduction to Information Retrieval.* Online edn. Cambridge, England: Cambridge University Press.
29. Mäntylä, M.V., Graziotin, D. and Kuutila, M. (2018) *The evolution of sentiment analysis—A review of research topics, venues, and top cited papers.* Computer Science Review, 27, pp.16-32.
30. May. (2020) *Is twitter structured data or unstructured data?* Available at: https://datadition.com/is-twitter-structured-data-or-unstructured-data/ (Accessed: 20th July 2022).
31. McKinney, W. (2018) *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. 3rd edn. USA: O'Reilly Media, Inc.
32. Medhat, W., Hassan, A. and Korashy, H., 2014. S*entiment analysis algorithms and applications: A survey*. Ain Shams engineering journal, 5(4), pp.1093-1113.
33. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.
34. Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. [Online]. Available at http://arxiv.org/abs/1301.3781.
35. Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T. and Trajanov, D., 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, *8*, pp.131662-131682.
36. Nabipour, M., Nayyeri, P., Jabani, H., Mosavi, A. and Salwana, E., 2020. Deep learning for stock market prediction. *Entropy*, *22*(8), p.840.
37. Nguyen, T.H., Shirai, K. and Velcin, J., 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), pp.9603-9611.
38. Nti, I.K., Adekoya, A.F. and Weyori, B.A., 2020. Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence from Ghana. *Appl. Comput. Syst.*, *25*(1), pp.33-42.
39. Okasha, M.K., 2014. Using support vector machines in financial time series forecasting. *International Journal of Statistics and Applications*, *4*(1), pp.28-39.

40. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, pp.2825-2830.
41. Phan, H.T., Tran, V.C., Nguyen, N.T. and Hwang, D., 2020. Improving the performance of sentiment analysis of tweets containing fuzzy sentiment using the feature ensemble model. *IEEE Access*, *8*, pp.14630-14641.
42. Ren, R., Wu, D.D. and Liu, T., 2018. Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, *13*(1), pp.760-770.
43. Ruan, Y., Durresi, A. and Alfantoukh, L., 2018. Using Twitter trust network for stock market analysis. *Knowledge-Based Systems*, *145*, pp.207-218.
44. Shah, D., Isah, H. and Zulkernine, F., 2018, December. Predicting the effects of news sentiments on the stock market. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 4705-4708). IEEE. Available at https://doi.org/10.1109/BigData.2018.8621884. (Accessed June 2022).
45. Shah, D., Isah, H. and Zulkernine, F., 2019. Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2), p.26.
46. Siddiqui, T. and Tiwary, U.S. (2008) *Natural Language Processing and Information Retrieval*. 1st edn. India: Oxford University Press.
47. Siqueira, H. and Barros, F., 2010, October. A feature extraction process for sentiment analysis of opinions on services. In *Proceedings of International Workshop on Web and Text Intelligence* (pp. 404-413).
48. Stenqvist, E. and Lönnö, J., 2017. Predicting Bitcoin price fluctuation with Twitter sentiment analysis.
49. Tsytsarau, M. and Palpanas, T. (2012) *Survey on mining subjective data on the web*. Data Min Knowl Disc 24, 478–514 (2012). Available at https://doi.org/10.1007/s10618-011-0238-6, (accessed 18th July 2022).
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, *30*.
51. Wong, E.L.X., 2021. Prediction of Bitcoin prices using Twitter Data and Natural Language Processing.

52. Xing, F.Z., Cambria, E. and Welsch, R.E., 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, *50*(1), pp.49-73.
53. Zhang, L., Wang, S. and Liu, B., 2018. *Deep learning for sentiment analysis: A survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), p.e1253.
54. https://companiesmarketcap.com/united-kingdom/largest-companies-in-the-uk-by-market-cap/

APPENDIX

Five steps of Porter's Stemming Algorithm

Step 1: Deal with plurals and past participles.

Step 1a:

Action Rules	*Examples*
SSES → SS	grasses → grass
IES → I	species → speci
SS → SS	fitness → fitness
S → null	boys → boy

Step 1b:

Conditions	*Action Rules*	*Examples*
($m > 0$)	EED → EE	proceed → procee
		meek → meek
(*υ*)	ED → null	scared → scared
(*υ*)	ING → null	sleeping → sleep
		ding → ding

Step 1b1:

Conditions	*Action Rules*	*Examples*
Null	AT → ATE	rotat(ed) → rotate
Null	BL → BLE	struggled → struggle
Null	IZ → IZE	recogniz(ed) → recognize
(*d and not)	double letter	ding → ding
(*L or *S or *Z)	→ single	call(ing) → call
		miss(ing) → miss
(m = 1 and *o)	Null → E	trail(ing) → trail
		pil(ing) → pile

Step 1c:

Conditions	*Action Rules*	*Examples*
(*υ*) Y → I	lazy → lazi	
	spy → spy	

Step 2:

Conditions	*Action Rules*	*Examples*
(m > 0) ATIONAL → ATE		rotational → rotate
(m > 0) TIONAL → TION		proportional → proportion
(m > 0) ENCI → ENCE		valenci → valence
(m > 0) ANCI→ ANCE		hesitanci → hesitance
(m > 0) IZER → IZE		recognizer → recognize
(m > 0) ABLI → ABLE		stabli → stable
(m > 0) ALLI → AL		practicalli → practical
(m > 0) ENTLI → ENT		differentli → different
(m > 0) ELI → E		vileli → vile
(m > 0) OUSLI → OUS		previousli → previous
(m > 0) IZATION → IZE		privatization → privatize
(m > 0) ATION → ATE		predication → predicate
(m > 0) ATOR → ATE		dictator → dictate
(m > 0) ALISM → AL		socialism → social
(m > 0) IVENESS → IVE		comprehensiveness → comprehensive
(m > 0) FULNESS → FUL		successfulness → successful
(m > 0) OUSNESS → OUS		obviousness → obvious
(m > 0) ALITI → AL		realiti → real
(m > 0) IVITI → IVE		transitiviti → transitive
(m > 0) BILITI → BLE		reliabiliti → reliable

Step 3:

Conditions	*Action Rules*	*Examples*
(m > 0) ICATE → IC		replicate → replic
(m > 0) ATIVE → null		formative → form
(m > 0) ALIZE → AL		conceptualize → conceptual
(m > 0) ICITI → IC		electriciti → electric
(m > 0) ICAL → IC		aeronautical → aeronautic
(m > 0) FUL →		hopeful → hope
(m > 0) NESS →		goodness → good

Step 4:

Conditions	*Action Rules*	*Examples*
(m > 1) AL →		revival → reviv
(m > 1) ANCE →		allowance → allow
(m > 1) ENCE →		preference → prefer
(m > 1) ER→		hardliner → hardlin
(m > 1) IC →		endoscopic → endoscop
(m > 1) ABLE →		adjustable → adjust
(m > 1) IBLE →		defensible → defens
(m > 1) ANT →		irritant → irrit
(m > 1) EMENT →		replacement → replac
(m > 1) MENT →		adjustment → adjust
(m > 1) ENT →		dependent → depend
(m > 1 and *S or *T)) ION →		adoption → adopt
(m > 1) OU →		homologou → homolog
(m > 1) ISM →		communism → commun
(m > 1) ATE →		activate → activ
(m > 0) ITI →		singulariti → singular
(m > 0) OUS →		anologous → homolog
(m > 0) IVE →		defective → defect
(m > 0) IZE →		equalize → equal

Step 5a:

Conditions	*Action Rules*	*Examples*
(m > 1) E →		equate → equat
		date → date
(m = 1 and not *o) E → null		cease → ceas

Step 5b:

Conditions	*Action Rules*	*Examples*
(m > 1 and *d and *L)		null → single letter
		controll → control
		roll → roll

WITH SENTIMENT VARIABLE

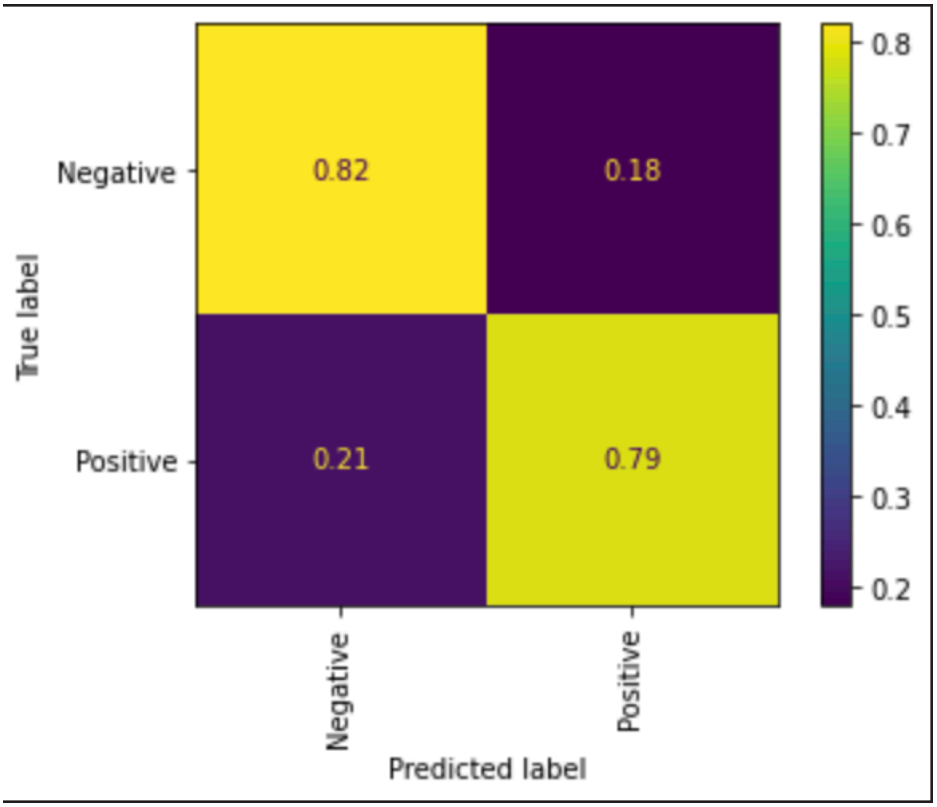


Figure 33: SVC Classification Report for AZN equities future trend.

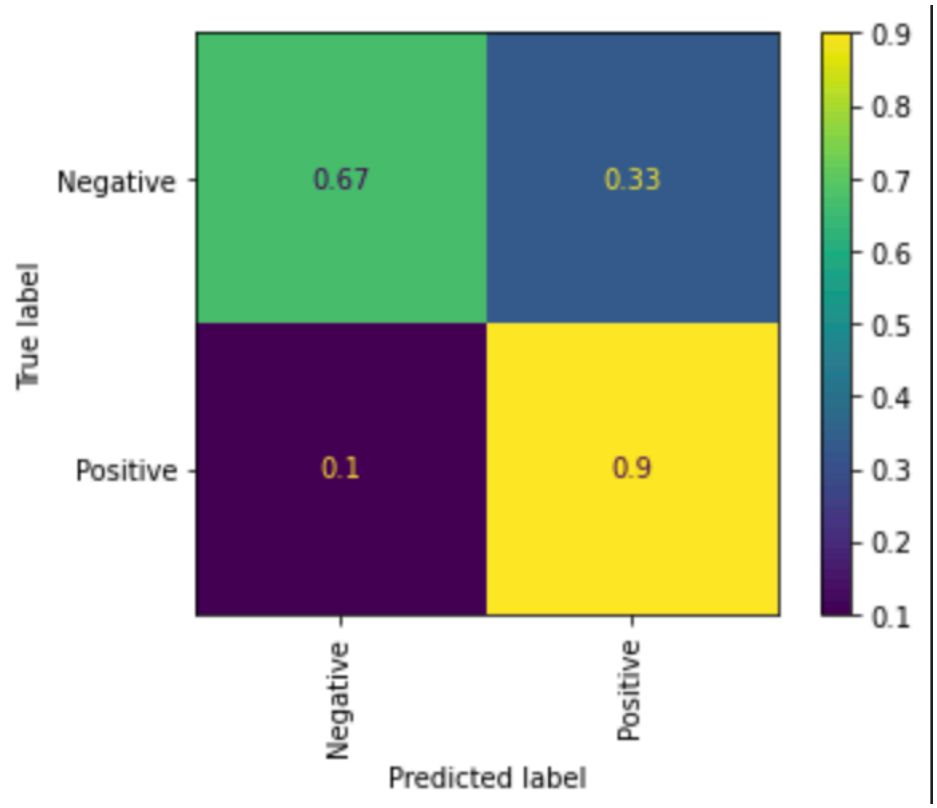


Figure 34: Random Forest Classification Report for AZN equities future trend.

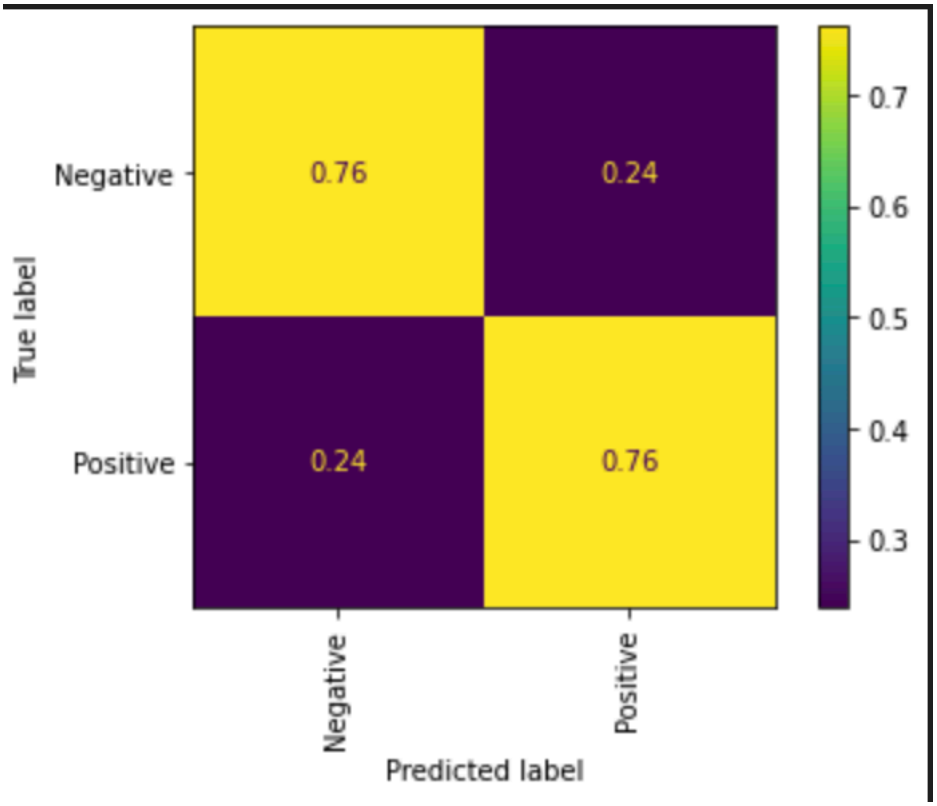


Figure 35: SVC Classification Report for BP equities future trend.

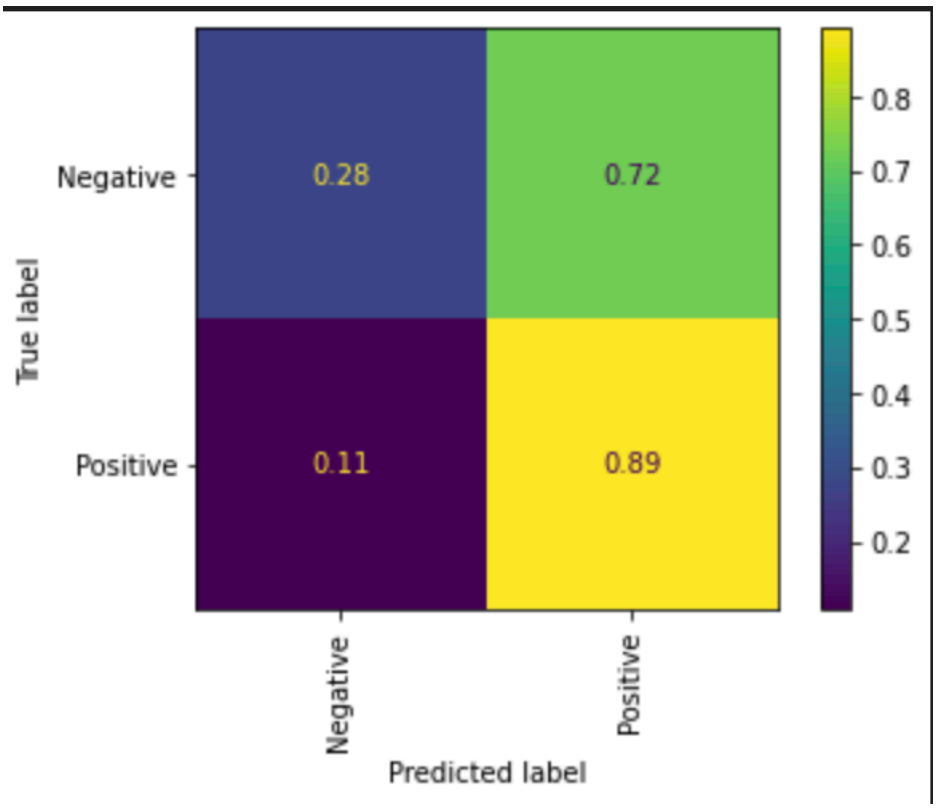


Figure 36: RF Classification Report for BP equities future trend.

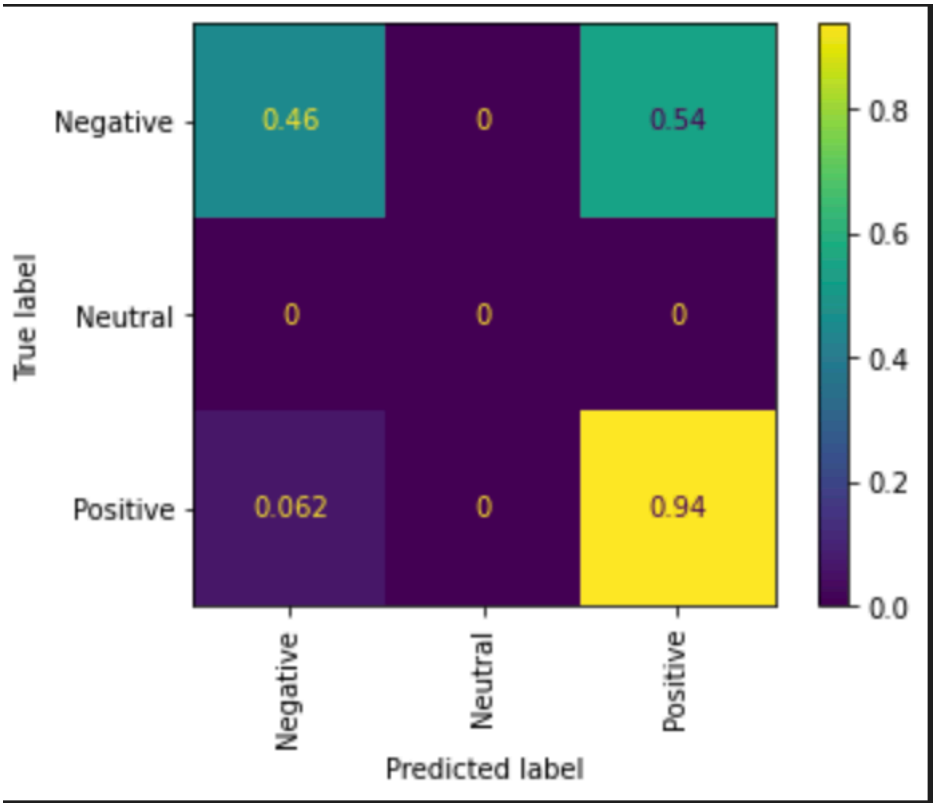


Figure 37: SVC Classification Report for GSK equities future trend.

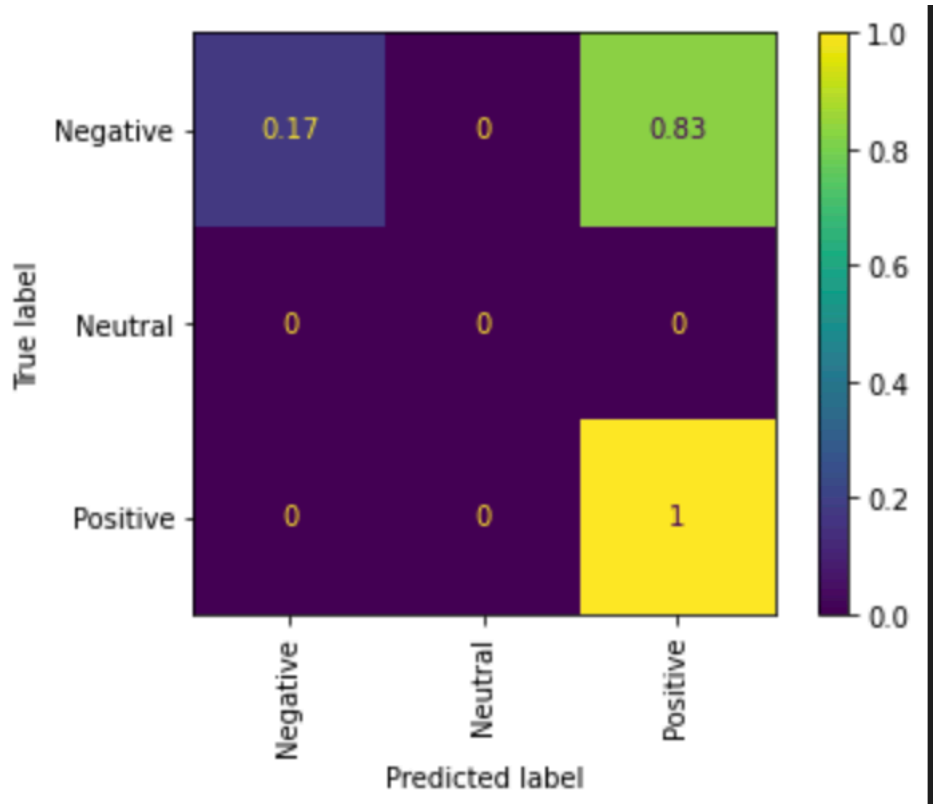


Figure 38: RF Classification Report for GSK equities future trend

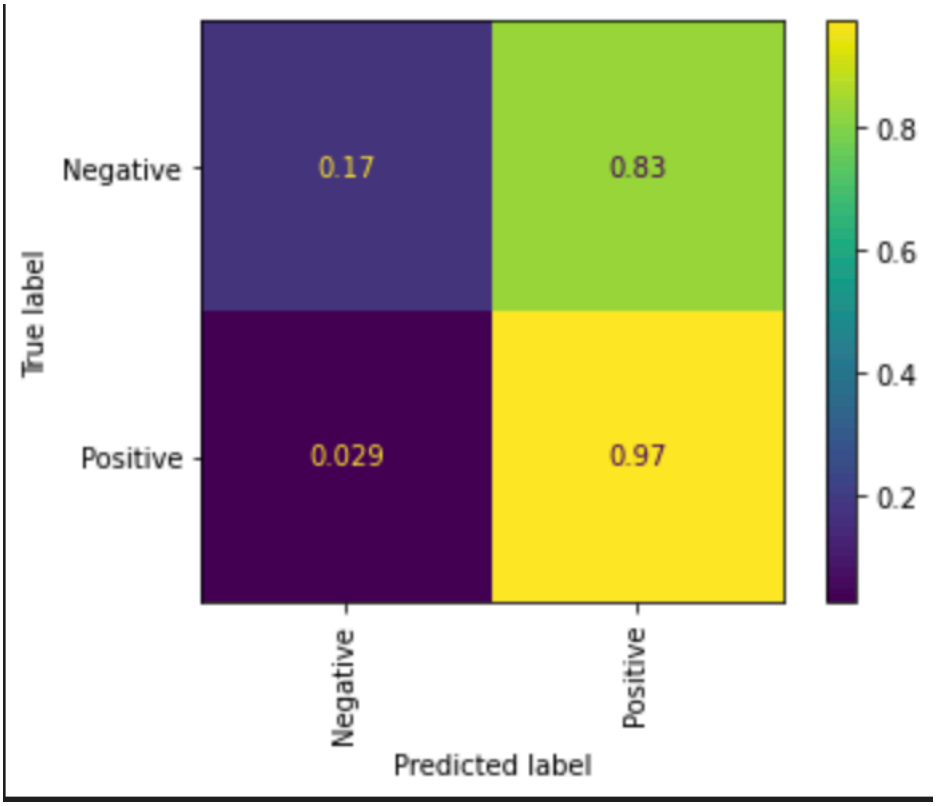


Figure 39: SVC Classification Report for HSBC equities future trend.

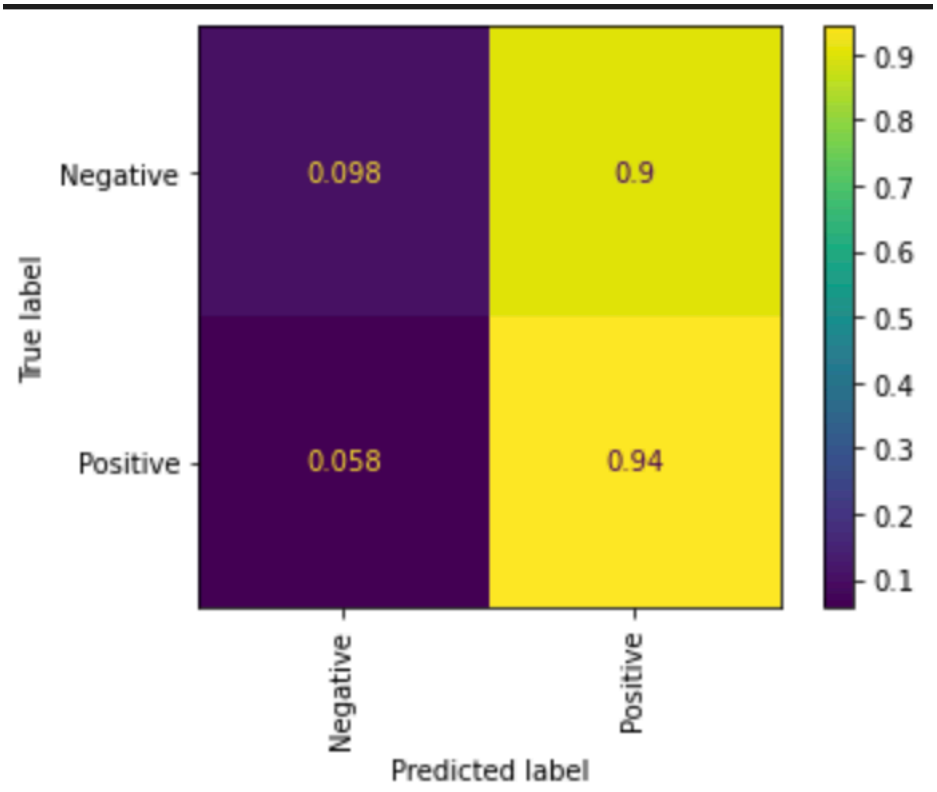


Figure 40: RF Classification Report for HSBC equities future trend.

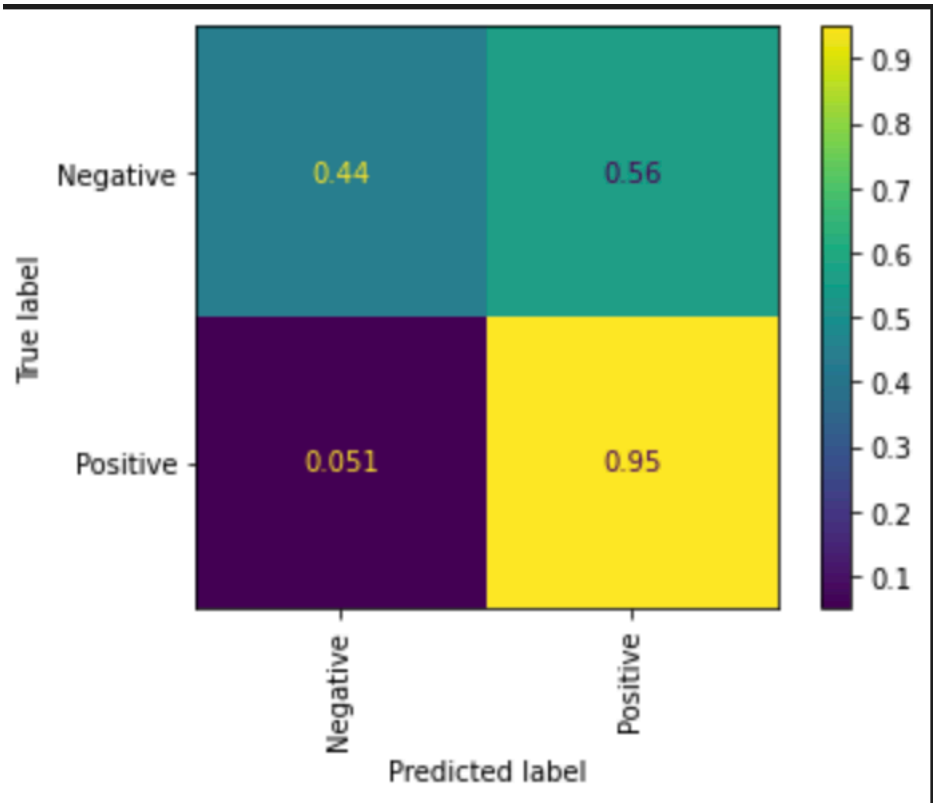


Figure 41: SVC Classification Report for Vodafone equities future trend.

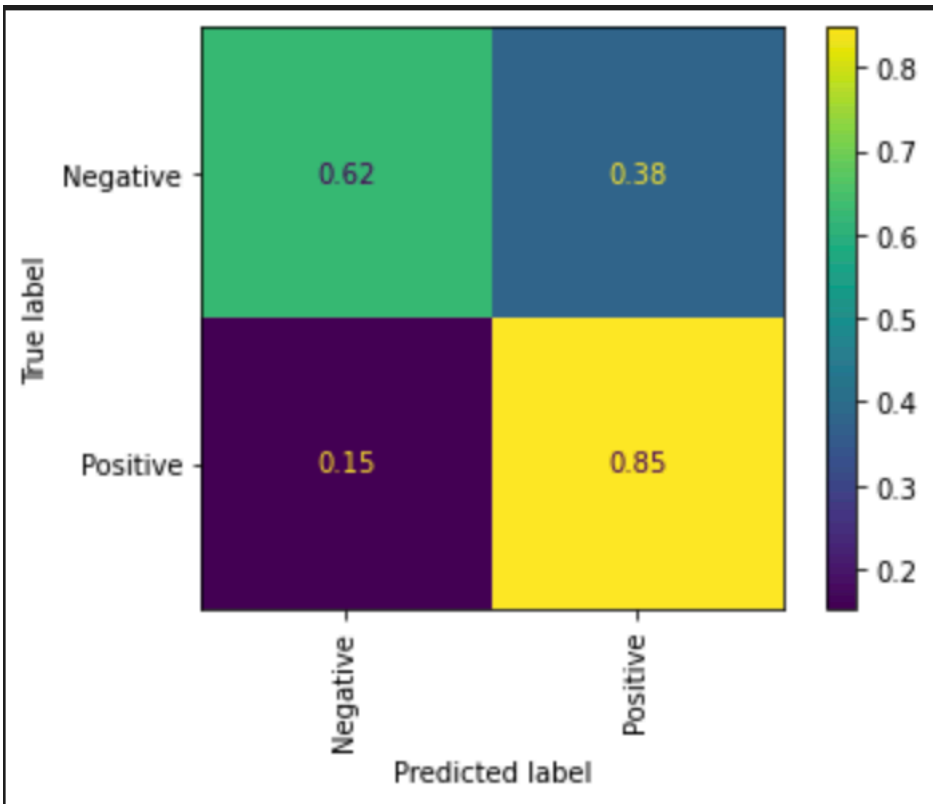


Figure 42: RF Classification Report for Vodafone equities future trend.

WITHOUT SENTIMENT VARIABLE

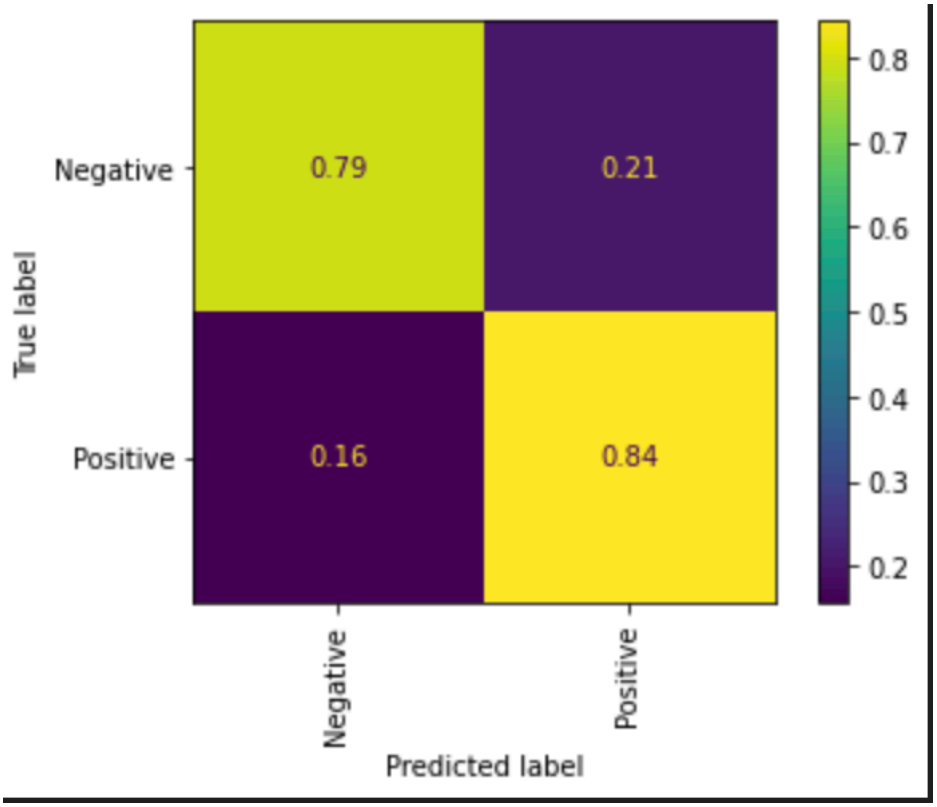


Figure 43: SVC Classification Report for AZN equities future trend.

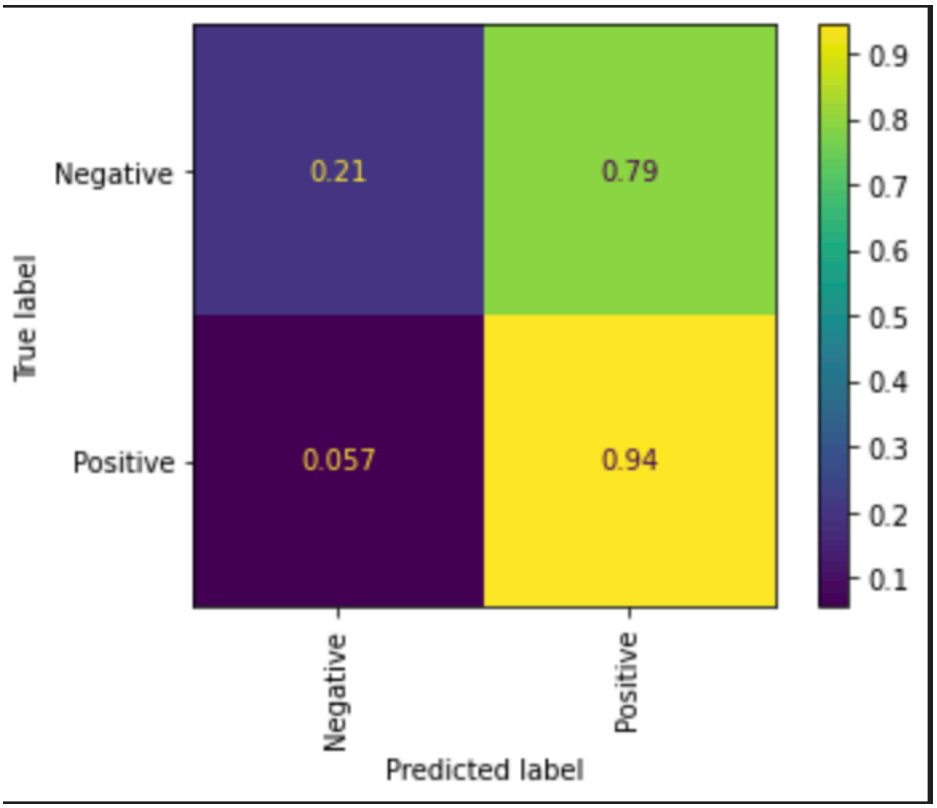


Figure 44: RF Classification Report for AZN equities future trend.

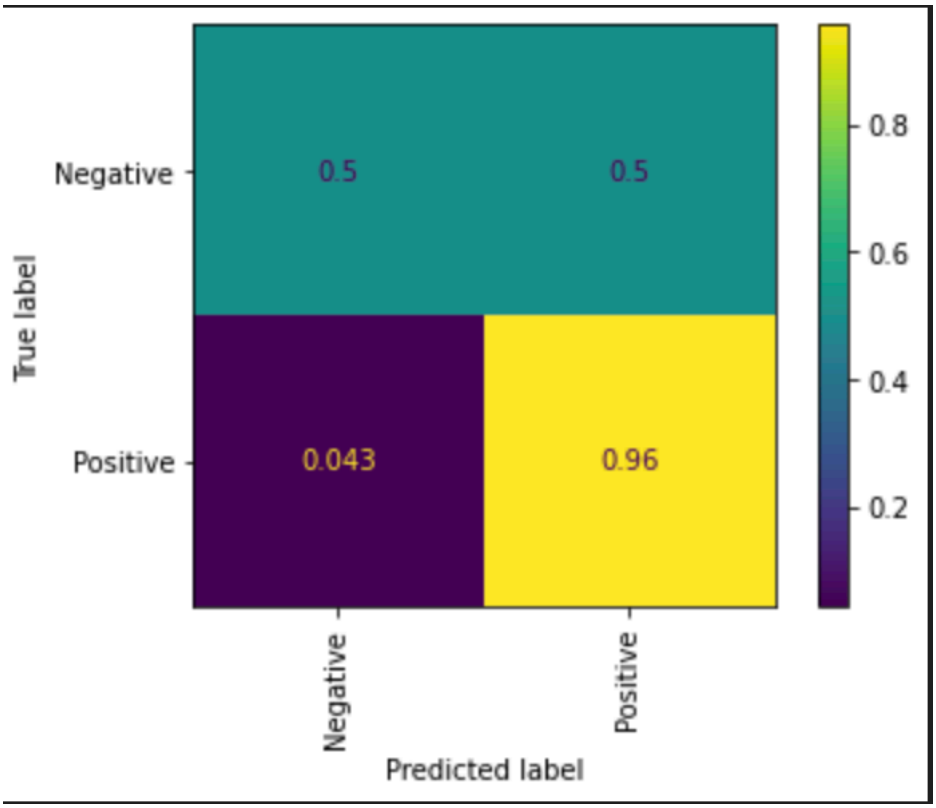


Figure 45: SVC Classification Report for BP equities future trend.

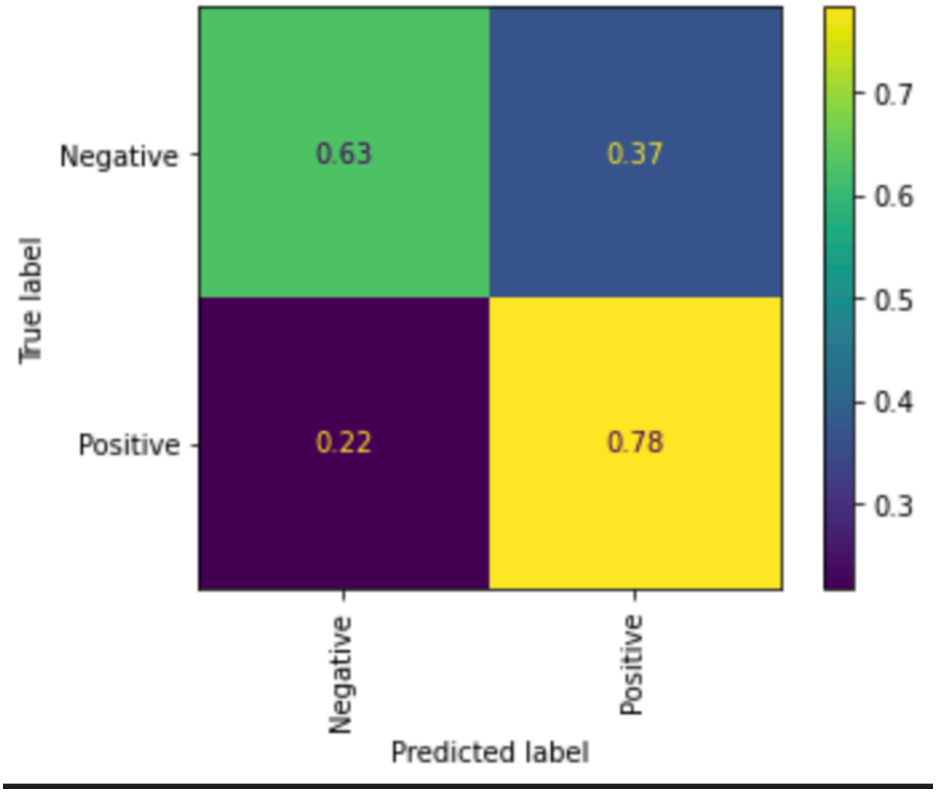


Figure 46: RF Classification Report for BP equities future trend.

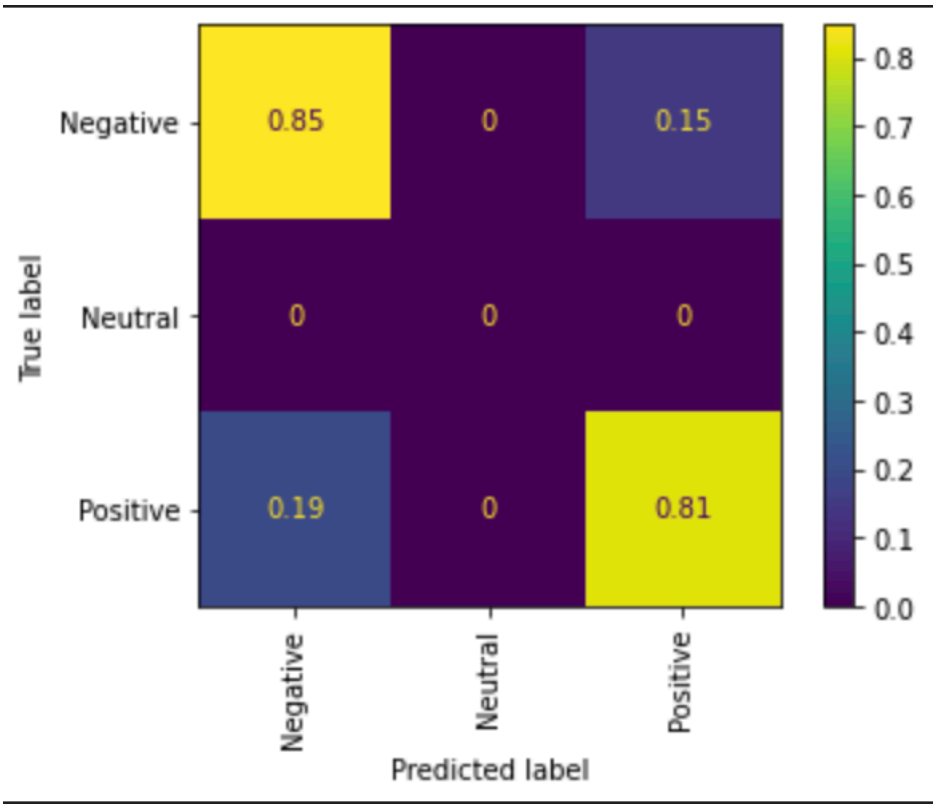


Figure 47: SVC Classification Report for GSK equities future trend.

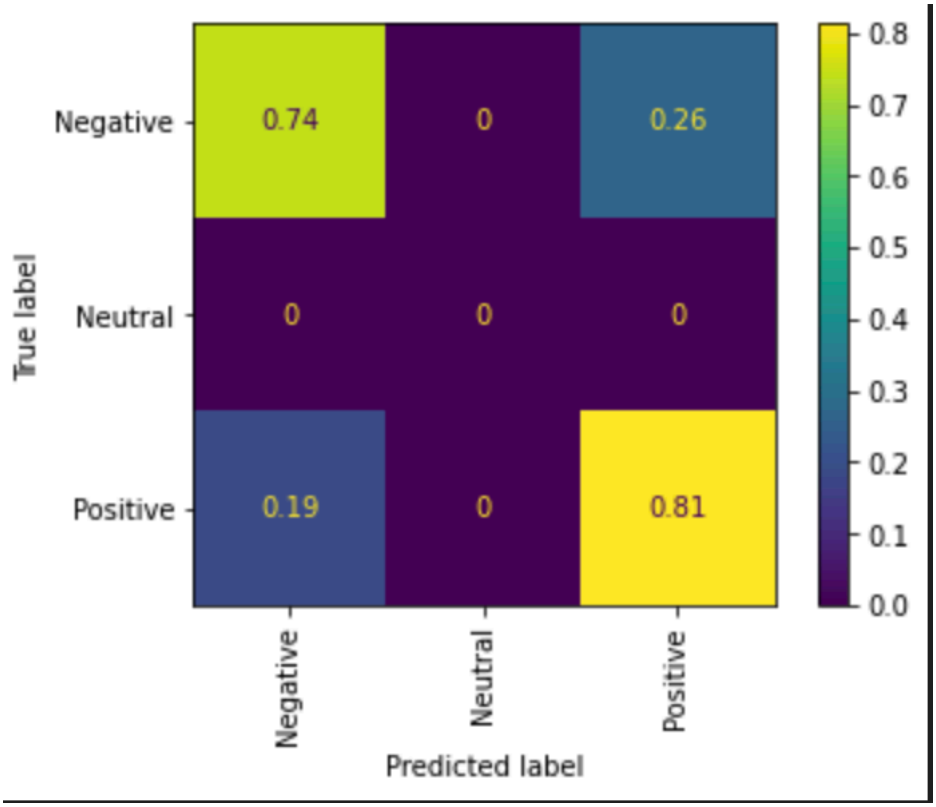


Figure 48: RF Classification Report for GSK equities future trend.

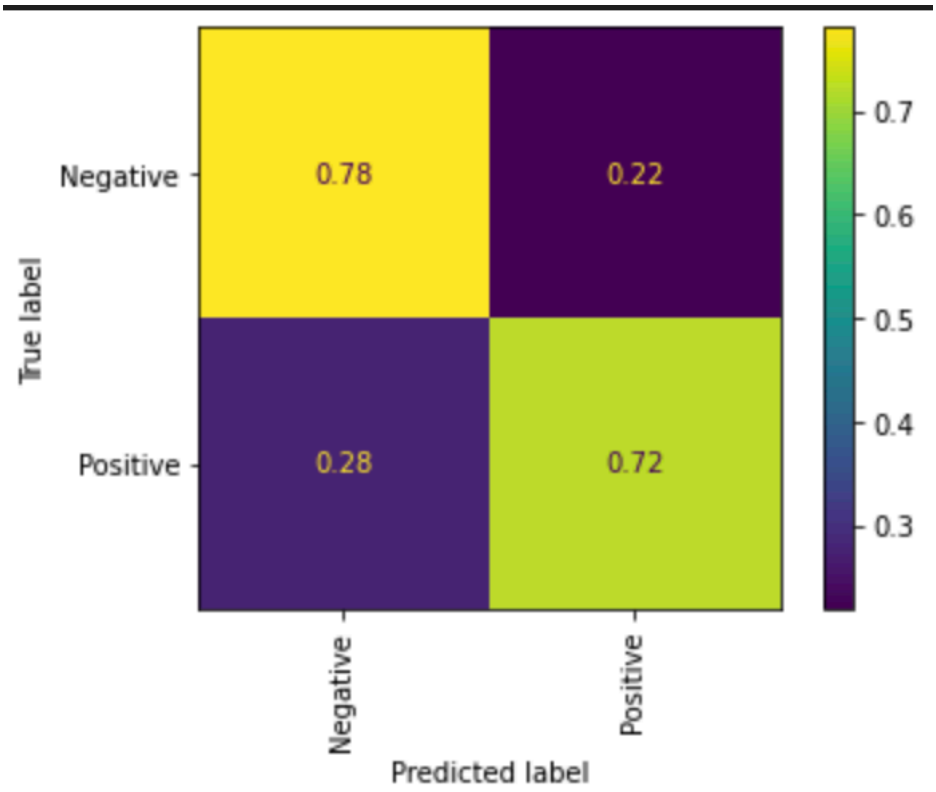


Figure 49: SVC Classification Report for HSBC equities future trend.

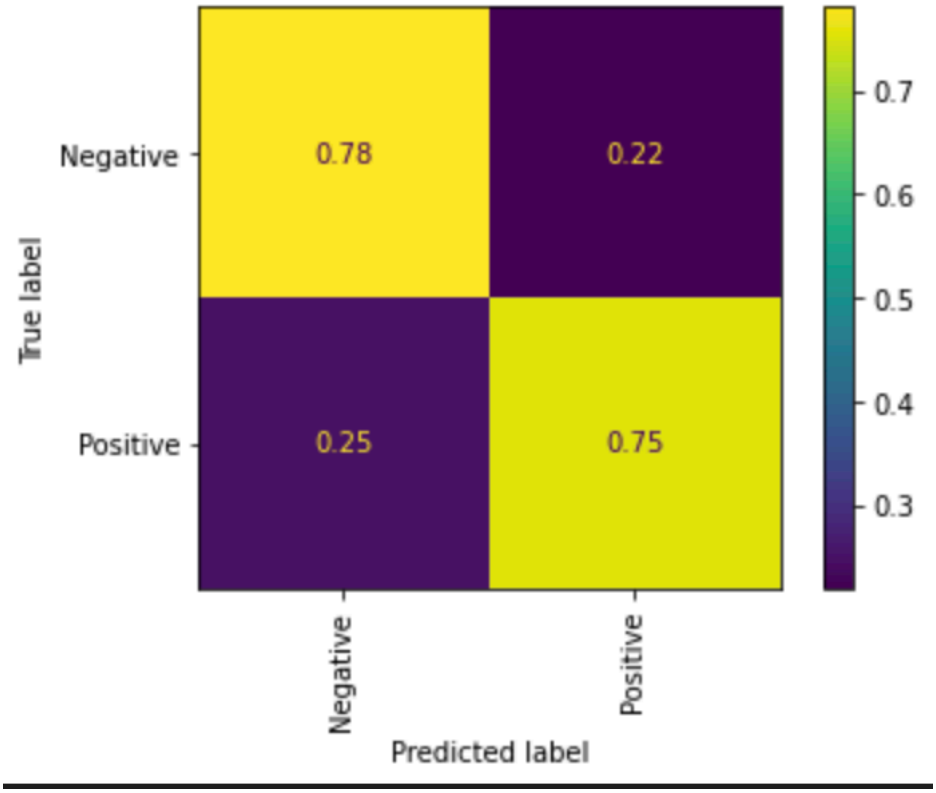


Figure 50: RF Classification Report for HSBC equities future trend.

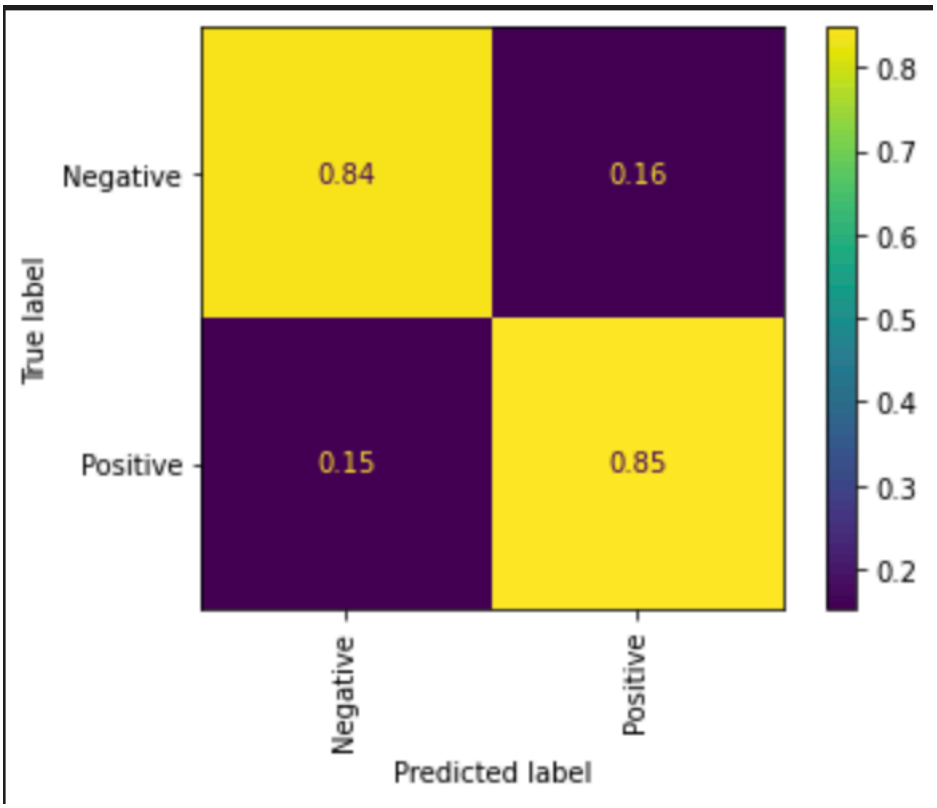


Figure 51: SVC Classification Report for Vodafone equities future trend.

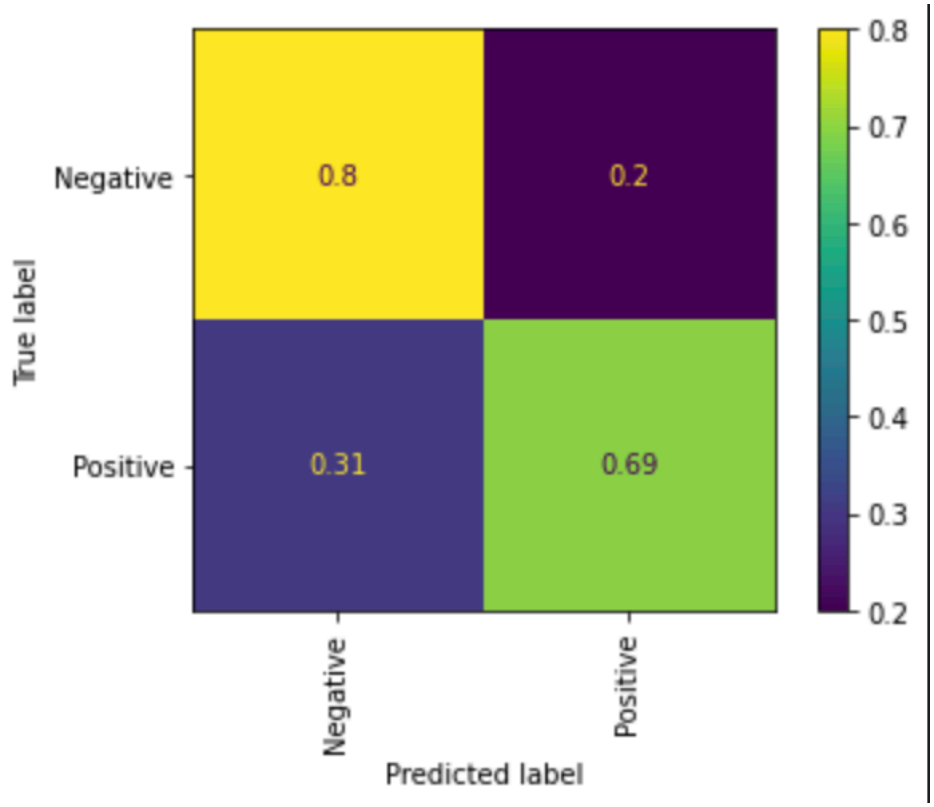


Figure 52: RF Classification Report for Vodafone equities future trend.

FEATURE IMPORTANCE

```
ema_5: 0.1799983638183909
Adj Close: 0.16353411395063916
sma_5: 0.13774236935174775
Close: 0.12895381720297922
Low: 0.11057244535645551
High: 0.09898864739317186
Open: 0.08158436311508938
Volume: 0.05341584171625331
Neutral: 0.02858645922910011
sentiment: 0.01662357886617282
Positive: 0.0
```

Figure 53: AZN feature importance.

```
ema_5: 0.2146660921847547
sma_5: 0.1490517856930352
Close: 0.12211463857321327
Low: 0.11312818548778543
Adj Close: 0.10152969510388996
High: 0.09731425038752783
Volume: 0.09633388364124684
Open: 0.08237476944466213
sentiment: 0.014628924693771967
Neutral: 0.008673798174654475
encoded_trend: 0.0001839766154580689
```

Figure 54: BP feature importance.

```
ema_5: 0.17801584515772415
sma_5: 0.14401565688859294
Adj Close: 0.1390711316435448
Close: 0.1297018209285612
High: 0.10171483339855858
Low: 0.09834476032710258
Volume: 0.08603451155191302
Open: 0.08585833401379699
Neutral: 0.01981765621718742
sentiment: 0.01575919991584711
Positive: 0.0016662499571712 95
```

Figure 55: GSK feature importance.

```
ema_5: 0.1631045021655808
sma_5: 0.1413389319332508
Close: 0.1322071642503783
High: 0.130645479043995 57
Adj Close: 0.106450208519716 95
Volume: 0.10153338382035833
Low: 0.0780784379408125
Open: 0.07117638284126443
Positive: 0.0645909612967719
sentiment: 0.010874548187870299
```

Figure 56: HSBC feature importance.

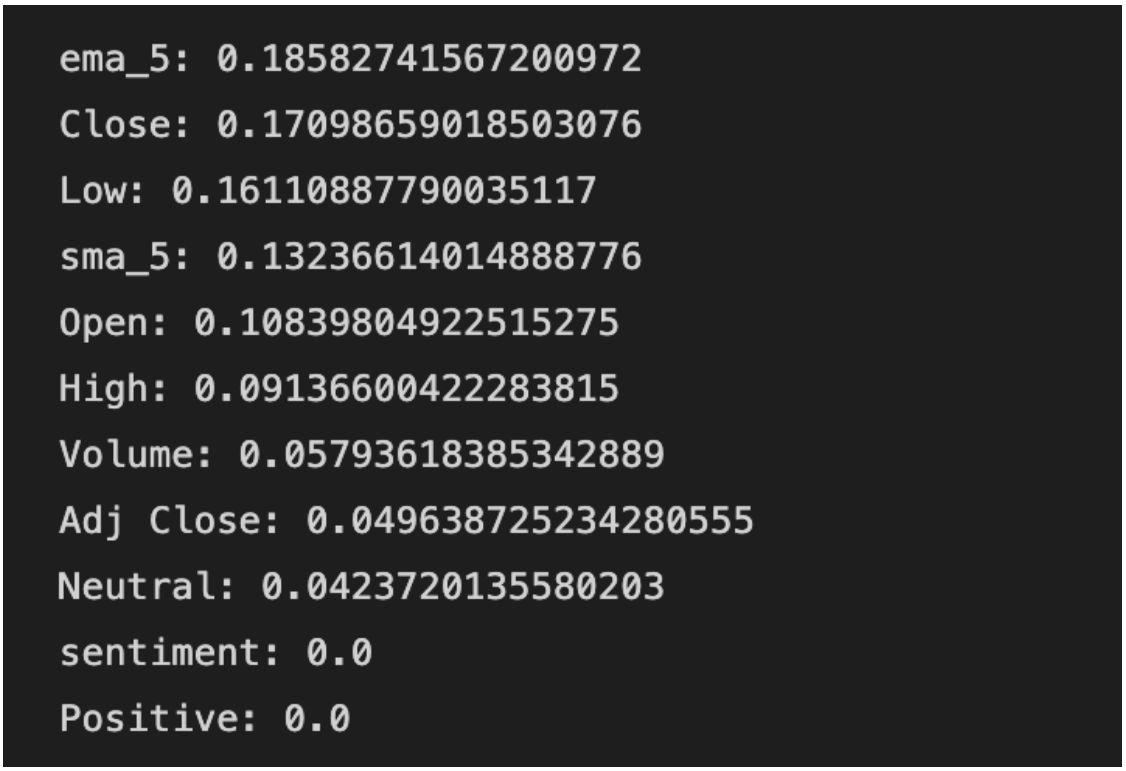


Figure 57: Vodafone feature importance.