Leveraging Deep Learning for Stock Market Prediction: A Sentiment Analysis Approach

Author: Daniel Forrester, 210280456.
Affiliation: Newcastle University, School of Computing G400.
Start Date: February 2024.

Abstract:

This research project aims to enhance sentiment analysis models by incorporating a diverse range of data from Twitter and later Reddit [32], with the objective of improving stock price prediction accuracy. The novelty of this approach stems from the utilisation of real-world, organic data from social media platforms instead of the conventional pre-human-classified sentiment scores of words, without consideration for phrases. Which can be continuously updated and refined to enhance model performance in parallel to its usage, live over time. Methodologically, the SVM, LSTM and BERT models were trained on both individual and later combined datasets and then evaluated through Classification Reports, ROC curves, Confusion matrices and Accuracy and Validation Loss curves over each epoch to measure performance. Following this, the models were applied to generate predicted sentiment scores for the analysis dataset [33]. These scores were then used to train an SVR model, which predicted the stock prices for the last month of the analysis dataset's timeframe. The predictions were cross-validated with true stock price shifts at that time.

The findings indicated that the combined datasets led to a more robust sentiment analysis, with the models showing increased precision and accuracy. Particularly, the BERT model trained on both Twitter and Reddit datasets demonstrating near-perfect discriminatory abilities, correctly classifying 98.1% of the validation test set. However, SVR stock price predictions based solely on public investor sentiment expressed on Twitter showed limitations. This highlighted the demand for more complementary factors such as earning reports and other macroeconomic indicators. Therefore, it emphasises the importance of utilising diverse data sources for a more comprehensive and accurate market trend analysis. The combined dataset from Twitter and Reddit represented a wide range of opinions, perspectives, manners of expression, and different online cultures, ensuring sufficient diversity for a robust analysis and reducing model overfitting.

Thereafter, an extended research task was conducted to supplement the idea of another potential use case, where the best performing BERT model was chosen to analyse the investor sentiments focusing on external factors like political decisions, pandemics and wars. Against expectations of negative sentiment, the analysis showed universality of positive sentiments across tweets, which potentially was influenced by many factors as discussed. This relationship highlighted the complexity of sentiment analysis and interactions of socio-economic and psychological factors in shaping online discourse in regards to such significant events.

Table of Contents:

1. Introduction:.. 3

The Motivation:..3

The Aims of the project:... 4

SMART Objectives:..4

Reasoning behind the choices:... 5

2. Background:... 6

Related Work:..6

Datasets used:... 10

Resources used:...11

3. Methodology:..12

Ethical considerations:... 12

Exploratory Data Analysis:... 13

3.1 Theoretical Framework:..16

Support Vector Machine Model Framework:.. 16

Long Short-Term Memory Model Framework:..17

Bidirectional Encoder Representations from Transformers Model Framework:......... 19

Support Vector Regression Model Framework for stock price predictions:................20

4. Evaluation and Results of applied models:..22

4.1 Model training on Twitter Data:..23

Support Vector Machine Model Training Results:.. 23

Long Short-Term Memory Model Training Results:...24

Bidirectional Encoder Representations from Transformers Model Training Results:. 25

4.2 Analysis of sentiment predictions from model trained on Twitter data:..................... 26

SVM sentiment predictions:...27

LSTM sentiment predictions:..28

BERT sentiment predictions:..28

Visualisation of all sentiment prediction as comparison for the PG market:...............29

4.3 SVR stock price predictions from models train on Twitter data:..............................30

4.4 Training on Reddit and Twitter data:.. 33

SVM Model Training Results from the combined dataset:.. 34

LSTM Model Training Results from the combined dataset:.......................................35

BERT Model Training Results from the combined dataset:.......................................36

4.5 Analysis of sentiment predictions of models train on Twitter and Reddit data:.......... 38

4.6 SVR stock price predictions from models trained on Twitter and Reddit data............39

4.7 Extended Research: External factors study using the BERT model.......................... 42

5. Conclusion:.. 45

Discussion:...45

Summary of results:.. 46

Areas of improvement for the future:..47

6. References:...48

1. Introduction:

In financial markets, anticipating stock price movements has been crucial for investors looking to build their portfolios and as such has been an active research topic ever since. Traditionally, since the early 2000s, stock price prediction relied heavily on regression analysis of historical stock price data. However, with the advancements in deep learning in the 2010s, sentiment analysis has emerged as a powerful tool to integrate the effect of unstructured social information, such as investors' opinions, into stock price forecasting. A significant factor influencing public market stock prices is investor sentiment, which can be analysed through social media and financial news platforms. Natural Language Processing (NLP) tools are commonly used to evaluate the posts made by investors on social media, providing insights into their sentiments towards the markets they discuss.

The Motivation:

The motivation behind this study comes from our understanding of the ever increasing difficulty in analysing the complexity and volatility of financial markets to be used for future trend predictions. Investors continuously aim to seek new and innovative concepts and technologies to gain a more competitive advantage over their adversaries and to mitigate greater risks that may hinder their desires to improve their stock portfolios. Historically, the methods used for market analysis fell short in gathering a degree of understanding of the correlations between investor sentiment and true stock price shifts. Such as in the “Tweet Sentiment Analysis to Predict Stock Market’ [1] study having model accuracy scores from 70% to 83%, the latter being the score of the SVM model.

There are many studies that have proven that there is a strong positive correlation between online public and investor sentiment with stock price shafts that form the foundation of this study [42] [43] [44]. As such, utilising Natural Language Processing and Machine Learning algorithms could bridge this ever expanding gap by understanding the feasibility of using sentiment analysis for predictive modelling of stock market movements, to the benefit of a potential investor.

The Aims of the project:

This study aims to test the performance and accuracy of common processing technologies in predicting stock market shifts based solely on correlation between public investor sentiment against true stock prices in the same time frame. There are many different machine learning and deep learning algorithmic models to achieve such a feat, but the scope of the study is to compare three of the most common models used in industry.

The models of choice are the Support Vector Machine (SVM), the Long Short-Term Memory (LSTM), and the Bidirectional Encoder Representation from Transformers (BERT). These will be trained on historical data of Tweet sentiments over time to see whether they are positive, neutral or negative sentiments, and then these models will analyse a different dataset of time series sentiments with addition to stock prices, at the same time period of markets they're referencing. Thereafter, a Support Vector Regression (SVR) model will be trained on the provided time series sentimental data and stock price data for a specified market (control variable) and predict the latter month of the data's timeframe to compare the predicted stock price forecasts to the true stock price movement at that time.

Then if time allows, use the highest performing model from the cross validations to then predict the sentiments of the filtered tweets in relation to external factors such as war, pandemics and political decisions and investigate the effects this has on our training dataset.

SMART Objectives:

Develop and train three machine learning models: Support Vector Machine, Long Short-Term Memory, and Bidirectional Encoder Representation from Transformers - for sentiment analysis and stock market prediction using historical Twitter data and later Reddit data. Aim for a minimum sentiment prediction accuracy of 80% on the validation test set for each model as per GLUE (General Language Understanding Evaluation) benchmark standards [40] [41]. Compare models based on accuracy, precision, recall, and F1 scores. Gather historical Twitter data and stock price data, implement and fine-tune each model using appropriate tools, and ensure robustness through cross-validation. Contribute to understanding sentiment analysis in predicting stock market movements and evaluate model suitability for investors and researchers. Complete data visualisation and control variable selection within the first week; develop, train and evaluate three models within three weeks; visualise the model's predicted sentiment scores within another week; SVR predictions within two weeks and finally explore external factors' impact if time allows in the final week.

Reasoning behind the choices:

The reasoning behind the selection of models for sentiment analysis and subsequent stock market predictions comes from the comprehensive review of relevant background literature behind them which proves their efficacy in such similar studies.

The Support Vector Machine has demonstrated its effectiveness in sentiment analysis tasks due to its capability to handle high dimensional data and non-linear relationships between features and labels. Past research such as from the "Predicting stock movements with sentimental analysis of tweets with NNs"[5], highlighted the SVM's ability to produce high accuracy rates in sentiment analysis tasks related to stock market predictions.

The Long Short-Term Memory model, which is a form of recurrent neural network, excelled in gathering long term dependencies in sequential data, which made it well suited for sentiment analysis tasks involving time series data such as stock market sentiments. Studies like "Evaluation of deep learning techniques in sentiment analysis from twitter data"[7] underlined the LSTM's performance in sentiment analysis tasks, especially when facing 'temporal data'.

The Bidirectional Encoder Representation from Transformers, which is an attention based deep learning model showed high prominence for its exceptional performance in various natural language processing tasks across many studies. Such as from "Using BERT for sentiment analysis on news articles" [6], which highlights the successful application of the BERT model in sentiment analysis, as it yielded high F1 scores and accurately predicted stock movements like in the Dow Jones Industrial (DJI) stock market index.

Finally, the inclusion of a Support Vector Regression model appended to this study is supported by research such as in the "SVR modelling and parameter optimisation for financial time series forecasting" [21] study. Where they highlighted its effectiveness in forecasting parameters in the field of financial markets, as well as covering insights into how to optimise the parameters in such a model to reduce errors between predicted values and actual values in financial time series data. [22]

Changes since proposal:

There are three major changes since the project proposal:

- One, the complete redaction of using live sentiment data gathered through web scraping applications and API's to conduct our analysis on, which aimed to allow for the possibility of live predictions of stock prices for analysis. Due to ethical concerns around using web scraping for data gathering arose, it was changed to use historical data in its stead.
- Two, the implementation of an SVR model for stock price predictions against the gathered sentiment data with true stock prices, this was appended to the study as an experiment for a use case for a potential firm/investor with further research into the field of stock price predictions.
- Three, the addition of studying external factors such as pandemics, political decisions and wars, which were listed as a potential addition in the proposal.

2. Background:

Related Work:

Tweet Sentiment Analysis to Predict Stock Market Stanford. [1]

This study aimed to develop a Natural Language Processing model for stock market prediction using Twitter sentiment analysis with a transformer based neural network. They chose this method as they stated previous approaches often relied on statistical methods, rather than Machine Learning and Natural Language Processing techniques. They expressed a more unique focus of fine tuning the sentiment analysis component for improving the accuracy over anything else. This study also covered the importance of stock market prediction in modern economies; The difficulty to do so due to market volatility and unpredictability; The rapid reaction of news in stock prices on twitter and the suitability of twitter's real time analysis aligning with market movements; And how NLP advancements for sentiment analysis for nuanced sentiments in finance can classify bullish, bearish and neutral sentiment.

The study is relevant to our goal as it demonstrates the effectiveness of NLP and sentiment analysis in predicting stock market movements based on social media platforms, with both studies delving into fine tuning the sentiment analysis to improve accuracy. It also aligns with our project's aims of utilising similar techniques to analyse sentiment from various sources, including social media, for predictive modelling. But it has to be highlighted that it's a recent study with no known publish date (possibly 2023), with little to no direct citations, but will still be considered somewhat valid for this project due to its recency and relevance.

Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets. [2]

This study's research also focused on predicting stock market movements using historical prices and sentiment data, and the impact of computing and machine learning on research speed and new possibilities. They used two models, an LSTM model with historical prices as the independent variable and a Random Forest model incorporating sentiment analysis from the 'intensity analyser'. They also implemented additional macro parameters like Gold, Oil prices, USD exchange rates and Indian Government Securities yields were included to enhance the accuracy for their use case. They specifically targeted their predictions against the Reliance, HDFC Bank, TCS, and SBI markets and evaluated them using the RMSE metric (the square root of the mean of the square of all of the errors).

This study is quite relevant as it provides useful insights into utilising sentiment analysis and deep learning techniques to predict stock market movements. Although it specifically focuses on the Indian market, it can have interchangeable insights with our project and whilst using LSTM and Random Forest models to achieve this. While our project shares the goal of predicting stock market movements through sentiment analysis, it differs in its focus on building NLP sentiment analysis models using real time data from social media platforms as we will be using historical data, but still it offers a different perspective on predictive modelling. The study was published in 2022, making it only two years old and was cited 21 times, equating it to be very valid for our purposes.

Stock Trend Prediction Using News Sentiment Analysis. [3]

In this study, the focus was the challenging efficient market hypothesis with research on stock prediction using 'non-quantifiable data' like financial news articles. They aimed to study the relationship between news sentiment and stock trends. To do this they used three different classification models created to classify news polarity as positive or negative (but not neutral). They identified that the Random Forest and the Support Vector Machine performed exceedingly well in testing, with 'Naïve Bayes' showing good results but fairly inferior to the other two. The results gathered were considered encouraging, with prediction model accuracy surpassing 80%, which is up by as high as 30% compared to random labelling.

This paper is very relevant to our project, as it delved into the influence of nontraditional factors, such as sentiment from financial news articles, on stock market behaviours. Since our project also uses sentiment analysis from financial news along with social media data, it provides valuable insights into the efficacy of using news sentiment for stock market prediction making as well. This study validates the use of sentiment analysis from multiple sources, including financial news which supports our project's approach of adopting a similar analysis with model comparisons and cross validation. The paper is quite old, being published in 2016 making it eight years old now, but was cited roughly 224 times as of the date of recording, making it very valid for our desires in this project but its age will be considered.

Investor Sentiment and the Cross-Section of Stock Returns. [4]

This historical paper showed the challenge to classical finance theory's dismissal of investor sentiment. Their theory was: The investor sentiment affects a cross section of stock prices due to mispricing from uninformed demand shocks and exchange constraints. The study investigated the summary of the U.S. market sentiments all the way back from 1961 to the internet bubble. They approached this by examining the cross sectional predictability patterns in stock returns based on 'proxies' for beginning of period sentiment. This formation of sentiment indexes aligned with historical accounts of bubbles and crashes and their findings showed that the cross sectional variation is stock returns based on beginning of period sentiment with significant implications for different firms characteristics and decision making.

This paper is somewhat relevant to our project as it investigates the relationship between investor sentiment and stock returns, which aligns with our goal of understanding and utilising investor sentiment to predict stock market behaviour. While the source focuses on specific categories of stock returns in different markets, the overall theme of analysing the effects of investor sentiment remains consistent with our project study. Although the source may differ in the markets it studies, the overarching insights into the impact of investor sentiment on stock returns can provide some valuable context and support for our project's objectives. Now, this is a very old study, being released as far back as 2003 making it twenty one years old but only cited 44 times. Making it somewhat valid for our purposes but the lack of citations for its time spent in the public domain raises a few questions on its validity, but accepted nonetheless.

Predicting Stock Movement with Sentiment Analysis of Tweets with NNs. [5]

This study focused on utilising social media, specifically Twitter, for stock price predictions. They trained models on sentiment of 140 Twitter datasets and the SVM they used yielded the best results at 83% accuracy for sentiment analysis. The predictive models like the Boosted Regression Trees and Multilayer Perceptron Neural Networks were used for predicting specifically the stock prices of the AAPL and DJIA markets. They concluded that Neural Networks outperformed traditional models for stock price predictions and should thus be further developed.

This paper is highly relevant to our project as it investigates the use of sentiment analysis on Twitter data to predict stock market movements, aligning with our goals of using social media data for predictive modelling. Additionally, the source study's findings in regards to the effectiveness of SVMs provide valuable insights for our project, as we can use an SVM as a benchmark for comparison with other models in our study. This reinforces the significance of SVMs as a prime candidate for being the least resource intensive sentiment analysis model for stock market predictions. The paper was published in 2020, making it only four years old and cited 43 times, thus making it very valid towards its concepts and insights being implemented into our project.

Using BERT for sentiment analysis on news articles. [6]

This study utilised Natural Language Processing algorithms for automatic text analysis and used BERT for their sentiment analysis of new articles to aid in stock market decision making. They pretrained their BERT model on general domain documents, fine tuned it on manually labelled stock news articles for sentiment analysis. This resulted in achieving a 72.5% F1 score and the model's output used to predict subsequent movements of Dow Jones Industrial (DJI) stock market index as well.

This paper is very relevant to our project as they aimed to leverage sentiment analysis techniques, particularly using BERT just as we are, to analyse investor sentiment expressed through social media and financial news platforms for making informed investment decisions. Predicting stock market movements and utilising publicly available data sources to enhance predictive models and provide valuable insights for stakeholders and investors. It was also published in 2020 and cited roughly 39 times, also making it very valid for its insights to be implemented into our project.

Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data [7]

This study compared different deep learning methods for sentiment analysis in Twitter data and provided valuable insights and enhancements into how to evaluate and compare convolutional neural networks (CNN) and recurrent neural networks (RNN), particularly Long Short Term Memory (LSTM) networks. They identified the performance of these architectures in sentiment analysis tasks. Additionally, the comparison of word embedding systems such as ‘Word2Vec’ and ‘GloVe’ could offer additional thought to what could be implemented to the model as understanding the impact of different embedding techniques on sentiment analysis accuracy could be beneficial to us. The study also ensured a standardised dataset for evaluating model performance, enhancing the credibility of the project results. Moreover, the analysis of advantages and limitations of each method provided hints into model selection and insights into potential implementation challenges, additionally offering a standardised framework for rigorous evaluation against already established benchmarks.

This paper provided fairly relevant material for our project, including evaluations of different deep learning methods for sentiment analysis on Twitter data. Now, our project focuses on the usage and comparison of results from the SVM, LSTM and BERT models. The reference paper studies the effectiveness of CNNs like LSTM networks specifically, of which the insights will benefit the development of our own LSTM model. We both use natural language processing and systematic evaluation techniques to analyse online sentiments, just in different contexts. Both our studies will analyse the performances, advantages and limitations of many deep learning methods which the insights and results of which would benefit our study. The paper was published in 2019, only five years old, and was cited roughly 90 times, leading us to be very confident over its valid study to base on this.

Sentiment Analysis using Machine Learning and Deep Learning [8]

This paper discusses methods of sentiment analysis for data generated from internet platforms and offers more insights into the analysis of public sentiment, but it particularly focuses on the decision making processes for defence and government organisations. They leveraged data from platforms like Twitter, Reddit, and Facebook and they addressed the need to understand public sentiment to guide organisational actions and ensure decisions align with public sentiment. Especially during critical national events which could benefit our extended research project, which is quite similar, but it’s also quite similar to the overarching scope of our project, but we focus mainly on the financial sector and predictive powers such NLP systems that could benefit investors on stock market predictions.

They utilised machine learning classifiers and deep learning models for ‘polarity based sentiment analysis’, which enables the classification of user tweets as either positive or negative (but not neutral), giving some nuanced understanding of public sentiment. The inclusion of a variety of model architectures to train on diverse opinions and thoughts expressed on social media platforms, which enhances the robustness of the sentiment analysis. Furthermore, the classification models proposed can be implemented to classify live tweets on roughly any topic, extending the applicability of sentiment analysis techniques to real time data analysis (but we will be based on historical data analysis).

This paper aligns fairly well to our project as well, in terms of its focus to sentiment analysis on social media platforms. We both recognise the significance of analysing user generated data for decision making, and understanding public sentiment during critical national events. Our projects both use machine learning techniques for sentiment analysis but we concentrate our sentiment analysis models on stock market predictions where they analyse them for government organisation and decision makers, similar nonetheless. The paper was published in 2020, making it a fairly new paper only being four years old with 37 citations, leading to fair confidence in its validity towards utilisation into our study.

SVR Modeling and Parameter Optimization for Financial Data Forecasting. [21]

Last but not least, this paper was a study that compared developed countries' financial markets and optimised the forecasting ability over these markets using a SVR with their time series data. They modelled a SVR to conduct parameter optimisation research and extracted future predictions of markets and reduced the error between the predicted values and the accrual values as much as possible. This study will form the backbone for the stock price predictions using the true stock price and predicted sentimental scores produced by the previous SVM, LSTM, and BERT models.

It follows our project fairly well as its research is focused on SVR modelling, and more specifically, optimising it for financial time series (FTS) forecasting but this is in regards to state national financial markets that are still developing. This aligns with our project as we aim to utilise a SVR model for stock price predictions based on sentiment scores predicted from the previous different machine learning models.

This paper is relevant as it does state the importance of developing methods of financial market forecasting technologies for state benefits, but our project will use these insights into building and optimising an SVR to apply for our benefit of enhancing stock price prediction accuracy. The paper was published in 2023, making it extremely new in relation to other papers stated previously, only being a year old since the creation of this study, but was only cited once. The near to no citations to this study brings worry as its results and methods don't seem to be fairly well discussed or even cross validated, but is expected with newer papers. This will be taken into consideration during the creation of our SVR model but the insights can still be utilised for our development nevertheless making it somewhat valid.

A few notable others. [28] [29] [30] [31]

There were a few additional studies that also aided in the development of this study but not significantly, rather they acted as a benchmark for similar projects with similar goals that would help frame the structure of this study. They ensured that this study followed on their work but no technical insights was utilised from them, only more broad concepts.

Datasets used:

The Twitter and Reddit Sentiment Analysis Dataset, created for a university project using PySpark, contains 162,980 tweets and 37,250 Reddit comments labelled with sentiment scores (-1, 0 & 1). The data was gathered from Twitter and Reddit using Tweepy and PRAW APIs. This will be used to train our sentiment analysis models. [32]

The Stock Tweets for Sentiment Analysis and Prediction dataset comprises over 80,000 tweets related to the top 25 most watched stock tickers on Yahoo Finance from 30-09-2021 to 30-09-2022, along with corresponding stock market price and volume data. Each entry includes the date and time of the tweet, the full text of the tweet, the stock ticker name, and the company name. Inspired by similar datasets, this dataset can be used for sentiment analysis experiments, stock price prediction, and exploring the relationship between public sentiment and stock price movements. [33]

Resources used:

Google Colab-pro was utilised throughout this project to allow steady and consistent development of the software aspect of the project, this includes data visualisation, model training, cross validation, visual analysis of results, and the extended research project. Most visualisation methods were conducted and the free to use CPU provided by colab and the model training was conducted on the T4 GPU provided through the Colab-pro subscription to vastly reduce training time of the models, and instead of using onboard processing in house.

Software used:

- Python: The programming language used throughout the project for its versatility, ease of use, and extensive libraries for AI modelling. [45]

- Matplotlib: A library for creating static, animated, and interactive visualisations in Python. It was used for data visualisation tasks throughout the project. [46]

- Seaborn: Seaborn is a statistical data visualisation library based on matplotlib. It's used for creating informative and attractive statistical graphics in Python. [47]

- Scikit-learn (Sklearn): A machine learning library for Python used for various machine learning algorithms, including classification, regression, clustering, and ‘dimensionality reduction’. [48]

- Tensorflow.keras: Keras is a high level neural networks API that runs on top of TensorFlow. In the project, TensorFlow's Keras interface is utilised for building and training neural network models. [49]

- PyTorch (torch): Another machine learning library that provides a flexible framework for building and training neural networks. PyTorch was used for deep learning tasks and model experimentation. [50]

3. Methodology:

Ethical considerations:

Before any work is carried out, the ethical considerations surrounding the project's methodology have to be carefully and correctly considered. Firstly, to ensure data anonymity and compliance with open source regulations [26], which go against privacy concerns associated with gathering data from online social media platforms. By not using or even gathering user names we will be respecting government copyright laws [27] on handling such data. Additionally, the project maintains ethical integrity while using publicly available historical financial and sentimental data. [32] [33]

As such, we must consider the possible negative impact on investors' behaviour arising from the uncontrolled usage of data findings gained from this study. Hence a concise, clear and transparent communication strategy has to be implemented to ensure that the study's findings are read and understood only in the intended way. This is due to the very speculative nature of the analysis we are conducting and the possibility of influencing the market volatility, which adds to preventing unintended consequences such as promoting informed decision making amongst investors and the public. Measures will be taken to counter such unintended consequences, like misinformation or panic within the market by thoroughly judging the potential impact of external factors that could contribute to such an effect, such as pandemics, political decisions and wars, the research aims to minimise these repercussions as much as possible and stand with common ethical standards when releasing our research findings.

Thereafter, safeguards are produced to prevent market manipulations by conducting the entire study discreetly and whilst only using historical processed data for our training and analysis instead of live data that isn't censored. Thereby maintaining strong confidentiality within the study and preventing public release and access to results until after the completion of the study in question. The project will also aim to prevent competitors from exploiting the findings for their own financial gains, thus holding down the idea of fairness and integrity within the financial markets, no matter the size and impact of such findings.

Overall, the ethical considerations carried out show a strong, robust and fairly committed effort to conduct the research with reasonable responsibility and be ethically feasible, whilst adding research and development into the field of sentiment analysis for financial marketing.

Exploratory Data Analysis:

Firstly, an analysis was conducted on the chosen datasets [32] to identify any features that may skew or disadvantage the study's analysis and to select any control variables if needed. The Twitter training dataset for the models contains exactly 162,980 posts and the sentiment score of which are distributed as such:

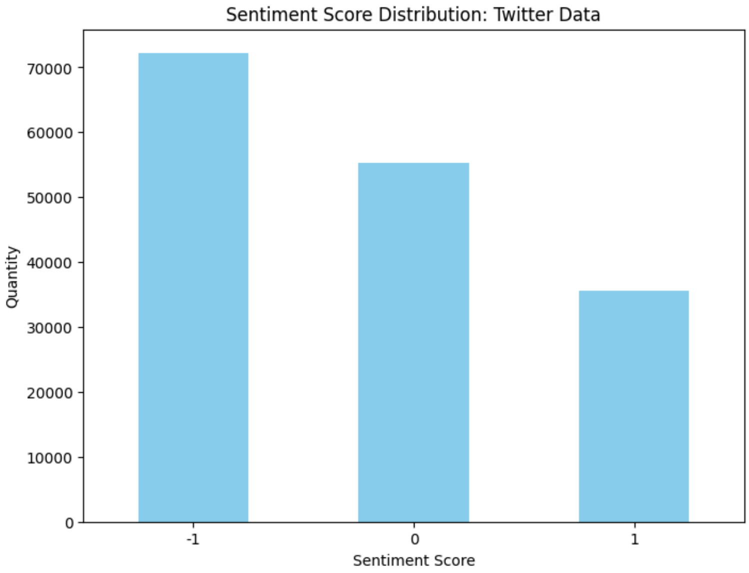


Fig. 1. Quantity distribution of sentiment scores in the Twitter dataset. [33]

As the data shows in Fig. 1, there is a discrepancy between the distribution of the sentiment scores, there is roughly twice as many negative sentiments than there are positive, with neutral sentiments placing somewhat in between. This may lead to a bias towards the negative sentiments during the training process, as the negative sentiments will have more comprehensive and more varied vocabulary relative to the other classes.

Here is the same visualisation but for the additional Reddit training dataset's sentiment score distribution, also showing a very similar uneven distribution of sentiment scores in Fig. 2:

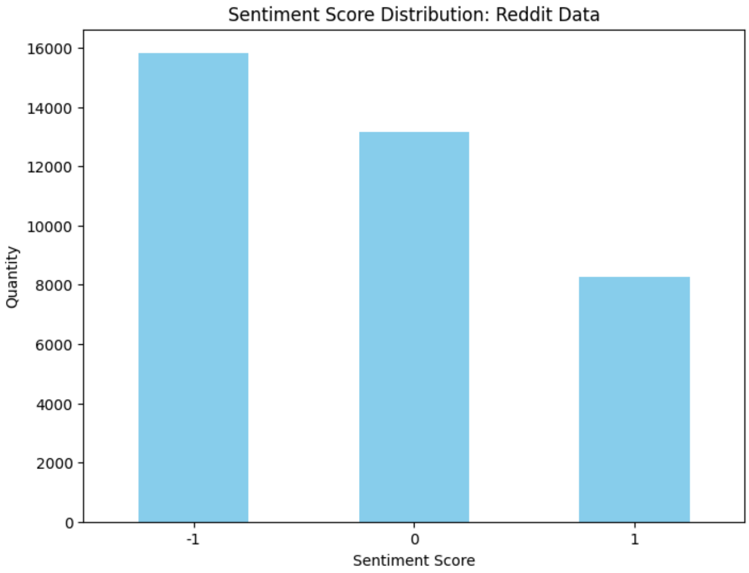


Fig. 2. Quantity distribution of sentiment scores in the Reddit dataset. [33]

Although these discrepancies could lead to biases during the training process, equalising the classes, i.e. cutting the negative and neutral sentiments to have equal quantity of sentiments, would lead to needless data loss that could still be vital for the model training. For this study, all data will be utilised as per the default stock quantity of the data classes provided. This decision was made as the quantity of the positive and neutral sentiments are still somewhat substantial for the models to use, as the excess of the negative sentiments will serve to provide a more accurate predicted classification for that class. But also, it will reflect the real world quantity of sentiment classes from posts gathered from online platforms such as Twitter and Reddit.

Additionally, the analysis dataset [33] that holds exactly 80,793 tweets, of which models will predict the sentiment scores of, are grouped by the stock markets of which the tweets are referencing. The distribution of these referenced markets (Ticker names) are as such :

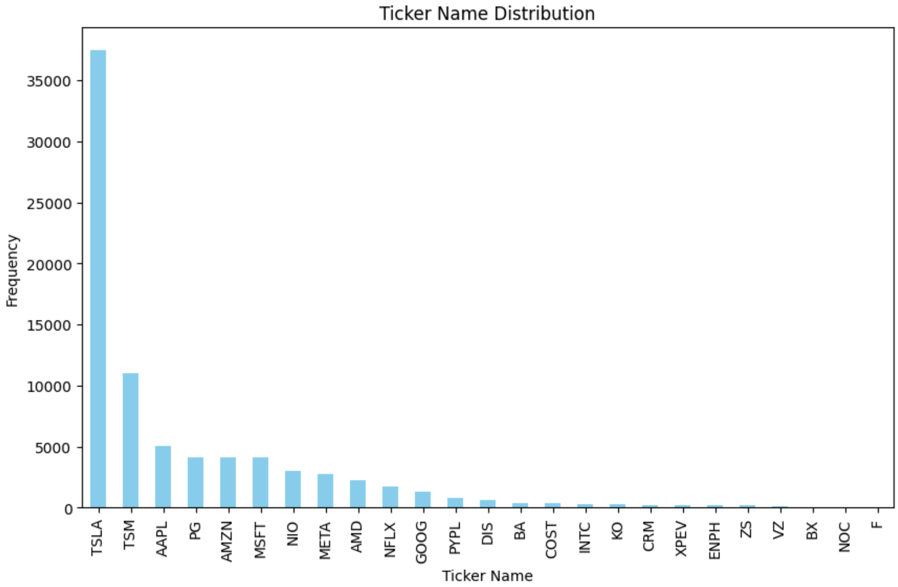


Fig. 3. Quantity distribution of stock market references in analysis dataset. [32]

The data shows in Fig. 3, that there is an extreme discrepancy between the quantity of markets the analysis dataset [33] contains. This indicates that experiments with markets on the far end with vastly less quantity of sentiments in reference to that market may lead to inadequate results over time, such as gaps and voids that could skew the mean average score of each week. To counter this, experiments and analyses should only encompass the greater half of markets shown in the graph, from “TSLA” to “GOOG” (tesla and google markets), leaving the latter half to be unused for experimentation.

Additionally, as seen in the data distribution, the TSLA market has a massive excess of sentiments in relation to its market. This may be due to the popularity and influence the founder of the company has on online platforms like Twitter, of which Elon Musk has ownership of, rebranding it as ‘X’ now. Because of this popularity, sentiments drawn from this market may not accurately represent Tesla’s decisions but rather the whole family of firms Elon Musk controls. As such it won't be utilised for our analysis due to these biases. So finally, the selected and most appropriate market to act as our control variable for the analysis of the models performances will be the PG market (Procter & Gamble Co).

Thereafter, an inspection was laid out over the time series stock price data to ensure no data is missing, as seen in Fig. 4. As if so, imputations would have to have been carried out over any gaps in the stock prices to prevent errors with expected sequence size of the input data for the model. But luckily, the dataset appears to be consistently complete as fully observant.

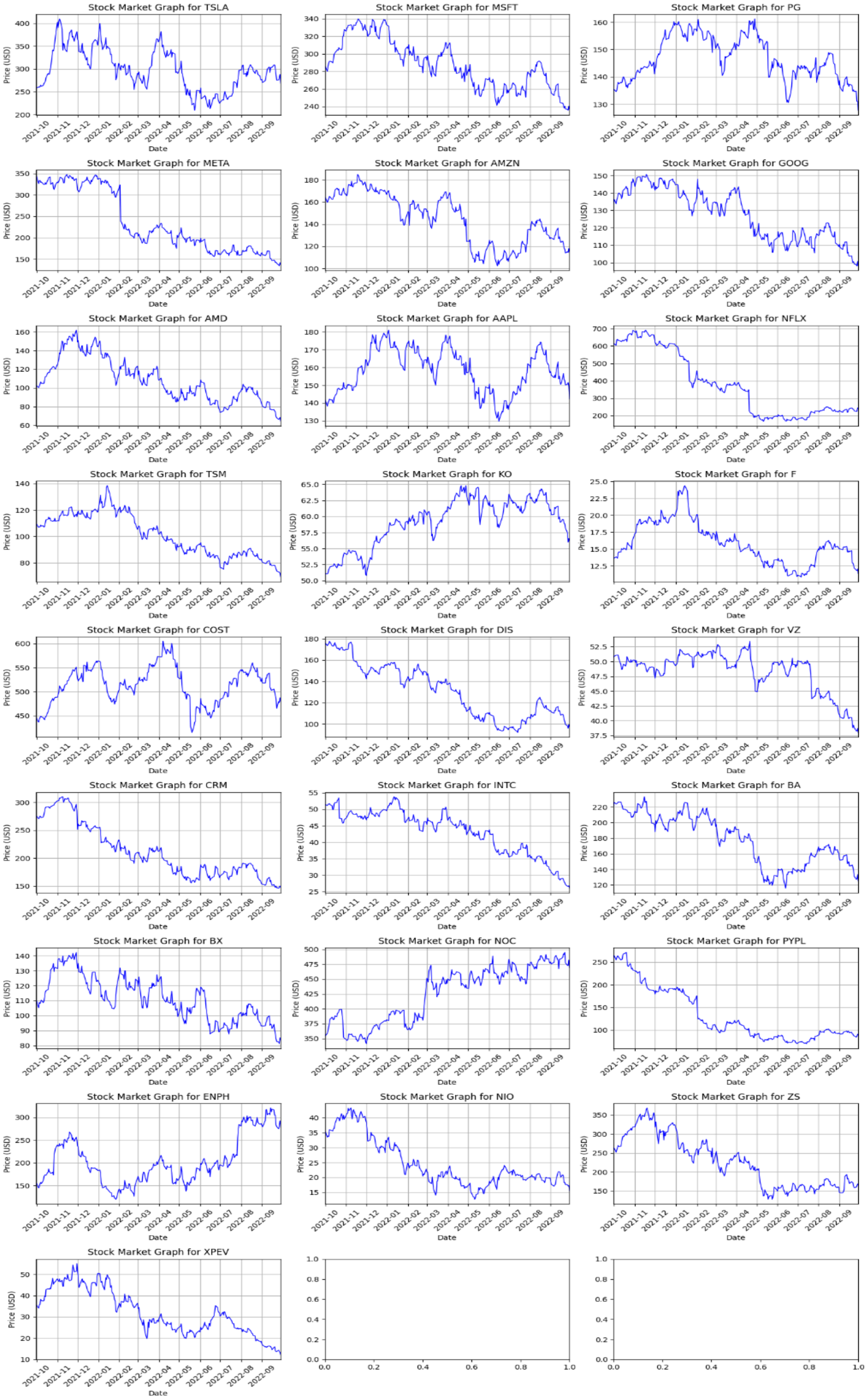


Fig. 4. Stock prices over time for each stock market in the analysis dataset. [32]

3.1 Theoretical Framework:

Here the theoretical overview of the selected model frameworks show how these architectures operate and differentiate from one another. Ranging from the simple regression task based SVM, to a gated RNN: LSTM model and even a Deep learning transformer based architecture model with bidirectional understanding of the data for greater contextual understanding as well as with pretrained models for english language understanding.

Support Vector Machine Model Framework:

SVMs are a powerful class of supervised learning algorithms used for classification tasks. It works by finding the optimal hyperplane that separates different classes in the feature space. The hyperplane is defined by a decision function:

$$f(x) = w \cdot x + b \,, \qquad (1)$$

where x is the input feature vector, w is the weight vector and b is the bias term. The main goal of the SVM is to maximise the margin, which is the distance between the hyperplane (1) and the nearest data point of each class, in our case -1, 0 and 1 for negative, neutral and positive classes. All in all, this is an optimisation problem where the aim is to minimise the 'squared Euclidean norm' [34] of the weight vector which is subject to the fact that each data point is correctly classified. This can be represented by the inequality for all points in i:

$$min_{w,b} \frac{1}{2}||w||^2 \,, \qquad (2a)$$

subject to these constraints for all points in i are true:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i \,, \qquad (2b)$$

where $||w||$ in (2a) represents the Euclidean norm of the weight vector. y_i in (2b) is the class label for the ith data point. x_i is the ith data point and b is the bias term.

In instances where the data Isn't linearly separable, a soft margin SVM is used to allow for some misclassifications, which is achieved by introducing slack variables ξ_i in (3a) which is the quantity of misclassification for each data point. The optimisation problem is then modified to include a regularisation variable C in (3a) which handles the tradeoff between maximising the margin and minimising the misclassification. This function can be represented as such:

$$min_{w,\,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} = \xi_i \qquad ,$$

(3a)

subject to these constraints for all points in i are true:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \,, \qquad (3b)$$

$$\xi_i \geq 0 \quad \forall i \,. \qquad (3c)$$

One of the main strengths of an SVM model is its ability to handle nonlinear classification tasks using the kernel trick. This is done by mapping the input feature space in a higher dimensional space, which the classes can then be separated linearly. The decision function is described as a linear combination of kernel functions that were applied to the input data points, weighted by what's called Lagrange multipliers collected from the optimisation problem. The decision function can be computed as such:

$$f(x) = \sum_{i=1}^{N} = a_i y_i K(x_i, x) + b \ , \quad (4)$$

where $K(x_i, x)$ in (4) is the kernel function, and a_i are the Lagrange multipliers collected from the optimisation problem. By utilising this kernel trick, the SVM can effectively handle more complex data distributions and achieve higher classification accuracy. [12] [13]

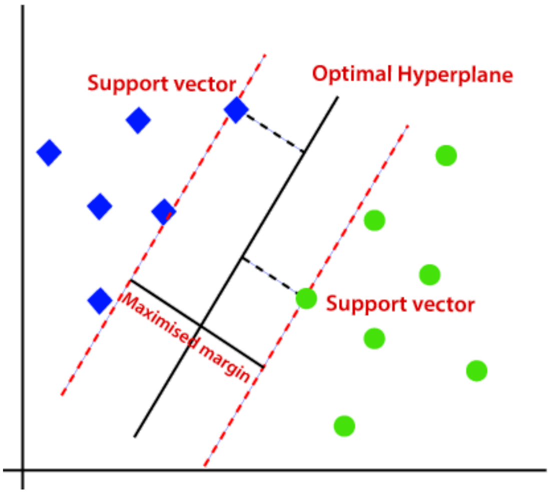


Fig. 5. Diagram of SVM classification against the linear optimal hyperplane [9]

Long Short-Term Memory Model Framework:

The Long Short-Term Memory (LSTM) model is a type of Recurrent Neural Network (RNN) developed to capture long range dependencies in sequential data. It differs from more traditional RNNs as it's built with methods to selectively hold or forget information over time, which makes it more suited for natural language processing and time series prediction, which is exactly why it's perfect for our task. Additionally, an LSTM model has a set of gating mechanisms that control what information is passed through which is quite useful and uniquely relative to the other models, they are as follows:

Input Gate:
The input gate determines the quantity of data from the current input that should be whitelisted into the cell state, it's computed as such:

$$i_t = ReLU(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \ , \quad (5)$$

where in (5), x_t is the current input, h_{t-1} is the previous hidden state, c_{t-1} is the previous cell state, W_{xi},W_{hi}, and W_{ci} are the weight matrices for the input, hidden and cell state respectively.
b_i Is the bias vector, and $ReLU$ or sometimes notated as $max(0, x)$, is the activation function which was chosen to be used for our practice, in replacement of the traditional sigmoid activation function σ.

Forget Gate:
The forget gate determines the quantity of data from the previous cell state which should be forgotten and is computed as such (carrying on with the same variables as before):

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \ . \qquad (6)$$

Output Gate:
The output gate controls the quantity of data where the current cell state that will be used to calculate the output and is computed as such:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \ . \qquad (7)$$

Cell State:
The cell state represents the cell’s internal memory at a specific time t, each time it's updated with the input and forget gates and a candidate value. It’s computed as such:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \bar{c}_t \ . \qquad (8)$$

where the c̄ in (8), which is commonly notated with a tilde (~) on top, is the candidate cell state.

Hidden State:
The hidden state is the output of the LSTM’s cell at time t. It differs from the cell state as it’s for predicting or can be passed to other layers, and is computed as such:

$$h_t = o_t \cdot tanh(c_t) \ . \qquad (9)$$

Candidate Cell State:
The candidate cell is the new candidate value that could possibly by appended to the cell state and is computed as such:

$$c = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \ . \qquad (10)$$

Update Cell State:
Finally, the update cell state is an amalgamation of the forget gate’s output data after the previous cell state is passed through and the input gate’s output data which is the new data to add. This combination can then be modified by the candidate cell state which results in the creation of the updated cell state at time t. [14] [15] [16]

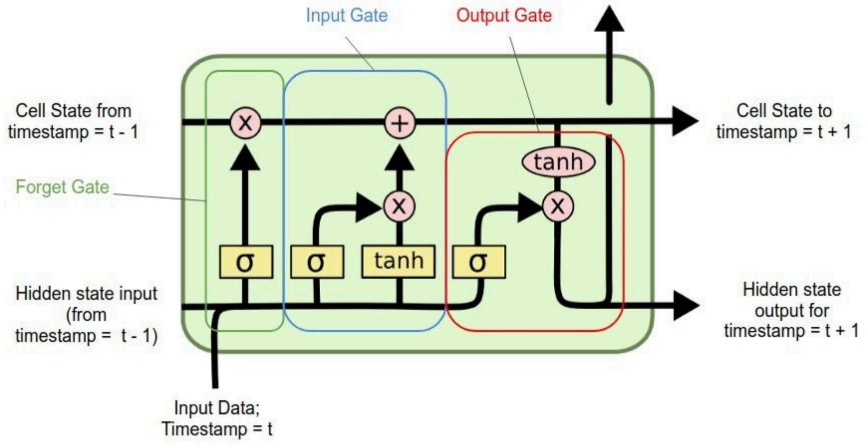

Input Gate
Output Gate
Cell State from timestamp = t - 1
Cell State to timestamp = t + 1
Forget Gate
tanh
σ
σ
tanh
σ
Hidden state input (from timestamp = t - 1)
Hidden state output for timestamp = t + 1
Input Data; Timestamp = t

Fig. 6. Diagram of the common LSTM gates [10]

Bidirectional Encoder Representations from Transformers Model Framework:

The Bidirectional Encoder Representation from Transformers, is a language representation model developed by Google. It utilises transformer architecture which is a form of neural network architecture best suited to handle sequential data fairly effectively and efficiently. BERT models on the other hand have a new innovation, that is the ability to understand the context of words in a sentence by comprehending both the left and right context relative to that word simultaneously, where the term bidirectional comes from. But in brief, a typical BERT model works as such:

Tokenisation process: where it first tokenises the input text into separate tokens, these may be as words, sub words or characters depending on the chosen methods, the chosen method for our implantation of the model was subwords. This means that words are broken down into smaller meaningful units called subwords; this was chosen to allow the BERT model to handle 'out of vocabulary' words by breaking them down into subwords that it recognises.

Input representation: where each token is then represented by an embedding vector which allows the embeddings to capture the semantic meaning of the token in context to the input data.

Transformer encoder: where it applies a stock of transformer encoder layers used to process the input tokens. With each encoder layer made of 'self attention mechanisms' [35] and feed forward neural networks.

A self attention mechanism [35] calculates the attention scores between pairs of all tokens in the input sequence, which would then compute how much each token should attend to other tokens in the sequence.

Additionally, there are contextualised embeddings which oversee the whole input sequence through the self attention process. The model then generates contextualised embeddings for each token. These embeddings capture the token's meanings but also its relationship with other tokens in the input sequence.

But the main factor of the BERT model is its utilisation of pre-trained objectives, for our implementation the model is initialised with a pretrained BERT model for sequence classification using the weights from the 'bert-base-uncased' model [36]. This allows the model to load the pretrained weights and architecture from the Hugging Face 'transformers' library. Which was built for BERT base model training, specifically on uncased English text, ideal for our requirements.

Finally, after the pre-training on English uncased text [36], the model will then be fine tuned on a downstream task, for example: text classification, named entity recognition, question answering or etc.., but for our application it was fine tuned for text classification for sentiment analysis for our Twitter and later Reddit data provided.

The self-attention mechanism in the Transformer architecture can be computed with an given input sequence token embedding:

$$X = \{x_1, x_2, ..., x_n\}. \tag{11}$$

Then the self attention mechanism computes attention scores A, computed as such:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right), \tag{12a}$$

where: $Q = XW^Q, K = XW^K, and\ V = XW^V,$ (12b)

are the predictions of the input embeddings X in (11), with weight matrices W^Q, W^K, W^V in (12b), with d_k is the dimension of the main vectors. The Softmax function is used to gather normalised attention scores for each token and the output of the self attention mechanisms [35] is competed as $Y = A \cdot V$, where Y holds the weighted sum of the vectors V which is based on the attention scores A. [17] [18] [19]

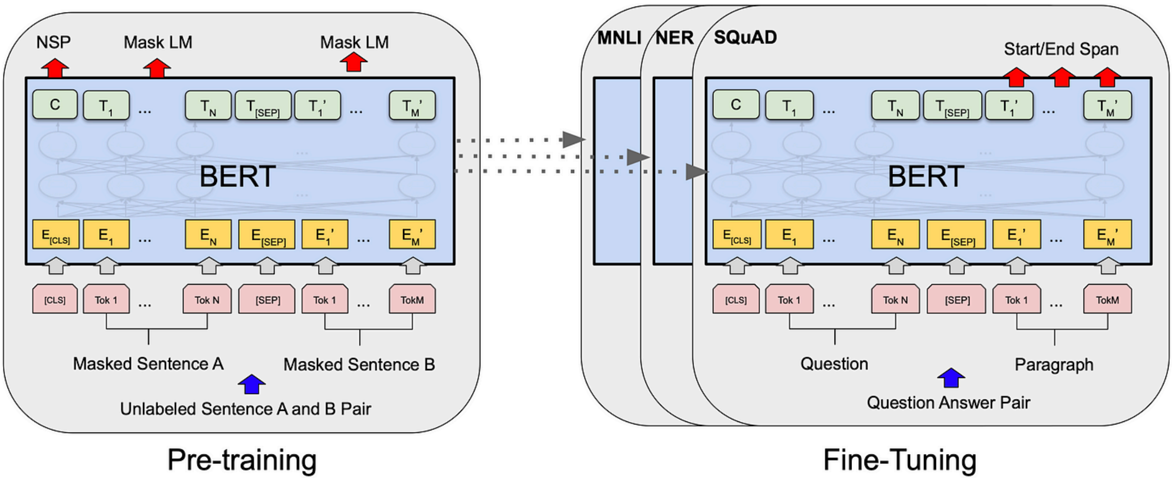


Fig. 7. Diagram of pre-training on unlabeled data and Fine-Tuning of labelled data and pretrained parameters a BERT model [11]

Support Vector Regression Model Framework for stock price predictions:

The support vector regression (SVR) model is a variant of the support vector machine (SVM) specifically designed for regression tasks. Similar to the SVM for classification, the SVR aims to find what's known as the hyperlane in the feature space that best fits the training data, whilst maintaining a maximum margin. But it focuses on finding the hyperplane that fits as many data points as possible in a specific margin instead of trying to fit all data points exactly. This is done by introducing a margin of tolerance ε, providing some deviation from the predicted value.

As such, this model will be used to predict the last month of stock prices in the time series dataset for a given market using the sentiment scores predicted by the previous three models with the true stock prices during the same time frame of that specified market. Now the theory behind it is as follows: Given the training data of {(xi, yi)} where xi is the feature vector and yi is the target value, the SVR aims to find the optimal hyperplane defined as:

$$f(x) \ = \ w \cdot x \ + \ b \ , \qquad (13a)$$

subject to these constraints:

$$yi \ - \ (w \cdot xi \ + \ b) \ \leq \ \varepsilon \ , \qquad (13b)$$

$$(x \cdot xi \ + \ b) \ - \ yi \ \leq \ \varepsilon \ . \qquad (13c)$$

Where, ε in (13b and 13c) represents the margin of tolerance, which allows for deviations between the predicted values and the actual target values, but the aim is to minimise the error while keeping within the margin of tolerance. Additionally, the SVR uses a parameter C in (14a) for regularisation, similar to the SVM for classification. This manages the trade-off between maximising the margin and minimising the error, this can be computed as such:

$$minimise \ 0.5 \cdot ||w||^2 \ + \ C \cdot (\Sigma\xi i \ + \ \Sigma\xi i \ *), \qquad (14a)$$

subject to these constraints:

$$yi \ - \ (w \cdot xi \ + \ b) \ \leq \ \varepsilon \ + \ \xi i \ , \qquad (14b)$$

$$(w \cdot xi \ + \ b) \ - \ yi \ \leq \ \varepsilon \ + \ \xi i \ *, \qquad (14c)$$

$$\xi i, \ \xi i \ * \ \ \geq \ 0 \ . \qquad (14d)$$

Where ξi and ξi * in (14a, 14b, 14c and 14d) are what's known as the ‘slack variables’ in regards to the i’th training instance which allows for deviations from the margin of tolerance.[20] [21]

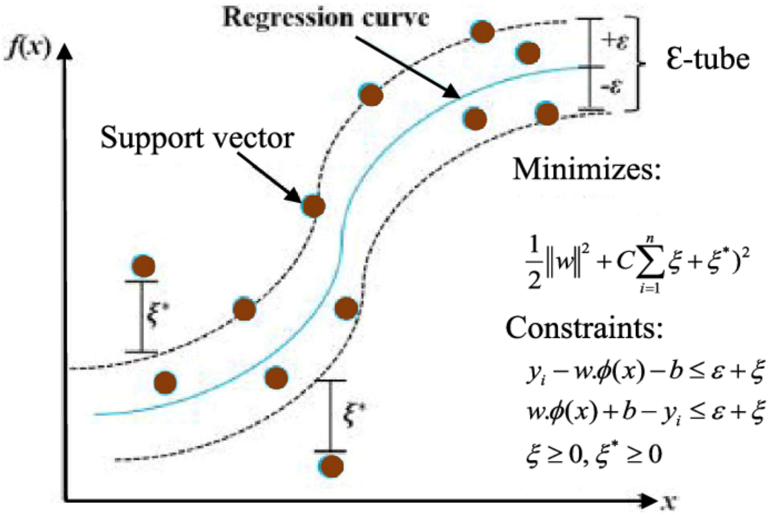

Regression curve
f(x)
+ε
-ε
ε-tube
Support vector
Minimizes:
$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi + \xi^*)^2$
Constraints:
$y_i - w.\phi(x) - b \leq \varepsilon + \xi$
$w.\phi(x) + b - y_i \leq \varepsilon + \xi$
$\xi \geq 0, \xi^* \geq 0$
ξ^*
ξ^*
x

Fig. 8. Diagram of SVR classification against a regression curve [23]

4. Evaluation and Results of applied models:

Here each model is trained based on a dataset of tweet sentiments and their true sentimental scores beside them. Each model is to predict a classification of each sentiment from 80% of the dataset acting as the training set and then test their results against the other 20% of the data, acting as the testing set. As per the Pareto Analysis standard [38] [39].

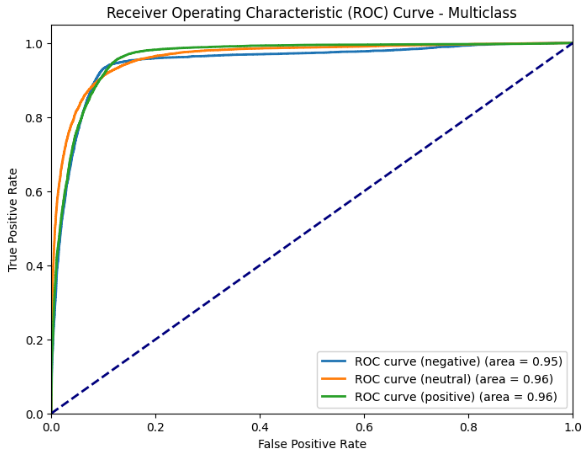
Now, to evaluate their performances, a multiclass Receiver Operating Characteristic curve (ROC curve) is used to plot the true positive rate against the false positive rate for each threshold. As such, it will show the performance of the model's classification of the three classes, -1, 0 and 1 for negative, neutral and positive classification of sentiments from the test dataset. The closer the Area Under the Curve (AUC) is to 1 (excellent score) the better, as it shows perfect discriminating abilities for classification with 0 being the worst possible score.

Additionally, a confusion matrix will be used to visualise the true value of each classification and graph whether the predicted values are accurate to the true sentimental values for each sentiment in the test set. The greater the scores are along the diagonal (top left to bottom right) the better, as the predicted classes align with the actual classes, with as few values beyond this diagonal being the most optimal as they represent false positive classifications. But false and true positives are terms used for Binary classifications (two classes) but we use three classes in our study. So we incorporate the 'One-Vs-All' approach to classifying the false and true positives from here on, representing correct and incorrect classifications evaluated by the models after training.

Finally, the LSTM and BERT models were also evaluated with Validation Accuracy and Loss graphs to visualise the increase of accuracy and drop in loss values after each epoch. This is to measure whether the accuracy could increase if there were any more epochs implemented. Or whether the returns after any further additions wouldn't be beneficial relative to the time and resources needed to train the model, which would be evident with plateauing curves on either of the graphs. But also proof that maybe the last epoch may not be the best performing one and as such would not be taken as the best performing model for the analysis steps.

4.1 Model training on Twitter Data:

Support Vector Machine Model Training Results:



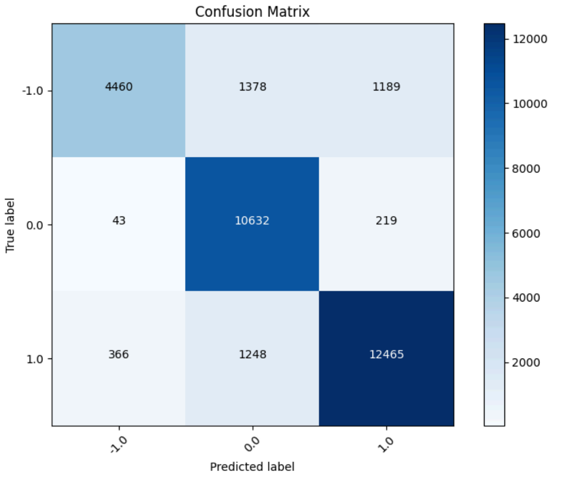


Fig. 9. ROC curve of SVM classifications, train on Twitter data.

Fig. 10. Confusion Matrix of SVM classifications, trained on Twitter data.

The Support Vector Machine took roughly three and a half minutes to train with the 160k sentiments passed through. The resulting classification report achieved 0.92, 0.80 and 0.90 precision scores; 0.63, 0.98 and 0.89 recall scores; and 0.75, 0.88 and 0.89 f1-scores for the negative, neutral and positive classes respectively, achieving an overall accuracy of 0.86.

The SVM's multiclass ROC curve for the analysis sentiment classification accuracy with the given data set indicates a high degree of performance across the three classes. With the Area Under the Curve (AUC) values in Fig. 9 of 0.95, 0.96 and 0.96 for negative, neutral and positive sentiments respectively, the model demonstrates robust discriminatory abilities. This suggests that the model effectively distinguishes between the different sentiment classes, with minimal false positive rates and high true positive rates. With such high AUC values typically identifying a strong predictive capability of the model across different thresholds.

In regards to the confusion matrix in Fig. 10, it revealed that the model performed well in correctly identifying negative and positive sentiments with evidence by the high counts along the diagonal for the three classes. However, there is evidence of some difficulty in accurately classifying neutral sentiments, as shown by the off diagonal counts. The model showed some characteristics of misclassifying neutral tweet sentiments as either negative or positive more frequently than the opposite. Despite this, the overall performance is commendable, with a majority of predictions aligning with the true sentiment labels and is more than sufficient for such acute training times.

These characteristics may lead to form an idea that even though the model is far from perfect, its sufficiency over the time needed to train from a large dataset, could be well suited for real time applications with new datasets provided over time to build upon a model of such sentimental classifications problems.

Long Short-Term Memory Model Training Results:

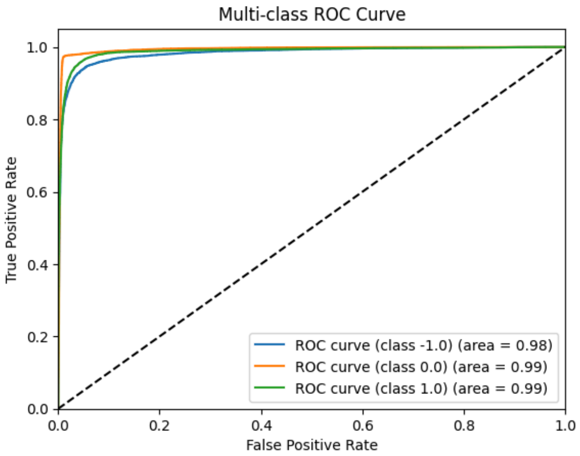


Fig. 11. ROC curve of LSTM classifications,
Trained on Twitter data.

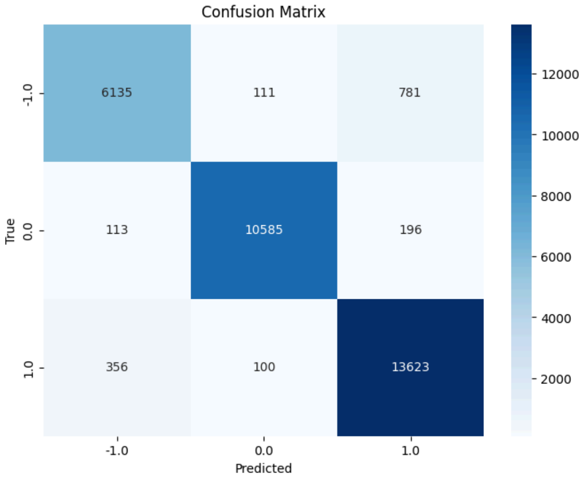


Fig. 12. Confusion Matrix of LSTM classifications,
Trained on Twitter data.

The Long Short-Term Memory took roughly thirty six minutes to train with on average six minutes per epoch for six epochs, as you'll see why six epochs were chosen as there was clear evidence of diminishing returns and lacked further improvement for such few gains. The resulting classification report achieved 0.93, 0.98 and 0.94 precision scores; 0.90, 0.97 and 0.97 recall scores; and 0.92, 0.98 and 0.95 f1-scores for the negative, neutral and positive classes respectively, achieving an overall accuracy of 0.95.

The LSTM's multiclass ROC curve in Fig. 11 showed high AUC values of 0.98, 0.99 and 0.99 for negative, neutral and positive classes respectively, identifying that the model demonstrated an outstanding discriminatory capability. With such high AUC scores, the model showed that it can effectively distinguish between different sentiment classes with minimal false positives and high true positives, highlighting its robust predictive capabilities.

In addition, examining the confusion matrix in Fig. 12 showed further clarification of the model's performance as it illustrates a strong ability to correctly classify tweets into their respective sentiment classifications. The LSTM model shows particularly impressive precision with identifying neutral sentiments with the majority of them correctly classified, but overall the confusion matrix underlines the model's ability in accurately predicting the sentiment labels with the data set provided.

Furthermore, an evaluation of the training in Fig. 13 and Fig. 14, validation accuracy and loss over the epochs was carried out to prove that further epochs would only provide diminishing returns and finding a valid amount with minimal training time:

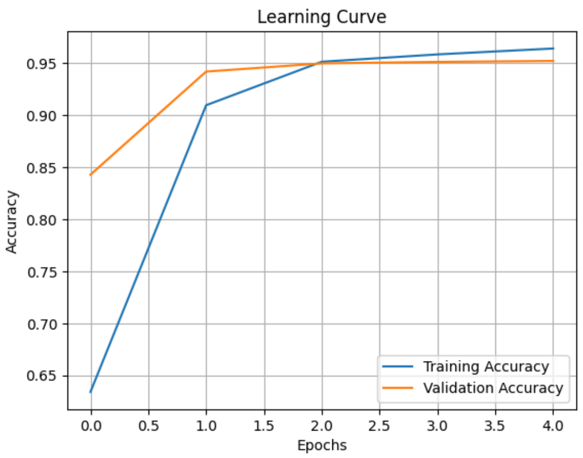


Fig. 13. Learning curve over the LSTM's epochs, trained on Twitter data.

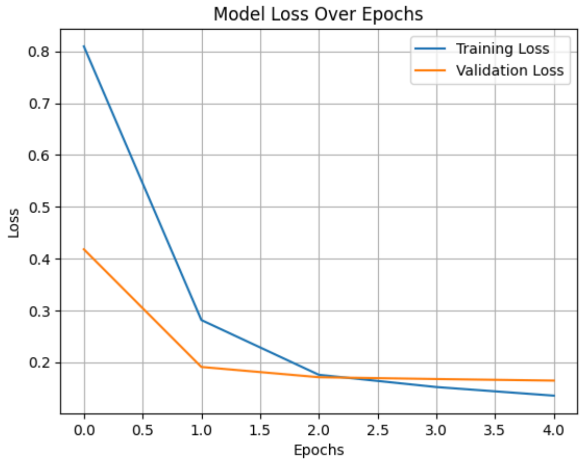


Fig. 14. Validation loss curve over the LSTM's epochs, trained on Twitter data.

Bidirectional Encoder Representations from Transformers Model Training Results:

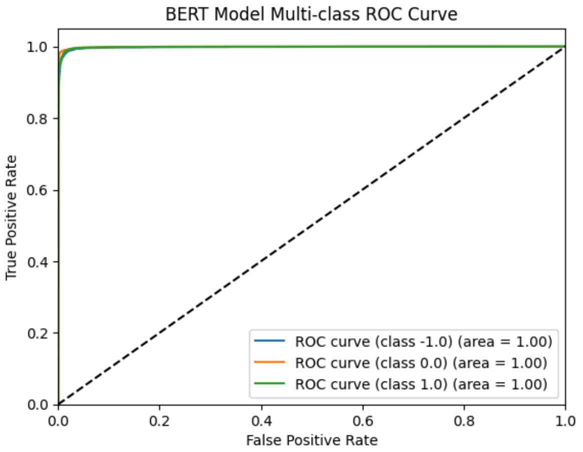


Fig. 15. ROC Curve of BERT classifications, trained on Twitter data.

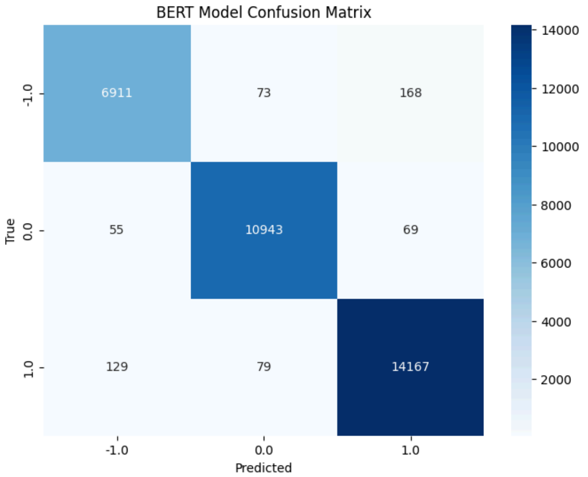
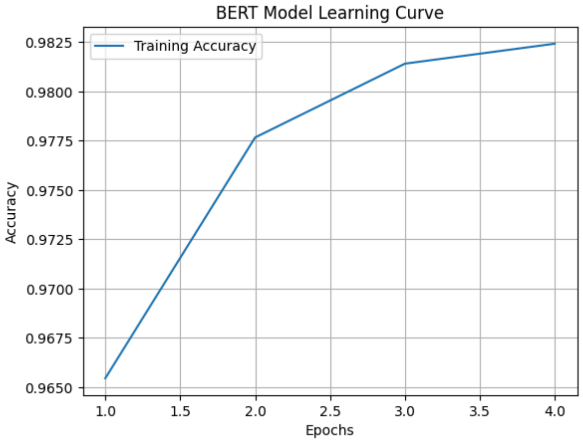


Fig. 16. Confusion Matrix of BERT classifications, Trained on Twitter data.

The BERT model showed fairly remarkable performance characteristics; the ROC curve analysis identified that there was a consistently perfect Area Under the Curve (AUC) in Fig. 15 with scores of 1 across all the classes. This indicates that the BERT model shows near perfect discriminatory abilities with distinguishing between the positive, neutral and negative sentiments. Additionally, in relation to the confusion matrix in Fig. 16, the model's performance further clarified this claim as the predictions are near perfect across all sentiment classes. Most notably, the model achieved 97.4% accuracy in identifying 6,911 out of 7,095 negative sentiments correctly predicted, 98.6% accuracy in identifying 10,943 out of 11,095 neutral sentiments and 98.3% accuracy in identifying 14,167 out of 14,404 positive sentiments.

Furthermore, an evaluation of the training accuracy and loss over the epochs was also carried out to prove that further epochs would only provide diminishing returns shown in Fig. 17 and Fig. 18. But an interesting factor from the BERT model over the LSTM model was its extremely high accuracy on the first epoch's results at approximately 0.96 where the LSTM model performed roughly only 0.65:



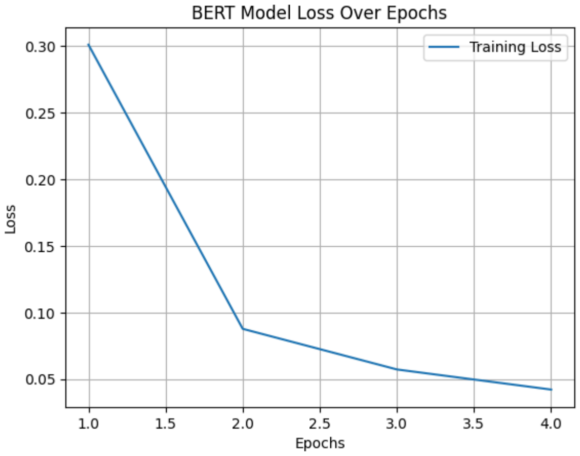


Fig 17. Learning curve over the BERT's epochs, trained on Twitter data.

Fig. 18. Validation loss curve over the BERT's epochs, trained on Twitter data.

4.2 Analysis of sentiment predictions from model trained on Twitter data:

Analysing the model's predictions against true historical data, a market has to be chosen to act as the control variable that had a sufficient quantity of sentiments in relation to that market over the course of the stock data time frame. Issues arose due to the disparities mentioned before with the markets in the data set and some other external factors such as constant positive sentiment for a market like TSLA (Tesla). This may be due to the founder's public opinion from other companies under the same branch, or even his involvement with owning the Twitter (X) social media application itself. So for the experimentation of comparisons between the predicted sentiments of each model, the PG market (Procter & Gamble) was chosen to be the most suitable control variable.

Additionally, some other pre and post processing had to be carried out to achieve these evaluations which may affect the results slightly in different manners. The quantity of sentiments in relation to the PG market is slightly above 4000 over the course of a year in the dataset, but the data points for the stock pricing over time differs from this. Which does not affect the human understanding visually of the relationship or even correlation between the two, but rather would be a cause of issue for further experimentation, i.e. using the correlation to construct predictions of the stock price solely from the sentiment and stock price relationship, which will be carried out later.

To counter this, both the predicted sentiment scores and the stock price of that market are represented by the mean values they hold for each week. This way no matter the change in control variable or quantity of sentiments or stock price data points, both of which will be equal in amount. With 48 points as the data time frame holds 48 weeks worth of data, meaning that there are 48 data points of mean weekly sentiment score and stock prices.

SVM sentiment predictions:

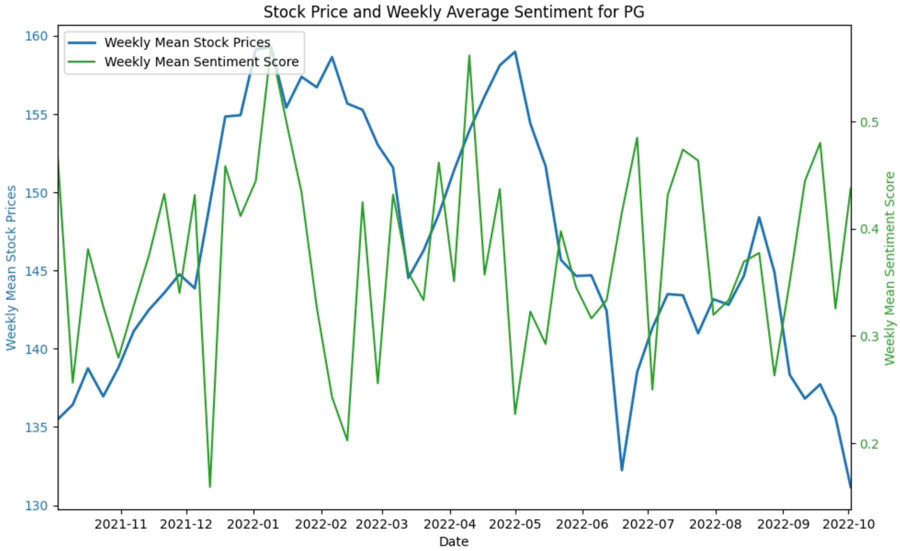


Fig. 19. Twitter trained SVM predicted weakly mean sentiment score in green and true weakly mean stock prices in blue.

Here, shown in Fig. 19, we can see the mean weekly sentiment scores of the SVM model over the mean weekly stock prices of the control variable PG market. It shows some correlations between peaks and troughs between the data frames, as seen with the high sentiment scores over the high stock prices at the weeks following 2022-01 and 2022-04 and three more volatile peaks following this. But some troughs in the sentiments appear to negatively impact the stock prices between two weeks to a month in advance as seen in the trough at 2022-20 with the stock price trough in mid 2022-03, and again with the sentiment trough in the later half of 2022-04 with the stock price dip in 2022-05 that lasted all the way up to mid 2022-06.

This does seem to be fairly volatile even though these scores are the mean scores for each week which should have stabilised these volatilities in retrospect. There do seem to be some weeks of extreme mean weekly sentiment values as seen in the first half of 2021-12 with a massive dip that does not seem to affect the stock prices at that time, which could indicate a local outlier.

LSTM sentiment predictions:

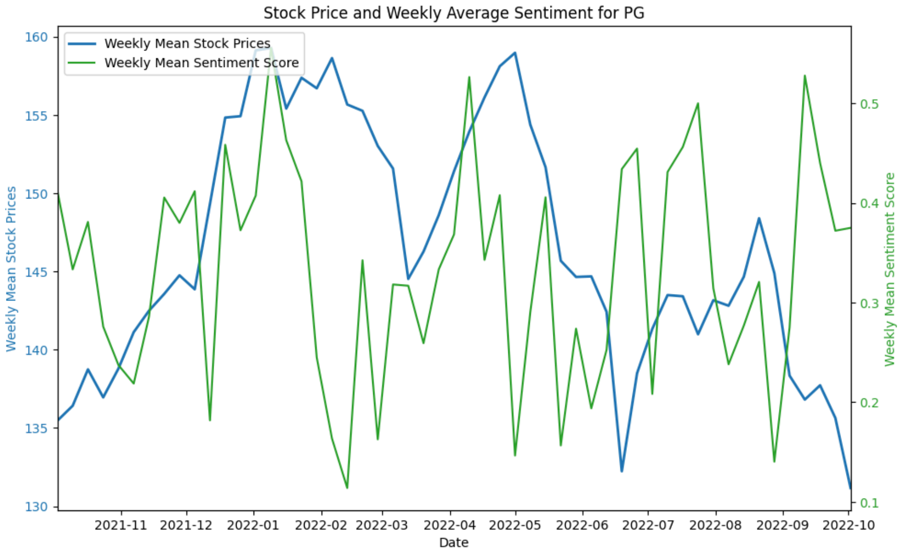


Fig. 20. Twitter trained LSTM predicted weakly mean sentiment score in green and true weakly mean stock prices in blue.

The LSTM model's sentiment predictions here in Fig. 20 are similar with major peaks in the month of 2022-01 and the first half of 2022-04 and three more volatile peaks following that after, just as the SVM predicted as well. There are some differences, such as with the last month 2022-09, where there is a predominantly greater mean sentimental score resulting in a greater difference between the stock price and the investor sentiment. This could affect the predictive models significantly if they are to operate in this month, which has to be taken into consideration beforehand. Overall, the model's predictions appear to be closely related to the SVM's predictions however, major troughs that could have been outlier weeks were softened in comparison and had less of an impact as seen in early 2021-12.

BERT sentiment predictions:

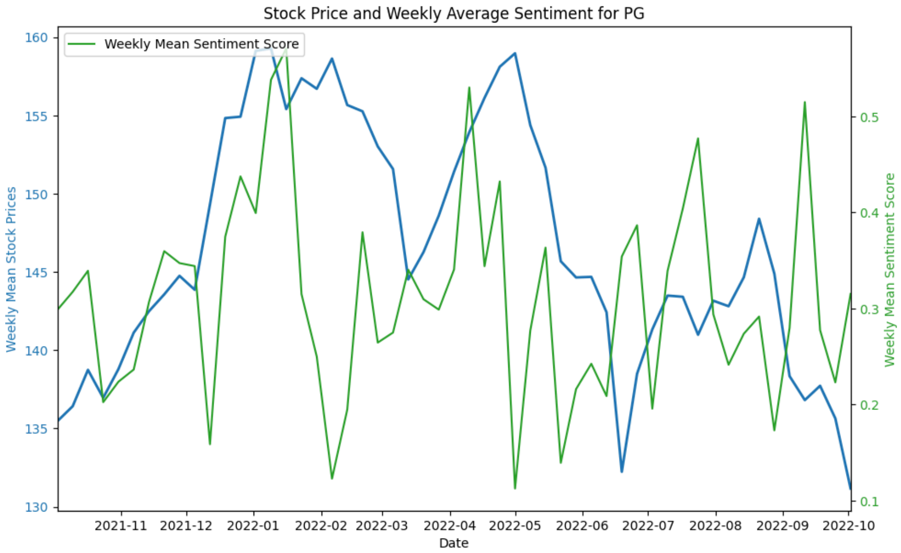


Fig. 21. Twitter trained BERT predicted weakly mean sentiment score in green and true weakly mean stock prices in blue.

Here, the BERT model predictions shown in Fig. 21 are still quite comparable with the previous two, with the two major sentimental peaks in 2021-01 and 2022-04, followed by three more volatile peaks. But a key change is in the last month with scores much lower than the other two models had predicted, which correlates more with the stock pricing at that time. This may be a correlation or a coincidence, but the overarching differences this model predicts are minimal in relation to the previous two, following the same structure with some disparities.

Visualisation of all sentiment prediction as comparison for the PG market:

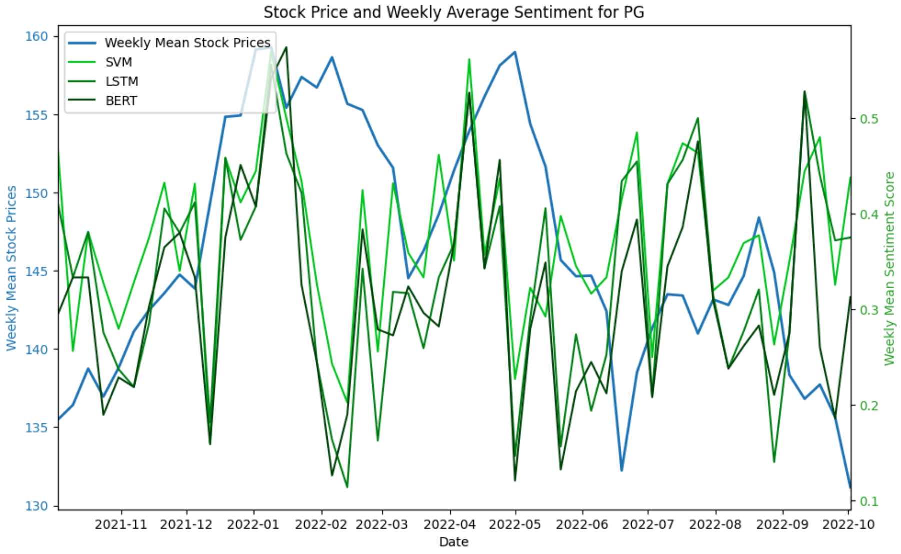


Fig. 22. Combined weakly mean sentiment scores across all models trained on the Twitter data.

Finally, a plot was made to compare the different structures of the predicted sentiment scores for the PG market over the dataset's time frame shown in Fig. 22. There seems to be a common understanding of the sentiment scores with all the major peaks and troughs of the sentiment scores being unanimous amongst the model predictions. With some exceptions with the SVM model performing most deviant predictions, especially during the month of 2022-03 and late 2022-05 and early 2022-06.

These unanimous predictions may be due to the fact that the models were trained on the same dataset, which may have been a limited dataset for vocabulary understanding. If the models were trained on a variety of training dataset. These predictions would have changed somewhat as the models would have gained a better understanding of human text sentiment.

But overall, the somewhat unanimous predictions can also indicate that the models performed as expected, as if there were massive deviations in predictions from each model, this could have led to the idea of errors or mistakes in the model's code. The fact that they are performing as expected, is a factor to be used as evidence to prove that the models worked well within the limitations of the training dataset.

4.3 SVR stock price predictions from models train on Twitter data:

Here, the model's sentimental predictions scores were then used to predict the latter month's stock prices in the dataframe based solely from the correlation between the sentiment scores and the stock prices. Mind you, only using investor sentiments to do such predictions as stated before is only a small cog in the overall predictive process of stock price shifts. This is because other factors such as: Earnings Reports, News and Events, Macroeconomic Indicators, The Competitive Landscape, Regulatory Changes, Technological Advances, Management Changes, Market Trends, Global Events among others have pivotal effects on the markets.

The usage of only Market sentiment will not return accurate results whatsoever, but the effects of which should still not be understated. This task aims to show, based solely on investor sentiment, what the trends of the stock price in the market should resemble if such correlations exist. This information could be used in the overall analysis of a market and appended to studies and research from other fields that affect the markets such as the factors stated before.

To this end, a common support vector regression (SVR) model is trained for all three models whose regression is based on the relationship between the two datasets, the weekly mean stock price and the weakly mean sentiment scores. This is the reason why the stock price and sentiment scores were calculated by the weekly median rather than raw data, as to aid this model in its regression prediction tasks, providing an equal number of segments to calculate from for each set.

To start, an 'End data' was implemented, which was chosen to be located at the beginning of the last month of the dataset, annotated by the red dashed line. This will serve as the end of the SVR's training set, or in other words, all the data on the left side of the red line acts as the training dataset and everything on the right acts as the validation for the results. The SVR then starts its plots where the stock price was at that instance (i.e. where the end date line intersects the stock price line) and plots its predictions with the knowledge of the sentiment of that week's sentiment datapoint to aid it. This will result in four weeks worth of predictions being plotted after the end date line based on the relationship of the data sets before the end date line.

SVM model for predicting last month of stock price using SVR based on the correlation:

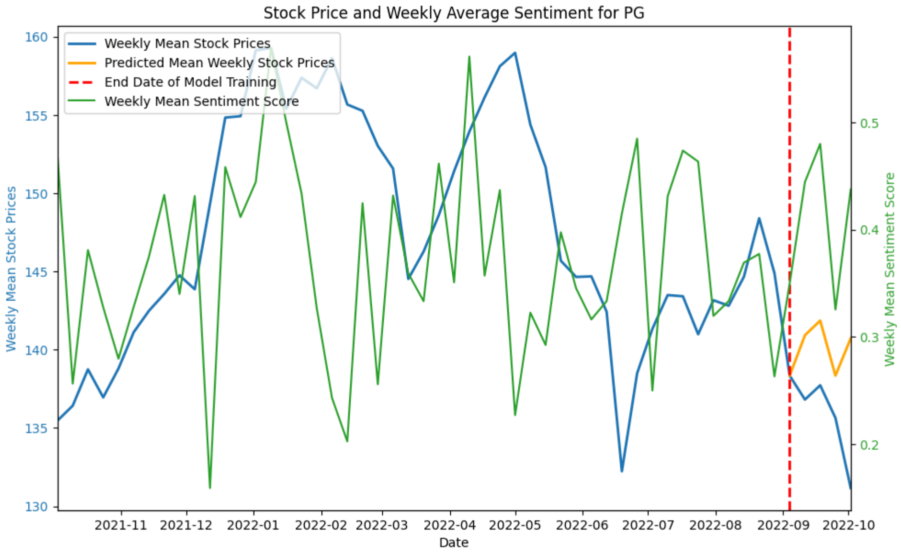


Fig. 23. SVR's predicted stock prices from the Twitter trained SVM's predicted weakly mean sentiment in green and true stock prices in blue.

As seen here in Fig. 23, the predictions the SVR derived from the sentimental analysis provided by the SVM model predicts stagnation from the end date, when in reality the true stock price drops to an all time low in the dataset's timeframe. This is due to the predicted sentiment scores rising and falling and rising again in a stagnant but volatile manner with one week of predicted mean stock price drop.

LSTM model for predicting last month of stock price using SVR based on the correlation:

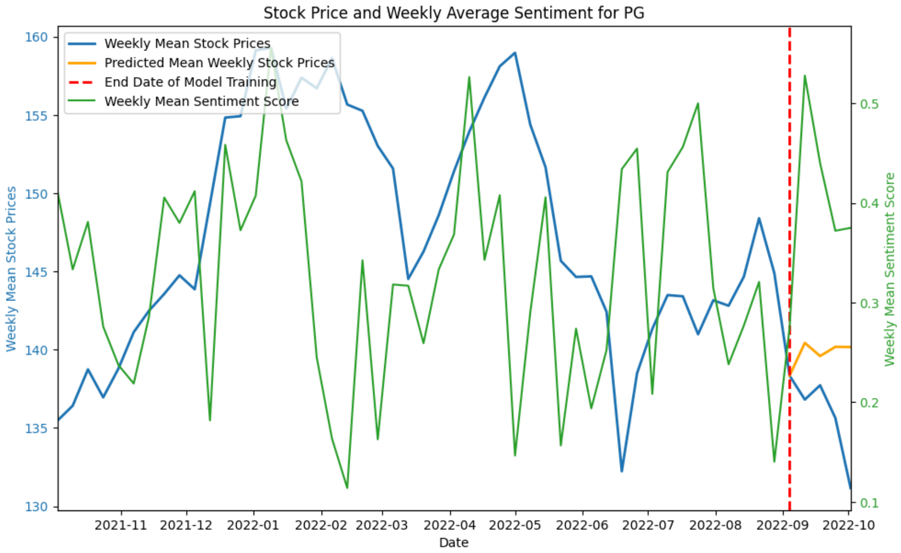


Fig. 24. SVR's predicted stock prices from the Twitter trained LSTM's predicted weakly mean sentiment in green and true stock prices in blue.

In Fig. 24, the predictions the SVR gathered from the sentimental analysis provided by the LSTM model also predicts stagnation from the end date, when in reality the true stock price drops to $130 in the dataset's timeframe. The predicted sentiment scores rose drastically and fell for the latter half of the month, but overall not that volatile of a prediction in relation to the SVM's predictions with the stock prices staying the same price throughout the month.

BERT model for predicting last month of stock price using SVR based on the correlation:

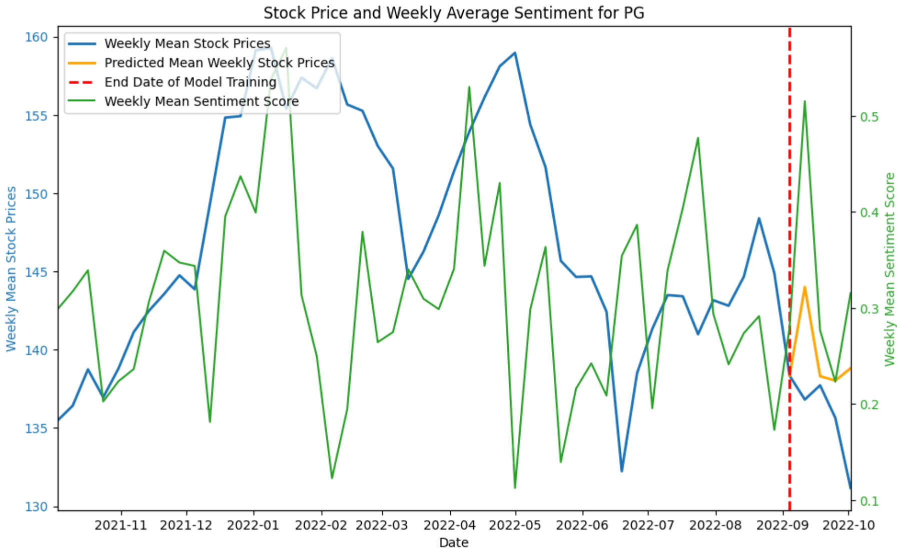


Fig. 23. SVR's predicted stock prices from the Twitter trained BERT's predicted weakly mean sentiment in green and true stock prices in blue.

And finally in Fig. 23, the SVR's stock price predictions derived from the BERT model's sentiment predicted scores showed a strong spike in the first week followed by that latter three weeks of stagnant stock price scores floating around $138. This results in the overall stagnant growth for the final month for the PG market as well, even though the true data shows a more drastic decline in stock prices.

In conclusion, while utilising sentiment predictions derived from the models built before, the SVM, LSTM, and BERT to forecast stock prices provided valuable insights. It is vital to acknowledge the limitations of relying solely on investor sentiment for such predictions. The analysis reveals that although these sentiment-based models offer some predictive abilities, they do not fully encapsulate the complexities of the market dynamics. Other crucial factors such as earnings reports, news events, macroeconomic indicators, and regulatory changes significantly influence stock price movements. Therefore, while sentiment analysis contributes to understanding market trends, it should be complemented with more verified methods from various fields that impact market behaviour to develop a more panoramic view of the stock price trends.

Additionally, upon the examination of SVM, LSTM, and BERT models, the results show variations in their predictions, underlying the importance of considering multiple sources of analysis for a robust decision making and wider view and comprehension in such financial markets. Ultimately, integrating sentiment analysis into broader market research enhances the understanding of market sentiment, but it should be utilised cautiously alongside other fundamental and technical analyses for more accurate predictions and informed investment strategies.

4.4 Training on Reddit and Twitter data:

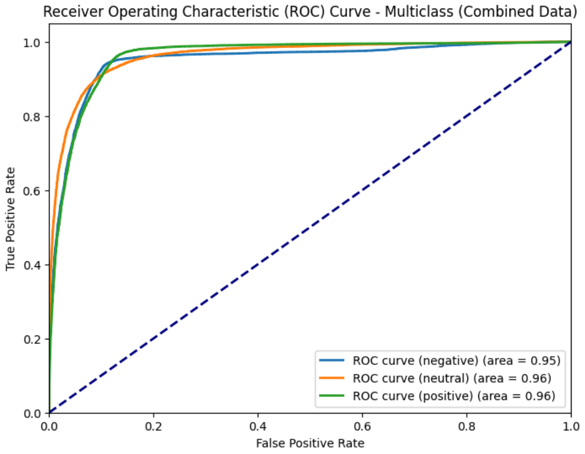
Expanding on the scope of the analysis and to further deepen the model's understanding, a second dataset was implemented into the training process. In addition to the initial body of tweets, a new dimension was introduced through the incorporation of Reddit posts ranging in size and complexity in relation to specific stock market conversation. This was implemented with the aim to leverage the diverse conversational nature of human speech and expand on the content the models understand through the addition of Reddit discussions, complementing the sentiments captured from Twitter.

This inclusion of Reddit posts should broaden the spectrum of the sentiments the models are capable of comprehending, in addition to aiming to limit or even mitigate the possibility of overfitting. This is because of the increased diversity as a different source of data the models are exposed to will widen the range of linguistic styles, topics and sentiments. This diversity should prevent the models from memorising the patterns from specific to the original dataset, reducing the likelihood of overfitting. This also encourages the models to learn more generalised patterns and features that are applied across different contexts. This broader understanding enables the models to generalise better to unseen data.

By amalgamating the Twitter and Reddit posts for the model training process (200k posts in total), a more comprehensive understanding of sentiment dynamics should be captured and even new nuances that could have been missed if trained on a single set of data. This would enhance the depth of the sentiment analysis but also provide a robustness and versatility to the models in classification sentiment from patterns across different platforms and online communities. Additionally, some VRAM issues occurred due to the size of the reddit posts as they were mostly paragraphs, so they were partitioned into smaller lengths beforehand.

This section will follow the same format as previous to compare and contrast the differences between the cross-validation results, sentiment prediction results and SVR predicted stock price results. To identify whether this addition of a secondary training dataset has truly positively impacted model's abilities to capture and classify sentiments in relation to stock price movements online.

SVM Model Training Results from the combined dataset:



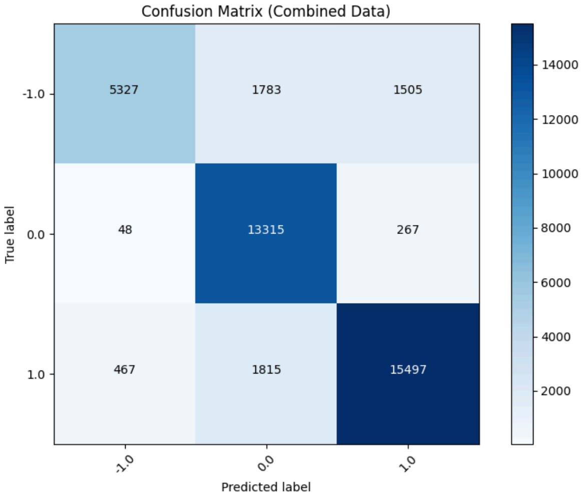


Fig. 24. ROC curve of SVM classifications, trained on both Reddit and Twitter data.

Fig. 25. Confusion matrix of SVM classifications, trained on both Reddit and Twitter data.

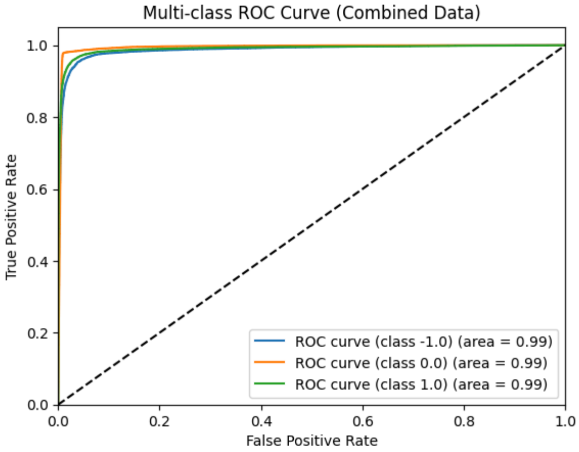
The Support vector machine model trained on the combined datasets, showed a fair performance of 85%. The model demonstrated promising precision scores of 0.91, 0.79 and 0.90 for negative, neutral and positive sentiments respectively. Additionally, the recall scores were admirable, particularly with neutral sentiments, with scores of 0.62, 0.98 and 0.87 for negative, neutral and positive sentiments respectively.

The multiclass ROC curve shown in Fig. 24, further demonstrated the model's fair and equivalent efficacy to previously being trained on the single dataset. Showing robust discriminatory abilities across all sentiment classes. It resulted in a high Area Under the Curve (AUC) values of 0.95, 0.96 and 0.96 for negative neutral and positive sentiments respectively, this demonstrates how fairly strong the model's predictive capabilities are in effectively distinguishing between different sentiment classes with minimal false positive rates and high true positive rates.

The analysis of the confusion matrix in Fig. 25 revealed that this provided additional insights into the model's performance and how it showcased a strong correct classification ability for negative and positive sentiments, as evidenced by the high counts on the diagonal for these classes. However, similar to the results from the single dataset training, it faced challenges in accurately classifying neutral sentiments. This can be reflected by the off-diagonal counts, showing a tendency to misclassify neutral tweet sentiments as either negative or positive more frequently than vise versa.

Despite these challenges it faced, the overall performance of the Support Vector Machine model trained on the combined dataset was admirable. The majority of predicted results ran in parallel with the true sentiment labels in addition to requiring very little time and resources to train with even on CPU, underlying the model's suitability for real-time applications.

LSTM Model Training Results from the combined dataset:



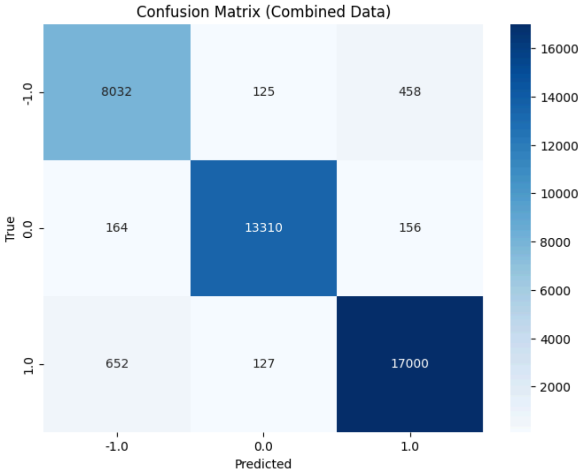


Fig. 26. ROC curve of LSTM classifications, trained on both Reddit and Twitter data.

Fig. 27. Confusion matrix of LSTM classifications, Trained on both Reddit and Twitter data.

The Long Short-Term Memory (LSTM) Model, trained on the combined datasets of Tweets and reddit posts, showcased an exquisite performance, achieving an impressive overall accuracy of 0.96%. It demonstrated fairly remarkable precision scores across the board of classes with values of 0.91, 9.98 and 0.97 for negative, neutral and positive sentiments respectively. Moreover, the recall scores are equally high with values of 0.93, 0.98 and 0.96 for negative neutral and positive sentiments respectively.

The multiclass ROC curve in Fig. 26 showed the LSTM model's outstanding discriminatory capabilities with high AUC cores of 0.99 of all sentiment classes. With such AUC scores, it indicates that the model has a very effective ability to distinguish between different sentiment classes with few false positives and high true positives, further underlining its robust predictive capabilities.

Additionally, from analysing the confusion matrix in Fig. 27 for further validation of the model's performance showed a strong ability to correctly classify Tweets and Reddit posts into their respective sentiment classifications. Most notably, the LSTM gave impressive precision in identifying neutral sentiments with the majority of them correctly classified, a stark contrast to the SVM. Overall, the confusion matrix spotlights the models accuracy in predicting sentiment labels across the combined dataset.

Furthermore, evaluation of the training process in Fig. 28 and Fig. 29 showed some insights that identified that the model's accuracy curve reached an accuracy of 0.96 at the second epoch and growth stagnated as it reached the fifth and last epoch. Together, the loss over epochs graph showcased a steady progressive decline and falling to near 0.14 by the fifth epoch until further results after more epochs would only return diminishing results not worth the computational cost and time they are worth, as seen here:

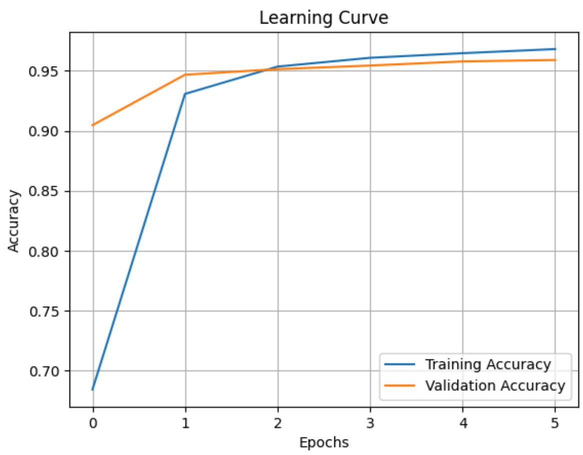


Fig 28. Learning curve over the LSTM's epochs, trained on both Reddit and Twitter data..

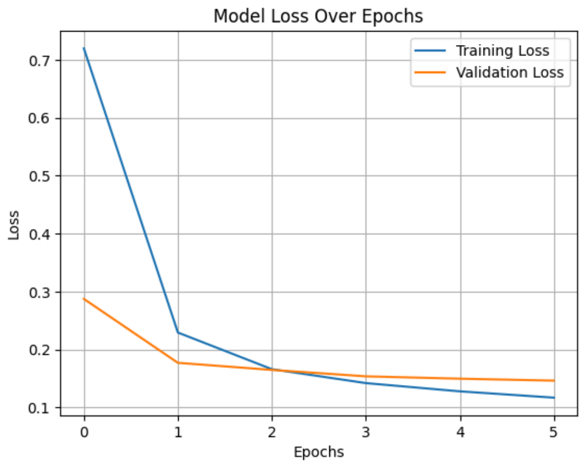
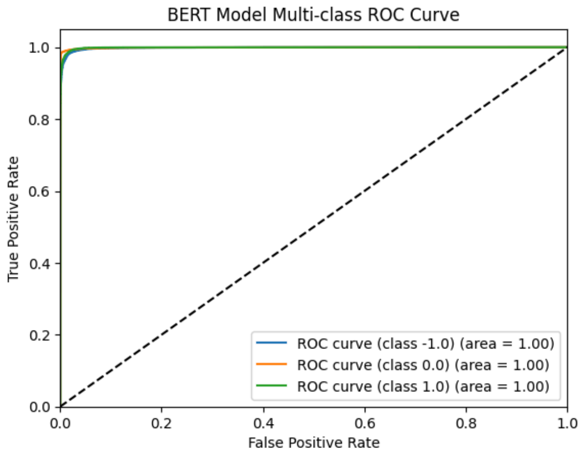


Fig. 29. Validation loss curve over the LSTM's epochs, trained on both Reddit and Twitter data.

BERT Model Training Results from the combined dataset:



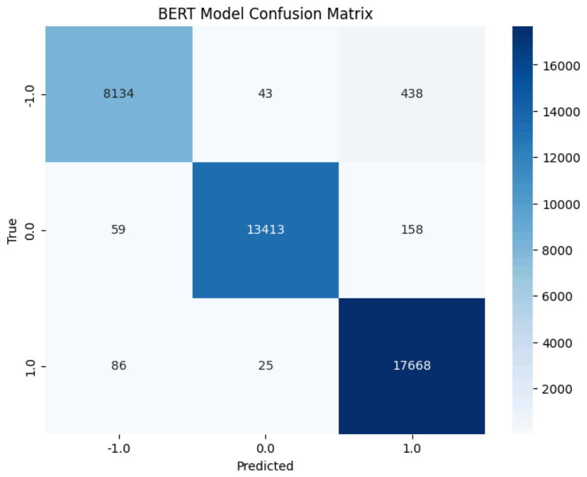


Fig. 30. ROC curve of BERT classifications, trained on both Reddit and Twitter data.

Fig. 31. Confusion matrix of LSTM classifications, Trained on both Reddit and Twitter data.

The Bidirectional Encoder Representation of Transformers (BERT) Model trained on the combined datasets showed remarkable performance with the ROC curve in Fig. 30 showcased a consistently perfect Area Under the Curve (AUC) scores of 1s across all sentiment classes. This identifies that the BERT model is near perfect when it comes to its discriminatory abilities with distinguishing between positive, neutral and negative sentiments.

Additionally, the analysis of the confusion matrix in Fig. 31 showed further validation of the model's exquisite performance, with near perfect predictions across all sentiment classes, with high accuracy rates of negative (98.9%), neutral (98.2%), and positive (97.7%) sentiments. This highlights the model's robustness and accuracy in classifying sentiment across the diverse textual datasets but also out performing the previous SVM and even LSTM models cross validated earlier.

During the training process, the BERT model showed consistent improvement in accuracy over the epochs shown in Fig. 32 but peaking at the third with a score of near 0.98. More noticeable was the drop in accuracy that followed that at the fourth and final epoch. This indicates that the model seemed to have reached its potential and possibly stabilisation or even reached a point of convergence for the model’s performance. This further suggested the possibility of no beneficial returns through further epochs.

Moreover, evaluation of the training accuracy and loss throughout the epochs highlighted the model's efficiency and effectiveness during the learning process shown in Fig. 33. The average training loss showed a steady decline throughout the training epochs, indicating effective convergence towards minimising loss and optimising performance. This is in stark contrast to the LSTM model, where the BERT reached an accuracy of 96% as early as the fifth epoch with the LSTM reaching only a menial 68%. Showcasing the effectiveness of utilising pretrained language representations and capturing complex contextual information, leading to a rapid convergence and superior performance event at the initial stages of training:

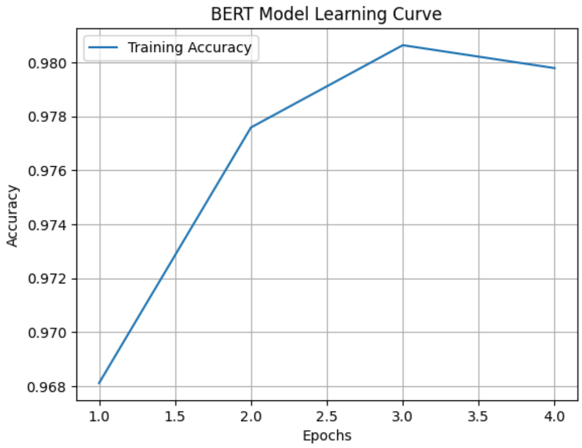


Fig 32. Learning curve over the BERT’s epochs, trained on both Reddit and Twitter data..

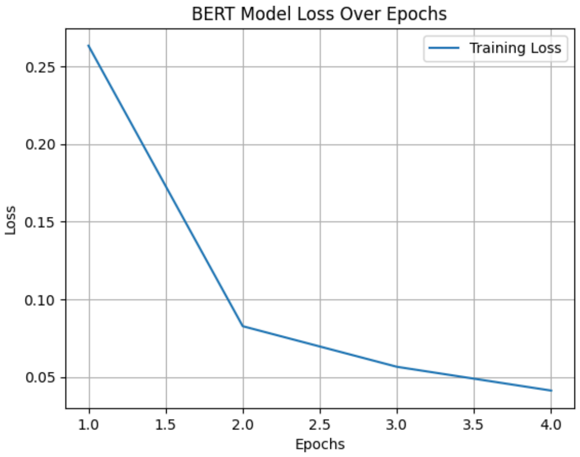


Fig. 33. Validation loss curve over the BERT’s epochs, trained on both Reddit and Twitter data.

4.5 Analysis of sentiment predictions of models train on Twitter and Reddit data:

We will skip over the entire process of evaluating each individual model's sentiment predictions for the PG market over the timeframe of the dataset. This is because the common features (peaks and troughs in sentiment scores) are very similar to the results gathered from the models trained on the single dataset:

The combined Reddit and Twitter datasets sentiment predictions of each model:

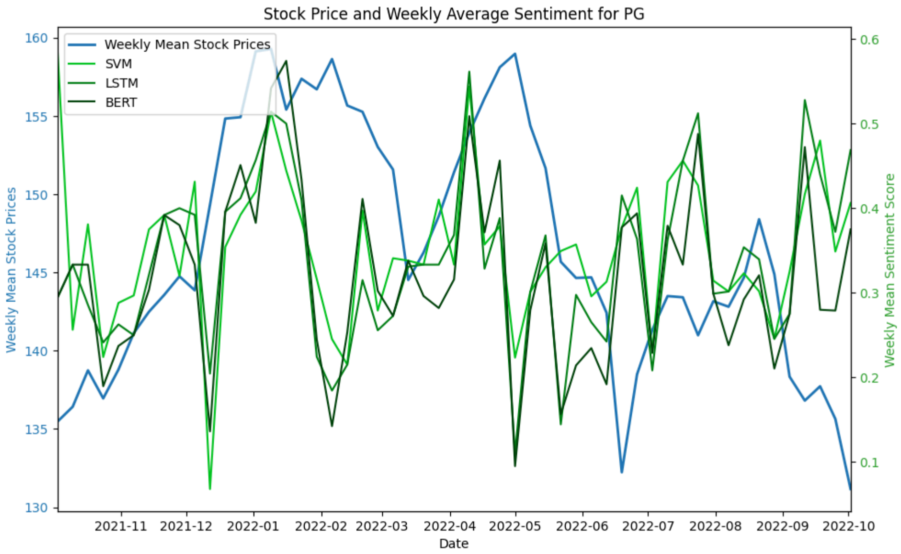


Fig. 34. Combined weakly mean sentiment scores across all models trained on both Reddit and Twitter data.

The previous Single Twitter dataset sentiment predictions of each model for comparison:

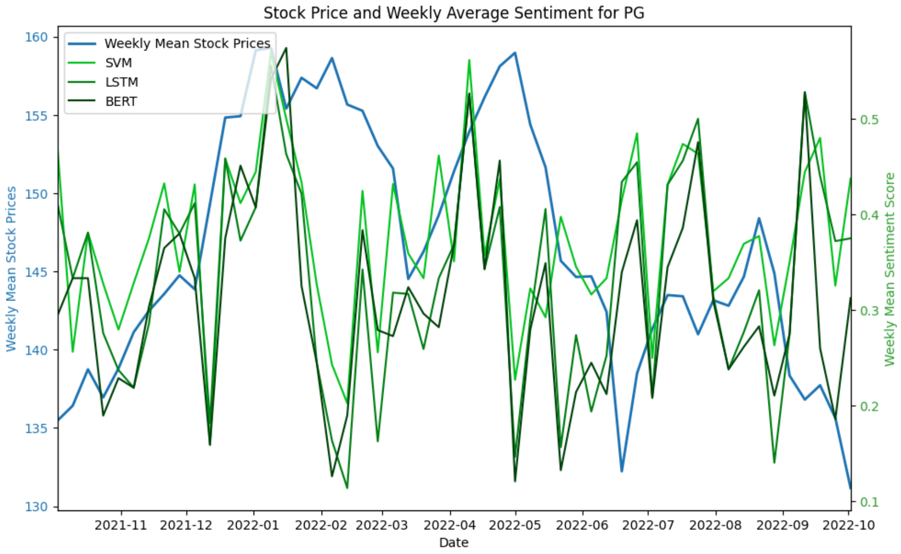


Fig. 35. Combined weakly mean sentiment scores across all models trained on Twitter data. Same as Fig. 22

As can be observed in Fig. 34, the results from the models trained on the combined dataset appear to be very similar to the previous single dataset models but overall appear to be much more unanimous with their predicted sentimen scores with the SVM, LSTM, and BERT models. With the previous single training dataset, only the LSTM and BERT models appeared to have somewhat unanimous decisions on the sentiment classification of the

Tweets in relation to the PG stock market. Here with the combined dataset training dataset, the SVM classification abilities appear to be overall much more unanimous with its competitors, with few exceptions such as the latter half of 2022-05 and the first week of 2022-06.

This phenomenon seems to underline the effectiveness of incorporating and utilising a diverse range of training data from different online communities. By doing so, it appears that the models gain a more nuanced understanding, leading to a convergence in their predictions. This convergence not only strengthens the confidence in the individual model predictions but also adds credibility to the collective sentiment analysis as a whole. They serve as compelling evidence that leveraging the broader spectrum of training data can significantly enhance model performance and enhance the understanding into such sentiment classification dynamics.

4.6 SVR stock price predictions from models trained on Twitter and Reddit data.

Combined dataset SVM model for predicting last month of stock price using SVR based on the correlation:

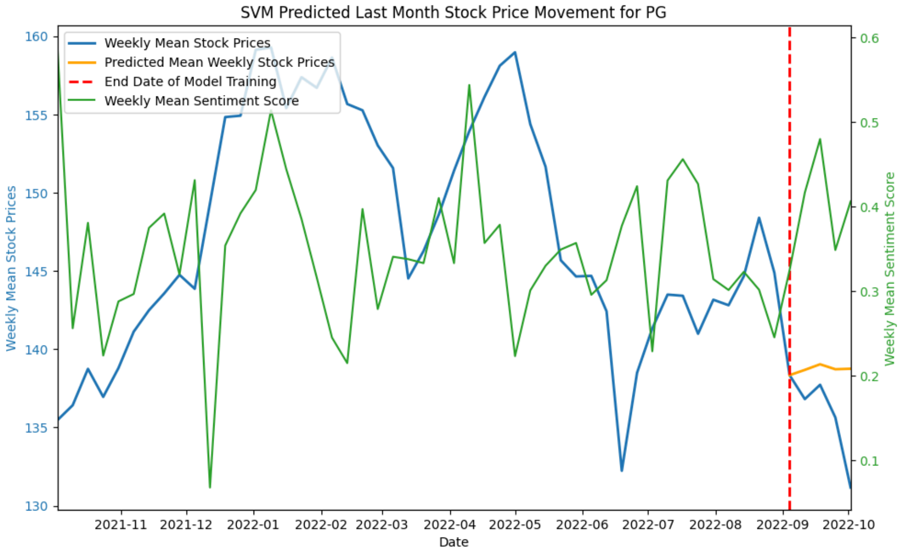


Fig. 36. SVR's predicted stock prices from the Reddit and Twitter trained SVM's predicted weakly mean sentiment in green and true stock prices in blue.

Here in Fig. 36, the SVR's predicted stock price for the last month of the time frame generated by the SVM model appears to be extremely stagnant, which is far from the true data that shows the drop of true stock prices at that time. This prediction of no change in stock price is not even similarly present throughout the whole time frame which shows a completely unexpected change in the character of this market's stock price. But still predicted a more stagnant month in relation to its pair SVM sentiment prediction model trained on one dataset, which is quite peculiar.

Combined dataset LSTM model for predicting last month of stock price using SVR based on the correlation:

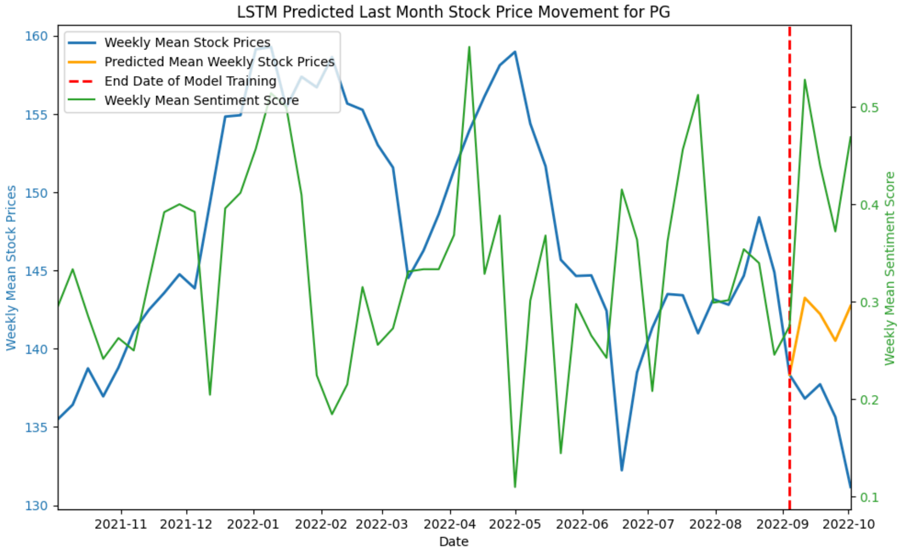


Fig. 37. SVR's predicted stock prices from the Reddit and Twitter trained LSTM's predicted weakly mean sentiment in green and true stock prices in blue.

Here in Fig .37, the SVR's predictions for the last month of stock prices based on the LSTM's sentiment predictions showed a gradual growth. This is a much more optimistic forecast in relation to the SVR's predictions produced from the predicted sentimental scores from the LSTM trained on the single dataset, which showed a strong stagnant forecast. This is also wrong as seen by the true stock price data that drops below it, which indicates a possible mismatch in the relationship between the investor sentiment on twitter and the true stock price.

Combined dataset BERT model for predicting last month of stock price using SVR based on the correlation:

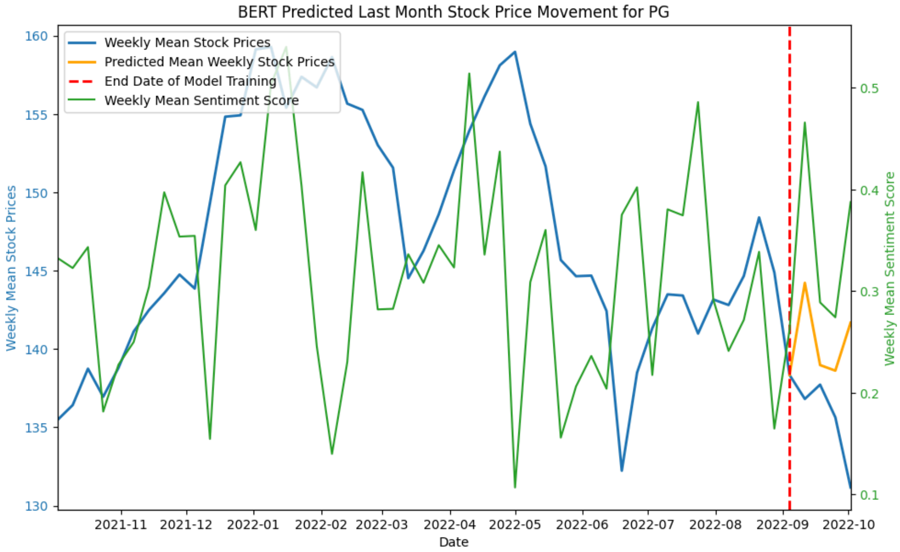


Fig. 38. SVR's predicted stock prices from the Reddit and Twitter trained BERT's predicted weakly mean sentiment in green and true stock prices in blue.

Now finally, In Fig. 38 the SVR's forecasted stock prices for the BERT's sentimental scores also shows an overall increase, but with a much more drastic increase for the first and last week in that month. This is much more in parallel with the sentiment scores as it seems to mimic them fairly equivalently, which is in stark contrast to the SVR's stock price predictions derived from sentimental scores the BERT model that was trained on the single dataset. Where the previous model predicted an increase similar to this one for the first week but then dropped and stabilised at a value at the beginning of the month, showing no growth, where this model showed some growth at the last week.

In conclusion, the analysis of the combined dataset SVM, LSTM, and BERT models for predicting the last month of stock prices using SVR revealed notable disparities with the predicted and actual outcomes, just as before. The SVM's models predictions depict a remarkably stagnant stock price, which diverges significantly from the true data, indicating a drop during that period. Conversely, the LSTM model's predictions forecasted a gradual growth in stock prices, which reflects a more optimistic outlook compared to the stagnant forecast derived from the model trained on the single dataset. However, this optimistic projection still fails to align with the subsequent decline in true stock prices, which raises questions about the relationship between the investor sentiment on Twitter and the stock prices of a market. Similarly, the BERT model's predictions showed an overall increase in stock prices which mirrored the sentiment scores more closely, especially during the first and last weeks of the forecasted month. This is contrasted with the single dataset model;s predictions which showed limited growth potential.

Overall, this analysis underlines the complexity of forecasting our stock prices based solely on sentiment gathered from Twitter in relation to that market, which shows the importance of considering diverse data sources listed before. While the combined dataset models showed promise in capturing sentiment dynamics more accurately, further refinement and research is needed to improve this accuracy of stock price predictions. This could be done by testing other predictive models rather than the SVR which might be limited to stark correlation predictions, for future research and development of this project.

4.7 Extended Research: External factors study using the BERT model.

Using the BERT model as the best performing model against this dataset, further research was carried out to study the effects of external factors like political decisions, pandemics and wars against these investor statements. This was done because, if factors such as these occurred during the time frame of the data set (2021 to 2022), they could skew the overall sentiment scores used in the analysis and predictive models as shown before.

To analyse the effects of such topics on the investor sentiments, a list of keywords were collected as a hard coded variable in relation to each topic such as ‘Ukraine’ for war related topics and ‘Covid’ for pandemic related topics etc. Obviously the quality and accuracy of the analysis heavily relies on these keywords as missing keywords could lead to uncaptured points of interest, but some effort was put into this list and kept it in reference to the time frame of the dataset. I.e. Not including topics like invasion of Crimea and Chechnya as these are relatively historic topics and should not be mentioned nor affect the investor sentiments, and results shown here in Fig. 39.

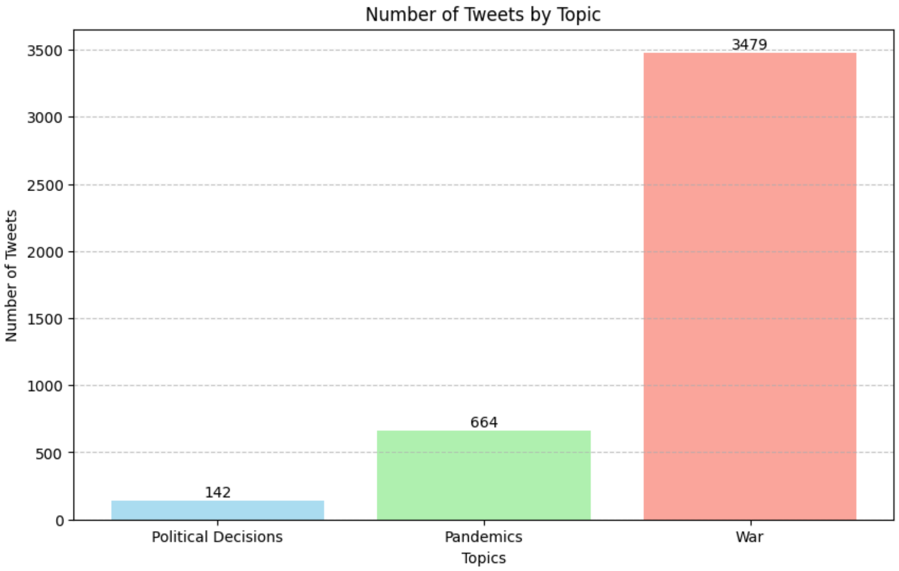


Fig. 39. Quantity distribution of filtered tweets that include keywords relating to the external factors.

From this there seems to be a large amount of tweets in the dataset that are in reference to wars and conflict at roughly 3,500 tweets with only half a thousand tweets in relation to pandemics, or a mere hundred or so tweets in relation to some political decisions. This indicates that out of the roughly 80,700 tweets, war related tweets seems to be a cause of main concern and its effects on the sentiment as it accounts for 4.3% of the total tweets in the dataset.

To gather a further analysis from this, the topics are then split into the three sentiment scoring classifications, -1, 0 and 1 for negative, neutral and positive sentiments to analyse the amount of these tweets belonging to these classes. This is to identify if these topics have a generally negative, neutral or positive affect on the investor sentiments.

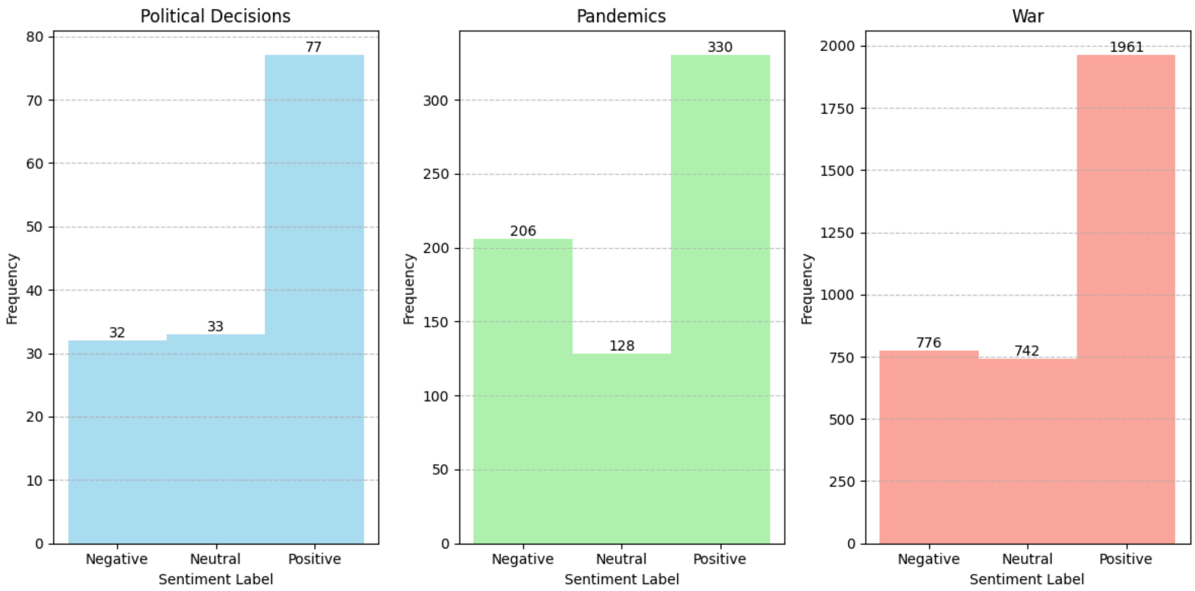


Fig. 40. Quantity distribution of predicted sentiment scores from the Reddit and Twitter trained BERT model for filtered tweets relating to the external factors.

As seen in Fig. 40, using the BERT model for classification of these filtered tweets, most of which were considered positive sentiments, indicates a quite substantial disparity of the effects such topics of discussion have and could disproportionately skew the results gathered before over the time frame. In addition, there seems to be a somewhat relevantly equal ratio of neutral and negative sentiments throughout the topics. But the overall results differ to what was expected, as the impacts of a global pandemic like Covid-19, the war in Ukraine and others were expected to negatively impact the overall sentiments that mention these topics. There are many factors that could lead to such conclusions, such as:

The truth: The fact that these external factors truly have a positive impact on the investor sentiment, this might be due to more bullish effects on the company's stance towards these topics. As seen in the high quantity of positive war related sentiments, companies may have shown a strong stance towards the war in Ukraine which occurred during the time frame of the data captures and as such the public showed keen admiration towards this. But such a factor like this is implied and to completely understand why this has occurred requires further investigation that might lead to reading and understanding each tweet one by one, which is beyond the scope of this study.

Survivorship Bias: Users who were active on social media platforms like twitter during crises like wars, pandemics, or political upheavals may represent a subset of the population that is more resilient or optimistic. Those who are disproportionately affected by negative events may be less likely to engage in online discourse and as such we don't have those sentiments in the dataset.

Coping Mechanism: Social media often serves as a vent for people to express their emotions and seek support during difficult times. In the face of distress, individual users may turn to positive affirmations or expressions of hope as a coping mechanism, which could manifest as positive sentiment in their tweets.

Information Dissemination: During crises, individuals may share positive news, stories of resilience, or messages of solidarity to boost the morale of others and counterbalance the win over the negativity in the news cycle. This dissemination of positive information could contribute to the observed distribution of sentiments in tweets related to these topics.

Selective Attention: While negative events may dominate headlines and kindle strong emotional responses, individuals may actively seek out or amplify positive perspectives to maintain a sense of balance or optimism in their feeds.

Bias in Data Collection: There may be a possibility of bias in the collection of the tweets through the filtering process. The methodology of filtering tweets by keywords is basic and limits the selection criteria for inclusion, and as such the models used for sentiment analysis could inadvertently favour positive content or even overlook certain types of negative sentiments not captured in the filtering process.

Temporal Dynamics: Sentiments expressed on social media platforms like twitter can vary over time, influenced by changing circumstances, news cycles, and public discourse. The timeframe of data collection (2021-2022) may capture specific events or trends that skew the sentiment distribution towards positivity by chance.
[37]

Despite the initial expectations of having a mainly negative sentiment in regards to the external factors due their negative connotations and effects on the world. The analysis revealed a more complex picture as seen in Fig. 40. Several potential explanations for what was a predominantly positive sentiment in relevance to the topics chosen have been identified. These come to the possibility that companies may have taken bullish stances on the issues, which evoked a positive reaction from investors. Additionally, survivorship bias, coping mechanisms, information dissemination, selective attention, bias in data collection, temporal dynamics contributed to the final shape of the sentimental data collected in relevance to the external factors that could pose an issue to the overall sentiment analysis conducted previously.

5. Conclusion:

Discussion:

The original aim of the research project was to leverage these sentiment analysis models, using data from social media platforms like Twitter and later Reddit, to enhance the accuracy of predicting stock price movements. The project highlighted the significant advancements in sentiment analysis, especially with the utilisation of diverse datasets and the utilisation of SVM, LSTM, and BERT models. Although the combined datasets led to more robust sentiment analysis, the efficacy of using public investor sentiment from platforms like Twitter alone for stock price predictions showed limitations. This suggests that a shift from the initial hypothesis was needed, and thus additional factors beyond social media sentiment, such as earnings reports and macroeconomic indicators, are necessary for more accurate predictions.

In regards to meeting the stated objectives, the project had successfully compared the performance of SVM, LSTM, and BERT models in sentiment analysis and stock price forecasting. Evaluation metrics such as classification reports, ROC curves, confusion matrices and epoch accuracy and loss graphing were enrolled, providing insights into each models' precision and accuracy. Furthermore, the project dived into the impact of external factors like wars, political decisions and pandemics on investor sentiment on the Twitter data and the reasoning for such results, aligning with the specified objectives. However, the extent to which these external factors were incorporated into the predictive models and their influence on sentiment analysis could have been developed further for a comprehensive understanding.

In the end, this research project made significant attempts into exploring the relationship between the predicted sentiment and stock price movements. But some changes to the project were carried out, such as having to partition the Reddit training sentiment datasets to reduce RAM usage which was unexpected. This highlights the complexity of training models for predicting sentiments but also stock market behaviour overall, solely based on social media sentiment where limitations were evident, which emphasised the importance of utilising multiple data sources and factors for more accurate forecasts. Further research is needed to address these limitations and enhance the predictive capabilities of sentiment analysis models to complement usage of an SVR for stock price prediction.

Summary of results:

Training, evaluation and application of the machine learning models for sentiment analysis against Twitter data using three prominent models, the Support Vector Machine (SVM), Long Short-Term Memory (LSTM) and Bidirectional encoder representations of Transformers (BERT). These were trained and assessed for their accuracy in classifying sentiments into negative, neutral and positive classes. Whilst the SVM model showed respectable performance, it ultimately struggled to accurately categorise neutral sentiments. On the other hand, the LSTM model displayed its superior accuracy, especially with classifying neutral sentiments, this highlights its potential for more nuanced sentiment analysis overall. However, the BERT model surpassed all as the standout performer, this was evident by the near perfect accuracy across all sentiment classes, which bolden's its robust discriminatory capabilities for natural language processing tasks such as this.

Furthermore, the addition of the Reddit data alongside the previous Twitter data into the training process displayed more promising results, leading to more consistent and unanimous sentiment predictions across all models. This combination of diverse datasets from different online communities contributed to a more comprehensive understanding of sentiment dynamics and as such, enhancing each model's performance relatively. But unfortunately, to dispute the models' proficiency in sentiment analysis, the stock price forecasting section of the research, based solely on the relationship between the historic stock price and sentiment time series data remained challenging. There were consistent discrepancies between the predicted and the actual stock prices for the latter month of the time series dataset which highlights the complexities of such a task but also was expected from the beginning. This was discussed previously why such a task could reach inaccurate results. But it was still carried out to aid in the overall goal for a potential firm or company to append such research and models towards a more developed study that incorporates more factors of financial markets that go beyond the scope of this project.

The study also identified the importance of critically analysing the models predictions to locate strengths, weaknesses and areas for improvement with such technologies already in use in the industry. While machine learning models provide more valuable insights into sentiment trends online, the drawbacks to predicting stock prices highlights the need for a more varied and deeper approach. Using diverse data sources for training and delving into other alternatives to predictive modelling beyond the selected Support Vector Regression (SVR) would provide more promise for more accurate and reliable predictions in financial marketing. Conclusively, while sentiment analysis has promise for understanding public investor sentiments online, its application into broader market research required very careful planning of combining various factors and strategies to gather more meaningful insights to give informed decisions to potential investors.

Areas of improvement for the future:

To develop the current findings from this project and work towards weaknesses found throughout, a few features could be implemented for future work. Firstly, appending additional data sources beyond Twitter and Reddit, such as financial news websites and other social media platforms like StockTwits, could provide a more deepening and diverse understanding of investor sentiment in relation to financial markets. Secondly, experimentation with other advanced NLP techniques like contextual embeddings and transformer-based models like Generative PreTrained Transformers (GPTs) similar to the BERT model could enhance the understanding of linguistic nuances. Thirdly, delving into newer and more varied modelling techniques and using features beyond sentiment analysis, such as market indicators and company specific events, that could improve predictive accuracy.

Furthermore, experimentation with other applications of deep reinforcement learning for developing trading strategies and implementing intelligible AI techniques for transparent models [25] show potential areas for improvement. Additionally, ‘cross-domain transfer learning’ [24] from similar domains and further research into model interpretability could add to a more robust sentiment analysis for the financial market. Overall, these future experimentations and research could enhance our sentiment analysis accuracy and provide a more informed decision for a client willing to invest in a sector or market, of which they want evidence of high degree potential of gainingfull investment.

6. References:

[1] Palomo, C. (2022). 'Tweet Sentiment Analysis to Predict Stock Market'. Stanford CS224N Custom Project. Available at: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-170049613.pdf.

[2] Darapaneni, N., Reddy Paduri, A., Sharma, H., Manjrekar, M., Hindlekar, N., Bhagat, P., Aiyer, U. and Agarwal, Y. (2021). 'Stock Price Prediction using Sentiment Analysis and Deep Learning for Indian Markets'. Available at: https://arxiv.org/ftp/arxiv/papers/2204/2204.05783.pdf.

[3] Kalyani, J. Bharathi, H. and Jyothi, R. (2016). 'Stock Trend Prediction Using News Sentiment Analysis'. International Journal of Computer Science and Information Technology, 8(3), pp. 67–76. Available at: https://doi.org/10.5121/ijcsit.2016.8306.

[4] Baker, M.P. and Wurgler, J.A. (2003). 'Investor Sentiment and the Cross-section of Stock Returns'. SSRN Electronic Journal, 61(4). Available at: https://doi.org/10.2139/ssrn.464843.

[5] Kolasani, S.V. and Assaf, R. (2020). 'Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks'. Journal of Data Analysis and Information Processing, 08(04), pp. 309–319. Available at: https://doi.org/10.4236/jdaip.2020.84018.

[6] Sousa, M.G., Sakiyama, K., Rodrigues, L. de S., Moraes, P.H., Fernandes, E.R. and Matsubara, E.T. (2019). 'BERT for Stock Market Sentiment Analysis'. Available at: https://doi.org/10.1109/ICTAI.2019.00231.

[7] Goularas, D. and Kamis, S. (2019). 'Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data'. Available at: https://doi.org/10.1109/Deep-ML.2019.00011.

[8] Yogesh, C. and Antoreep, J. (2020). 'Sentiment Analysis using Machine Learning and Deep Learning'. IEEE Conference Publication. Available at: https://ieeexplore.ieee.org/document/9083703.

[9] Saini, A. (2021). 'Support Vector Machine(SVM): A Complete guide for beginners'. Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

[10] Abdullah, R. (2023). 'LSTM Networks'. LinkedIn. Available at: https://www.linkedin.com/pulse/lstm-networks-abdullah-al-rahman.

[11] paperswithcode.com. (n.d.). Papers with Code - BERT Explained. Available at: https://paperswithcode.com/method/bert.

[12] Blanco, V., Japón, A. and Puerto, J. (2022). 'A mathematical programming approach to SVM-based classification with label noise'. Computers & Industrial Engineering. Available at: https://doi.org/10.1016/j.cie.2022.108611.

[13] MLMath.io (2019). 'Mathematics behind Support Vector Machine(SVM)'. Medium. Available at: https://ankitnitjsr13.medium.com/math-behind-support-vector-machine-svm-5e7376d0ee4d.

[14] Andrew, P. and Siwei, L. (2017.). 'LSTM with working memory'. Available at: https://ieeexplore.ieee.org/abstract/document/7965940.

[15] Yu, W. (2017). 'A new concept using LSTM Neural Networks for dynamic system identification'. In: 2017 American Control Conference (ACC), Seattle, WA, USA, pp. 5324-5329. Available at: https://ieeexplore.ieee.org/abstract/document/7963782.

[16] Saxena, S. (2021). 'LSTM | Introduction to LSTM | Long Short Term Memory'. Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/.

[17] Tan, Y., Jiang, L., Chen, P. and Tong, C. (2023). 'DQMix-BERT: Distillation-aware Quantization with Mixed Precision for BERT Compression'. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), Honolulu, Oahu, HI, USA, pp. 311-316. Available at: https://ieeexplore.ieee.org/document/10394642.

[18] Qi, W., Guo, X. and Du, H. (2023). 'LMIE-BERT: A Learnable Method for Inter-Layer Ensembles to Accelerate Inference of BERT-Style Pre-trained Models'. In: 9th International Conference on Big Data Computing and Communications (BigCom), Hainan, China, pp. 271-277. Available at: https://ieeexplore.ieee.org/document/10415602.

[19] Chauhan, N.S. (2022). 'Google BERT: Understanding the Architecture'. The AI dream. Available at: https://www.theaidream.com/post/google-bert-understanding-the-architecture#:~:text=BERT%20Base%3A%2012%20layers%20.

[20] Raimundo, M.S. and Okamoto, J. (2018). 'SVR-wavelet adaptive model for forecasting financial time series'. Available at: https://doi.org/10.1109/INFOCT.2018.8356851.

[21] Huang, D. (2022). 'SVR Modeling and Parameter Optimization for Financial Time Series Forecasting'. In: IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, pp. 1126-1130. Available at: https://ieeexplore.ieee.org/document/10016054.

[22] Tian, Y. "Stock forecasting method based on wavelet analysis and ARIMA-SVR model," (2017) 3rd International Conference on Information Management (ICIM), Chengdu, China, pp. 102-106, Available at: https://ieeexplore.ieee.org/document/7950355.

[23] Musa, A. (2022). 'A soft computing technique for predicting flexural strength of concrete containing nano-silica and calcium carbide residue'. Available at: https://www.researchgate.net/figure/Structure-of-the-SVR-model_fig3_361657233.

[24] Zang, S., Zhang, P., Guo, L., Ma, J., Chang, Y. and Ma, C. (2022). 'Transfer Extreme Learning Machine with Cross Domain Mean Approximation Projection'. In: 12th International Conference on Information Technology in Medicine and Education (ITME), Xiamen, China, pp. 490-496. Available at: https://ieeexplore.ieee.org/document/10086207.

[25] Pisirir, E. et al. (2023). 'A Process for Evaluating Explanations for Transparent and Trustworthy AI Prediction Models'. In: IEEE 11th International Conference on Healthcare Informatics (ICHI), Houston, TX, USA, pp. 388-397. Available at: https://ieeexplore.ieee.org/document/10337312.

[26] Gov.uk (2017). 'Open Source guidance'. Available at: https://www.gov.uk/government/publications/open-source-guidance.

[27] Gov.uk (2014). 'How copyright protects your work'. Available at: https://www.gov.uk/copyright.

[28] Jain, K. and Kaushal, S. (2018). 'A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis'. Available at: https://doi.org/10.1109/ICRITO.2018.8748793.

[29] Rahul, C.; Aman, G.; Prabhat, K.; Chandradeep, B.; Ishita, U. (2023). 'Fine Grained Sentiment Analysis using Machine Learning and Deep Learning'. IEEE Conference Publication. Available at: https://ieeexplore.ieee.org/document/10303481.

[30] Sindhu, S., Kumar, S., & Noliya, A. (2023). 'A Review on Sentiment Analysis using Machine Learning'. In: IEEE Xplore. Available at: https://doi.org/10.1109/ICIDCA56705.2023.10099665.

[31] Thian, L; N, Ravikumar, R; Poorna, C, R, A; G, Komala; Krishnanand, M. (2023). 'Detecting Sentiment Polarities with Comparative Analysis of Machine Learning and Deep Learning Algorithms'. In: International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, pp. 186-190. Available at: https://ieeexplore.ieee.org/document/10141741.

[32] Chaithanya, A. (n.d.). 'Twitter and Reddit Sentimental analysis Dataset'. Available at: https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset.

[33] Hanna, Y. (n.d.). 'Stock Tweets for Sentiment Analysis and Prediction'. Available at: https://www.kaggle.com/datasets/equinxx/stock-tweets-for-sentiment-analysis-and-prediction

[34] Fred, E. (2015). 'Euclidean Norm - an overview'. ScienceDirect Topics. Available at: https://www.sciencedirect.com/topics/mathematics/euclidean-norm.

[35] H2O.ai (n.d.). 'What is Self-attention?'. Available at: https://h2o.ai/wiki/self-attention/#:~:text=Self-attention%20is%20a%20mechanism.

[36] Hugging Face (2024). 'google-bert/bert-base-uncased'. Hugging Face. Available at: https://huggingface.co/google-bert/bert-base-uncased.

[37] The Decision Lab (2023). 'List of Cognitive Biases and Heuristics'. Available at: https://thedecisionlab.com/biases.

[38] Kenton, W. (2021). 'How Pareto Analysis and the 80/20 Rule Work'. Investopedia. Available at: https://www.investopedia.com/terms/p/pareto-analysis.asp#:~:text=Pareto%20analysis%20is%20premised%20on.

[39] MindTools (n.d.). ‘Pareto Analysis’. Available at: https://www.mindtools.com/afzbk2y/pareto-analysis.

[40] GLUE Benchmark (n.d.). Available at: https://gluebenchmark.com.

[41] GLUE Benchmark (2019) ‘GLUE: a multi-task benchmark and analysis platform for natural language understanding’. Available at: https://openreview.net/pdf?id=rJ4km2R5t7.

[42] Weidong Y; Gang L. (2018) ‘Study on the relationship between investor sentiment and stock bubble’. Available at: https://ieeexplore.ieee.org/document/8407304.

[43] Yu-Xin, Xing-Hua, L; Wen-Jin, W; Yi-Jiao, L. (2021) ‘A Study of Relationship between Investor Sentiment and Stock Price : Realization of Investor Sentiment Classification Based on Bayesian Model’. Available at: https://ieeexplore.ieee.org/document/9325323.

[44] Rui, L; DianZheng, F; Zeyu, Z. (2017) ‘An Analysis of the Correlation between Internet Public Opinion and Stock Market’. Available at: https://ieeexplore.ieee.org/document/8110268.

[45] Python Software Foundation (2019). Available at: https://www.python.org.

[46] Matplotlib (2012). Matplotlib: Python plotting Matplotlib 3.1.1 documentation. Available at: https://matplotlib.org.

[47] Pydata.org. (2012). seaborn: statistical data visualisation seaborn 0.9.0 documentation. Available at: https://seaborn.pydata.org.

[48] Scikit-learn (2019). scikit-learn: machine learning in Python. Available at: https://scikit-learn.org/stable/.

[49] TensorFlow (2019). Keras, TensorFlow Core. Available at: https://www.tensorflow.org/guide/keras.

[50] www.pytorch.org. (2024). PyTorch. Available at: https://pytorch.org.