A critical analysis of the role of social media sentiment in predicting stock price movements

# Introduction and Objectives

Talk about the amount of money and focus on analysing and predicting stock price movements.

Beyond financial fundamentals like profit margins and operating costs, there are alternative market forces which can drive change or volatility in a stock’s price. One which has arguably grown in prominence with increasing global interconnection through the internet, is public sentiment. People can now share their thoughts and opinions across extremely large audiences, influencing huge numbers of people who may go on to make trading decisions on the stock market, influencing stock prices. Emotions are not an insignificant factor in influencing a stock price, and can lead to rapid stock increases or devaluations, but cannot be accurately tracked by company fundamentals. This paper looks to review how sentiment analysis and machine learning techniques can provide insight into numerically translating market sentiment with a view to predict stock price movements.

## Beneficiaries

Beneficiaries of this project includes anyone working within or involved in the field of stock market trading. This can include both large organisations like banks, and individual day-traders who are interested in understanding and predicting stock market movements. Furthermore, this research may be relevant to academics interested in the application of natural language processing (NLP) techniques to social media platforms like Twitter. The specific topic for this project is related to the company Apple but a number of the concepts discussed and utilised within this project may be used with a range of other applications involving NLP and Twitter, like business owners better understanding public sentiment towards their company or products in an unbiased situation outside of a survey or focus group. Furthermore, this research may be of interest to Apple itself and its employees. As a publicly traded company, the success of Apple from their shareholders’ perspectives may be viewed at a core, if harsh level, as their stock price, being directly related to the value of these shareholders’ holdings. This research may allow Apple as well as their shareholders to better understand the degree to which their stock price is related to, in reaction or precipitation of, public social media sentiment. It may also provide the stepping stone to understanding how Apple may influence or otherwise be a part of social media sentiment involving their products and company.

## Success Metrics

This paper covers a strictly business problem, and therefore its overarching goal is to make profit by predicting the stock price movements as closely as possible. As this is primarily a regression problem, success can be measured by Root Mean Squared Error (RMSE), which measures the error between each true and its corresponding predicted value, which is then squared, the average taken, and then square root is applied to ensure the result is ambivalent towards positive or negative signage. This will be evaluated on the holdout dataset which will be two months’ worth of data after the initial dataset (June and July 2021).

As the core issue is profit, a simple experiment will be performed on the optimal machine learning model to review whether an investment strategy following this model would return a profit.

* + Better understand the forces which drive the Apple stock price
  + Machine learning and sentiment analysis is able to quantifiably

# Context

* + Get some facts and figures about people doing this at banks

The total value of US-based publicly listed companies is extremely large at $47 trillion (Siblis Research, 2021). Apple forms a significant 5.12% at $2.41 trillion and is widely accepted as one of the most popular and influential companies in the world. Large corporations like investment banks and hedge funds spend large amounts of money, time, and resources to project how these stocks are likely to move based on a wide variety of metrics, market understanding, and news. As mentioned earlier in this project’s introduction, a medium which has grown exponentially in importance and usability over the last decade is social media sentiment analysis. As computers have become more powerful, machine learning on large scales have become more viable and social media is an ideal source of emotions, news, and market sentiment. If, for example, fear is building within the market, this may be visible and quantifiably measurable on platforms like Twitter. This project looks to build on other contemporary work in this field as discussed below to attempt to measure and predict Apple’s stock price.

The concept of psychology’s impact on stock prices is not new, with the basis of stock prices being undervalued or overvalued almost completely resultant from overly negative or positive mass psychology driving the price beyond its quantifiably logical price level. This paper will summarise the contextual review from the proposal below and add depth where appropriate.

There is strong literature in this field of utilising social media as a representation of sentiment to predict stock prices. Zhang, Fuehres, and Gloor found that average Twitter user sentiment in general could indicate how the Standard & Poor 500 (S&P 500) index would move as a whole (Zhang, et al., 2011). This paper provides an important foundation for a great deal of literature hereafter.

Meesad and Li effectively applied a sentiment analysis and an SVM linear algorithm to tweets on Twitter and focused Apple only, achieving a 90.34% accuracy for price movement (Meesad & Li, 2014). This is highly relevant to this paper, which has also elected to focus on Apple. However, it should be noted that the accuracy represented above is related to binary classification where the price either moves in a positive or negative direction. Upon reflection, as this paper is focusing on what is a business problem, there is some benefit to attempting to predict not only overall positive or negative changes, but also the degree of these changes, given that if an unusually high prediction is made, then the model should adjust accordingly and suggest a higher investment to maximise profit.

This paper was particularly motivated by Coyne, Madiraju, and Coelho’s paper on ‘*Forecasting Stock Prices using Social Media Analysis’* (Coyne, et al., 2017), which examined 17 stocks for a year’s worth of price movement data and utilised ‘twits’ from the social website ‘Stocktwits’ to create a prediction model. They were able to achieve an average accuracy of 65%, illustrating a degree of predictive power. An interesting relationship they captured was the optimal number of days between stock market sentiment and stock price movement. This is also something which this paper will review and reimplement.

Oh and Sheng have a similar paper on the stock predictive power of micro blog sentiment, where they support the notion that online social platforms, while at times irrational, can be used to predict stock market movements, with a weighted f-measure of 0.851 for one of their models (Oh & Sheng, 2011). A particularly interesting facet of this problem explored by Oh and Sheng is around the difference between ‘bullish’ and ‘bearish’ sentiment, where bullish refers to a positive sentiment and bearish refers to negative. A relatively accepted psychological phenomenon is that people tend to focus on and give more weight to negative information than positive, typically known as negative bias (Ducette & Soucar, 1974; Fisk, 1980). Translating this to stock sentiment context, this paper looks to explore also whether bearish sentiments are weighed more highly or have more significant relevance to stock price movements. Oh and Sheng also addressed a point earlier from Coyne, Madiraju, and Coelho regarding the number of days between social media sentiment and stock price movement, suggesting that investors may hold to their prior beliefs for a period of time when faced with new news, causing a delay in stock price movement (Daniel, et al., 1998; Chan, 2003). This paper will address this by using a similar strategy to Coyne, Madiraju, and Coelho and creating multiple models for a range of time periods between social media sentiment and stock price movement. This will ensure the optimal range is met, while also illustrating more of the relationship between new news and time delays to the stock market.

However, this paper also considers opposing work like Oliveira, Cortez, and Areal’s paper which similarly uses data from Stocktwits but disputes previous work by authors like Oh and Sheng, suggesting that sentiment is not a strong predictor for stock market returns (Oliveira, et al., 2013). Furthermore, they suggest Oh and Sheng did not utilise a large enough test period for significant results. However, this paper notes that Oliveira, Cortez, and Areal utilised a multiple regression model in this analysis, which this paper suggests may not be complex enough to fully capture the depth of sentiment and price movement relationship. This paper will look to match this concern by trialling multiple neural network-based models like Multilayer Perceptrons and Support Vector Machines, which will add scalability from model depth. Furthermore, this paper will utilise data from a test period of a year’s length to ensure sufficient length.

## Context Summary

Overall, one key point this paper will attempt is to treat this as also a regression problem as opposed to purely a classification as a number of previous papers in the field do. While classification does provide understanding on the overall relationship between social media sentiment and stock price movement and the degree of strength of such a relationship, but this paper believes a regression approach will both provide a potentially superior method for maximising profit, as well as illustrate a greater depth of the relationship between sentiment and stock prices.

With regards to the frame of the business use case provided, this paper is open to trialling any appropriate machine learning model which may perform under the circumstances, and will experiment in particular with the Multilayer Perceptron, Multiple Regression and the associated Ridge, Lasso, and Elastic Net regression models.

Given time, resource, and scope restrictions, this paper has elected to focus on the Apple stock only. This allowed the researcher to dedicate a greater deal of time and effort in better understanding and modelling the specific intricacies related to the Apple stock, providing a more in-depth analysis.

* + Quants?

From the above contextual research, this paper has derived two main hypotheses which will drive project work. These have evolved fairly significantly from this paper’s initially stated research questions within the project proposal, but now more accurately fit this project’s scope and capabilities. Examining financial metrics would bring an additional dimension to the project, but would somewhat obscure focus on the main part of this project which is the sentiment analysis. Furthermore, thoroughly examining the differences between bullish and bearish sentiments’ impact is not necessarily appropriate for a project reviewing regression approaches.

**1st Hypothesis:**

**H0**: Sentiment derived from Twitter has no provable correlation to Apple’s stock price movement and cannot be competently modelled using this project’s selected model.

**H1**: Sentiment derived from Twitter has a provable correlation to Apple’s stock price movement and can be competently modelled using this project’s selected model.

**2nd Hypothesis:**

**H0**: Sentiment derived from Twitter cannot be utilised to predict Apple’s future stock price movements with any acceptable degree of performance using this project’s selected model.

**H1**: Sentiment derived from Twitter can be utilised to predict Apple’s future stock price movements with an acceptable degree of performance using this project’s selected model.

While these are similar hypotheses, this paper believes the addition of the time dimension with the second hypothesis sufficiently alters the question to require an additional hypothesis. While subsuming both under one hypothesis would be ideal, and it is true that logically the second hypothesis is dependent on the first, it is entirely possible that the first null hypothesis will be rejected but the second null hypothesis will be accepted, leading to an entirely different conclusion that must be clearly defined and examined. Separating these two topics into two hypotheses which flow from first to second allows this project to define and state this project’s results more clearly.

# Methods

## Data Gathering

Initially, this project looked to follow the work of Coyne, Madiraju, and Coelho by using text data from the trading social website Stocktwits (Coyne, et al., 2017). This site consists of essentially a trading-focused community where members may post their opinion into the public area and discuss. However, this paper found that since then, obtaining permission and access to Stocktwits’ database has become untenable through current company administrative limitations. Therefore, this paper reasoned that Twitter is a highly suitable alternative dataset. While it differs from the dataset used by Coyne, Madiraju, and Coelho, it is also likely to have a far larger base of information as Twitter’s monthly active userbase is approximately 310 million, compared to 1.5 million from Stocktwits (Roof, 2016; Statista, 2021). Furthermore, while Stocktwits is more likely to have an informed audience relative to the context of this project’s problem, this paper argues that more people invest in large blue-chip companies like Apple than is contained within Stocktwits, while also overall sentiment of the population and market is more likely to be better captured within a larger population like that of Twitter. Therefore, this paper concludes that Twitter is a highly appropriate source of sentiment data for the purposes of this analysis.

The primary dataset of text was sourced using the Twitter API to gather all tweets within the specified timeframe of 12/05/2020 to 11/05/2021 containing the hashtags “#apple”, or “#AAPL” (relating to Apple’s US stock market tag). Gathering this data was a significant activity and less straightforward than initially projected. The opening approach selected to obtain the tweets was the Postman tool. This has a user-friendly interface and an uncomplicated query builder. However, the researcher quickly noted that a core part of retrieving tweets from Twitter is the limitation on academic projects for a maximum of 500 tweets per request. While a key known as the “next token” provides a link to the subsequent part of the query result, it was clear the Postman tool was not easily optimised to automate gathering the full set of data. Therefore, this project reviewed alternative options to gather the necessary data. One option reviewed was Tweepy, a python library built to facilitate tweet retrieval. While fairly commonly utilised, this project had issues with customisation and implementation for this library, and was unable to accurately retrieve the desired results with the required fields in a timely manner.

The alternative which was implemented was to adapt manual Python code which directly interacts with the Twitter API without a ready-made interface, using the ‘searchtweets’ python module. This module was found to be slightly clearer than Tweepy, and this project was able to work to obtain the required results. However, this process was reasonably lengthy as there was a great deal of customisation required such as how to structure the loops, data formatting, and storage. Furthermore, the Twitter API has a number of restrictions on the number of tweets which may be requested over a variety of timeframes, and therefore a series of IF statements needed to be adapted to the code to pause it when necessary, ensuring the process would run uninterrupted and without any automatic throttling from the Twitter API. Several times, this project encountered difficulties with regards to this issue, where tweet retrieval would take an inordinate length of time as the program needed to keep pausing to ensure it remains in line with Twitter’s tweet retrieval restrictions. This project managed this issue by streamlining the time management IF statements, as well as retrieving part of the data initially to work on while running the program on a separate machine so as to mitigate the delay this caused.

However, this process was positive through relative comprehensibility and flexibility as the dataset was constructed on the familiar and flexible Python Jupyter notebook interface. This meant it was relatively easy for researcher to adjust and optimise sections of the process to best fit this project’s aims.

The Apple stock price dataset was sourced directly from the NASDAQ website which contained a variety of information regarding the stock price, including its daily open and close prices, the trading volume, and the high and low for each day. For the purposes of this project, the daily price change was extrapolated and merged alongside the relevant dates for each text entry retrieved from Twitter, allowing the project’s models to view each tweet in context of that day’s price change.

## Data Pre-processing

Data pre-processing for textual data which this project is personally sourcing was a significant section of this project’s work, involving several steps and judgement considerations.

One significant pre-processing step was to ensure the full text of each tweet was delivered. While standard tweets were satisfactory, retweets had their text automatically truncated, minimising the available text data for these tweets. A retweet is “a re-posting of a Tweet. Twitter's Retweet feature helps you and others quickly share that Tweet with all of your followers” (Twitter, 2021).

This paper considered and concluded that while a retweet is not the same as someone writing and expressing their opinion, it can be one’s way of expressing one’s opinion through the medium of another’s words. Therefore, this paper can assume that if a tweet is retweeted, the user has read the initial tweet and feels a similar sentiment which they wish to echo amongst their online social group. Therefore, this paper has concluded that these retweets are as important as original tweets and will be treated as such. Consequently, the retweet text field truncation would need to be rectified to ensure the machine learning model had the most accurate picture possible of the Apple-related tweets on Twitter.

This problem was ultimately solved by identifying which tweet the retweet referred to and copying the original untruncated text from the original tweet onto the retweet, ensuring there would be no truncation and compromise with the sentiment analysis. This paper achieved this by performing list comprehension within a lambda function to isolate each retweet’s referenced tweet number, which was then merged with a copy of the dataframe to obtain the retweet’s correct text on the correct row. From here, the correct text needed to be copied over to the correct row, only where necessary. Looping for this activity was briefly attempted with the “iterrows()” function but was found to take an unacceptably long time to process due to the large size of the dataset (3,390,454 rows). Therefore, this paper trialled several other methods and concluded that NumPy vectorisation was the optimal solution, at approximately 71,803 times faster than a standard loop and 224 times faster than the iterrows() function initially utilised (Droste, 2019). This was, however, noticeably more difficult to implement than the simpler loop, as it necessitated a different logic and perspective. This was nonetheless accomplished by creating a second dataframe with only the retweeted objects which could be merged with the relevant original tweets to obtain the correct original text. Noting this problem in a timely manner was another benefit of taking an initial sample of the dataset to work with as mentioned earlier, which allowed this project to note any unforeseen issues and ensure the full retrieval of the dataset contained the necessary additional fields required to solve this problem. The unnecessary additional columns created by the earlier merging as well as the ‘referenced tweets’ column was then removed, leaving only the now complete text and accompanying data.

As mentioned earlier, the relevant price data was modified to illustrate the price change for each day and then merged against the tweet data to show the price change on the day for each tweet. Matching the dates together required this paper to correctly pre-process the dates for the tweets into an acceptable format. For example, a raw retrieved tweet date may be “2021-05-11T11:43:58.000Z”, the accepted ISO 8601 format for writing an event’s date and time alongside time zone. However, this paper’s scope extended to only requiring the date itself, and therefore the following would need to be cut to present only the date in ISO 8601 format. Similarly, the stock price data from the NASDAQ was delivered in its raw format as the American format of date, “05/11/2021” for example. This also required pre-processing in the form of parsing the date into the required ISO format using the dateutil Python library.

One initial obstacle this paper noted was that the price change data was only available for days on which the market was open, meaning that for weekends and public holidays, there would be no relevant price data. However, this paper quickly concluded that removing tweets that did not match would not make sense, as while sentiment may not, at the time, have an immediate impact on the price, it is logical that news and sentiment over this time may build up and impact the next day the markets do open, both through people independently planning to make trades on the day as well as implementing buy or sell orders to occur as soon as the markets open. Therefore, this paper created a function which would take the tweet date, assess whether it was a weekend or public holiday, and create a new date column with the next available date when the markets would be open. While functional, this paper notes that an assumption has been made with regards to the relationship between tweets made on non-trading days and how that impacts the price. This paper has assumed that the relationship does not change and simply accumulates till the next available trading window, whereas it is possible this does not capture the full extent of the relationship, specifically on certain public holidays which may promote unique behaviour. However, this paper has made the judgement that any missed particularities in this relationship are likely to be minimal in terms of their impact on the price change, and therefore has been covered as per the function created.

## Data Cleaning

The tweets collected by this project, while all referencing “#apple” or “#aapl”, will come with a huge variety of words, phrases, and noise which are not constructive for sentiment analysis. The overall goal of sentiment analysis is to convey the key words, phrases, and meaning numerically which can be inputted to a machine learning model. The analysis inputs need to be as simple as possible while condensing as much conceptual meaning as possible into the text inputs. Therefore, before any meaningful text analysis, the data must be cleaned. While text subjectivity makes this a difficult process to carry out with comprehensive coverage, there are particular processes and checks to remove or modify any common errant issues with the text which would otherwise compromise the sentiment analysis.

### Emoji

Emoji and related emoticons are an integral part of many people’s modern internet parlance. They are typically utilised to convey typically facially significant emotions which otherwise are difficult to translate into words and messages. Considering this paper’s subject concerns sentiment analysis, emoji is a potentially powerful source of emotion which can be utilised by this paper to better understand and quantify a tweet’s sentiment. In order to utilise them, this paper utilises a Python library called ‘emoji’ which can convert emoji in a text string into the accepted text explanation. This can then be utilised as an input into TF-IDF, empowering the model when else it may be simply removed.

### Special Characters

Special characters excepting emoji like “@\*+^” were judged to be not specifically important to this project’s sentiment analysis and were therefore removed from the text strings where found. This also contributes to a simpler and cleaner sentiment analysis resource for the machine learning models to utilise.

### Links

Many tweets encountered in this project contained webpage links to articles and sites. While this are relevant for a Twitter user to read to gain more information and otherwise better understand the tweet’s sentiment, as far as sentiment analysis goes, these links are ineffective in discerning meaning. One way in which this paper could have potentially allayed this problem would be to search each link and provide a sentiment analysis on the contents of each. However, this introduces an unregulated number of additional sources which this project has determined is beyond its scope to manage. Therefore, these links when found by identifying where words start with “http” were removed from the dataset.

### Lowercase

Part of reducing the number of variations on words which are conceptually identical is ensuring all characters utilised in this data set are lowercase, further reducing unnecessary complexity to the TF-IDF process.

### Lemmatisation

One issue sentiment analysis may encounter is the significant variety of words available to describe conceptually similar ideas, or how affixes may alter the spelling of particular words, leading them to be read as separate entities by the machine learning model. Lemmatisation is the process of returning a word to its basic format as can be typically found in a dictionary. An example is changing “carried” to “carry”. This minimises the number of conceptually similar words being recorded as different objects and simplifies the machine learning model. It may also return stronger results as repetitions of the same concepts will now be represented more clearly and significantly, creating a more accurate model. This project utilised the Natural Language Tool Kit (NLTK) library in a small function to firstly tokenise the tweet strings, which separates each tweet into its individual words which is a requirement for lemmatisation. This allows the researcher to consider each word as an individual element in isolation, allowing for effective lemmatisation. Then a lambda function was utilised to efficiently lemmatise each word, which was then detokenised to recreate the tweets. This last stage is important to ensure the TF-IDF function and resultant model can clearly understand each tweet’s word structure and process them in a computationally efficient manner.

### Stopwords

Stopwords are common words like “the” or “is” which do not hold any conceptual meaning but are extremely common. Removing these are important to ensure these do not impair the sentiment analysis by being considered as features. These words were removed in this project by using the NLTK’s list of English stopwords, comparing against this and removing where appropriate.

## Data Analysis

### Requirements Analysis

Given the basis of this project is a fundamental business problem, the aim and subsequent requirements are clear. The overall goal is to maximise profit which can be gained by predicting the stock price movements as closely as possible. Since this is a regression problem, metrics like accuracy are not relevant, so success will be measured primarily through Root Mean Squared Error (RMSE). RMSE may be defined as the square root of the mean value taken from the summation of the squares of all recorded errors (Neill & Hashemi, 2018). Furthermore, this paper will also explore the subsequent two months (June and July 2021) after the initial data collection as a holdout set to test the data on and collect final results.

### Design

Following the requirements, as this project reviews a business problem and not a pure academic exercise, this paper will review a range of regression models to examine which best fits the problem and data. As mentioned earlier, this paper will not review classification methods as has been performed in literature, as this paper believes regression may offer a different and higher performing perspective with respect to profitability. This is because while the investment decision could be simplified to only three main options, buy, sell, or no action, this paper believes there is more scope to this problem, as the degree of investment is a highly relevant factor and can be exploited through a regression approach.

The following subsection will review the models this paper has elected to examine within the context of this project’s problem, discussing their strengths and weaknesses, and how they may be optimised.

### MLP

The Multilayer Perceptron (MLP) is a supervised learning feed-forward network with three layer categories, input, output, and hidden. Random weights are assigned to each neuron in the hidden layer to ensure they do not get stuck in local minima, influencing an output as values are passed through the system. Backpropagation then occurs when the output value is different to the true value, and the hidden layer’s weights are correspondingly updated to better fit the data. MLP is a popular tool for both classification and regression problems as it is accurate, scalable, relatively quick to train, and highly customisable.

Icon

Description automatically generated

Figure 1: A Concept of a Multilayer Perceptron

### MLP: Hyperparameters

This paper has discussed below the MLP hyperparameters which are available for optimisation.

#### Neuron Number

When discussing neuron number, there are three different relevant categories, the input layer, the hidden layer(s), and the output layer. The neuron number for the input and output layers are fixed to the dimensions of the dataset and desired output. The input neuron number will be fixed to the number of total words in the corpus after the TF-IDF function, which will be discussed further below. To summarise for this section, each unique word in the dataset will be considered with a separate input neuron which is fixed and cannot be optimised as part of this. Similarly, the output neuron number is fixed as there is only one desired output, an overall prediction for the stock price. The hidden layer however can be optimised to improve performance. Too few neurons may mean there is not enough capacity within the model to fully learn the dataset provided, leading to underfitting. Alternatively, too many neurons will likely lead to overfitting as the model adapts to noise unrepresentative of the population, as well as increasing computational time and effort. Heaton suggests the value could be 2/3’s of the input layer, plus the output layer, or also less than 2 times the input layer (Heaton, 2015). Based on this, this paper will attempt a variety of neuron numbers from 20 to 11,688 to understand what number of neurons returns the best RMSE score.

#### Learning Rate

Learning rate refers to the rate at which the model adapts to the results of each epoch. In each epoch, the model will review which of its predictions were incorrect and the error between each prediction and its corresponding true value. The model will then adapt to this, attempting to reduce the loss. Learning rate is the rate at which the model adapts. Increasing this may lead to a faster training model or a model which is overall superior through more training, but may also overtrain or train past what would have been optimal solutions, without allowing enough epochs between them to clearly examine them.

#### Momentum

Momentum is closely related to learning rate and refers to an additional factor added to the neuron weighting calculations to increase training speed and ensure the model does not get stuck in local minima. What this refers to is how when training, a model may reach what appears to be locally the minimum loss for the model. Without momentum, the model may remain in this local minimum, despite the possibility that the global minimum may be beyond. Momentum allows the model to move past local minima to more fully explore and find the global minimum. There is research however which does suggest that local minima are relevant and at times, spending additional effort to find the global minima may also result in an overtrained model (Choromanska, et al., 2015). Taking this into account, this paper will opt for a balanced approach, considering a range of momentum values.

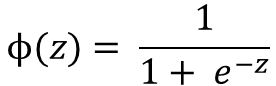
#### Epoch Number

Epoch number represents the number of times the model retrains with the aim of iteratively improving each time. In this regression problem, this is measured by training and validation loss. Increasing the number of epochs will likely improve model performance but may also cause overfitting. This occurs when a model is fit too well to a training dataset and has also modelled what is meaningless or counter-productive noise within that specific dataset. Therefore, when this model is applied to a test or holdout set, performance decreases as said noise is not present in the population, represented by the test set. To better manage this, this paper has elected to use early stopping as a form of regularisation. This observes validation loss and concludes model training when this metric has stopped improving in a meaningful way for a particular number of epochs known as the patience value. This is set at 5 epochs.

#### Activation Function

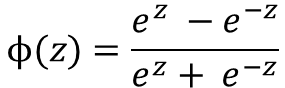
The activation function is a transformation applied to each hidden layer which supports neuron weight development. This is a key step as it is a significant part of how an MLP can be more sophisticated than a simple linear regression model. There are three main activation functions available to evaluate, rectified linear activation function (ReLU), Sigmoid, and Tanh. ReLU is a common modern activation function which does not suffer from the vanishing gradients problem like other neural networks. This problem occurs when a deep neural network (one with many layers) cannot properly complete backpropagation, meaning useful learnt information from the end of the model does not return to the beginning of the model. ReLU can be described as returning the input, or 0 if the input is 0 or less. Despite this initially seeming potentially very inconvenient for datasets containing a large number of negative values, weights are applied to each neuron, transforming them so they are likely still of use to this project’s dataset. Furthermore, literature has shown that ReLU trains up to six times faster than Tanh, another alternate activation function (Krizhevsky, et al., 2012).

The sigmoid activation function simply takes the input value and returns a value in the range of 0 to 1, with more positive values returning an output closer to 1 and more negative returning closer to 0. The function which governs this relationship is:

 (Sharma, 2017)

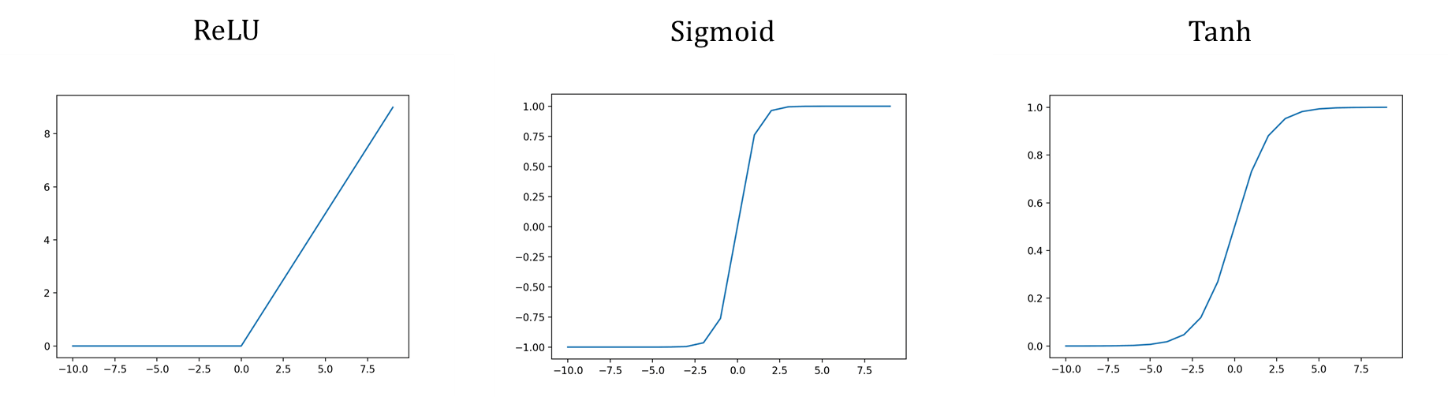
Where: z = the input value

The tanh activation function is similar to sigmoid but its lower bound is -1 as opposed to sigmoid’s 0. This is calculated as:



Where: z = the input value

These activation functions are graphically represented as below:



(Brownlee, 2021)

The above activation functions will be trialled with this project’s MLP to assess which is most suitable. Alongside performance with regards to RMSE, computational time and effort will also be assessed as this project is managing a significant level of data. Furthermore, the business application for this program may require quick retraining with new data available at a near-constant flow through new tweets, thereby reinforcing the need for a model which is computationally efficient.

#### Hidden Layer Number

The number of hidden layers is also a hyperparameter which may be optimised for this project’s MLP model. Increasing the number of hidden layers will increase the depth of the model, which may lead to greater development of the neuron weights, leading to a more accurate model. However, too many hidden layers may lead to the vanishing gradients problem, particularly when not utilising the ReLU activation function. Furthermore, the model may overfit to the training dataset, reducing test performance while also requiring unnecessary computational effort.

### Multiple Regression

Multiple regression is a form of linear regression which attempts to fit a linear model to the dataset which minimises the residual sum of squares between the true values and predicted. The term ‘multiple’ refers to how it considers multiple variables to model this relationship. In the case of this dataset, variables refer to the total number of words used in the dataset, also known as the corpus, where each unique word is represented by a variable. This is a simpler model when compared to MLP, but this project has determined it is worth investigation, particularly as Oliveira, Cortez, and Areal also utilised this model, which allows this project to compare results and review their conclusions more accurately. They suggested sentiment is not a useful indicator for stock market returns (Oliveira, et al., 2013), but as mentioned earlier in this project’s context review, they utilised a multiple regression model. This paper therefore believes it will be notable to also trial this model and review it in context of the others.

### Optimisation

### Ridge Regression

Ridge regression is similar to multiple regression but uses L2 regularisation to minimise any issues relating to multicollinearity which may arise from using multiple regression. This is performed by imposing a penalty on coefficient size equalled to the magnitude of coefficients, squared. This reduces the impact any collinearity may have on the results. Multicollinearity was expected by the nature of sentiment analysis where some words have conceptually similar meaning, and therefore it stands to reason that their ratings may be similar. However, this project has taken steps to mitigate this through lemmatisation as mentioned earlier, which reduces words to their core meaning, reducing any potential collinearity. However, it is possible that not all the words within the corpus were correctly identified and lemmatised, and therefore ridge regression is an acceptable model to trial for this project.

Optimising Ridge regression is also a relatively simple matter as there is only one meaningful hyperparameter available for optimisation, the alpha value. This refers to the regularisation strength, which is the penalty applied to the variables to reduce collinearity. This was optimised by iteratively searching a logarithmically spaced range between 0.00001 and 100 and scoring by RMSE, where the optimum alpha value was found to be 2.68.

### Lasso

Lasso regression is also similar to multiple regression but uses L1 regularisation, which applies a penalty to features which are determined to be less important as a result of the quantity of zero values they contain. Therefore, a rare word which does not appear much within the dataset is more likely to be ignored by the model to reduce complexity and improve performance. This model was selected this project has an especially large number of words/features (5844) in a sparse model with many zeros, which suits the usage of Lasso (Nagpal, 2017). This model has the same hyperparameter as ridge regression and therefore has been optimised in the same way, using a logarithmic scale to iteratively test alpha values. The optimal alpha value was found to be 0.000019.

### Elastic Net

Elastic Net regression combines L1 regularisation from Lasso and L2 regularisation from Ridge. Given the reasoning above for including Lasso and Ridge, it is pertinent to include Elastic Net within the models trialled by this project. Hyperparameters for this include alpha as discussed earlier, and also L1 ratio, which is the ratio of L1 to L2 regularisation this model utilised, with a high value emphasising L1, and a lower value emphasising L2. The optimal values for alpha and L1 ratio are 0.0013 and 0.56 respectively.

As SVM was utilised by some authors within this field like Meesad and Li, this paper reviewed implementing SVM within the scope of this project. However, while SVM has many strengths, this paper’s dataset consists of 3,390,454 data entries which made the prospect of training and utilising an SVM-based model discouraging. To test, this paper trialled a basic SVM model (SVR from Sklearn) with no optimisation whatsoever to gain an understanding of the computational effort required. This model took 34.8 hours to train once, which is an exceedingly long time, especially considering the optimisation process would require many iterations of this training process with variations on SVM hyperparameters. While this paper could take a fraction of the data to optimise the model, this may not produce the optimal parameters for the full dataset. Furthermore, additional experimentation with dimension reduction and other methods to mitigate computational effort did not yield fruitful results. Moreover, as per the business use-case for this paper, daily retraining of the model would be the optimal scenario to ensure the model has the most up-to-date information, and a model which takes more than 24 hours to train is unlikely to be fit for purpose, and therefore this paper elected to not include SVM as part of this project.

* TF-IDF

### TF-IDF

Sentiment analysis is essentially transforming qualitative text into numerically tangible data which can be read and analysed by a machine learning model. This is done by analysing each word given the context and providing it with a weight which denotes its importance. There are several different methods by which this sentiment analysis may be performed, with this paper reviewing three methods of word weighting term frequency (TF), inverse document frequency (IDF), and term frequency – inverse document frequency (TF-IDF).

TF utilises the frequency of terms used within the text to identify importance but lacks as it does not consider these terms with regards to the corpus as a whole, and therefore lacks perspective. IDF considers a word’s importance as the total document number divided by document number including the particular word. This is an advantage over TF since it considers the entire corpus as opposed to a single document which is unlikely to be 100% representative of the population, given the variety of tweets and language people may use.

TF-IDF considers the importance of a word relative to the document and the dataset corpus. Initially proposed in 1972 under the phrase “term specificity” (Jones, 1972, p. 11), it has been shown to be more accurate than TF (Cui, et al., 2015), likely because it considers the entire corpus as opposed to a single document, leading to a more grounded feature score. It was also found to be significantly more accurate than IDF as well, leading this paper to consider TF-IDF as its primary tool for sentiment analysis, combining benefits from TF and IDF.

The TF-IDF function from Sklearn returns a Scipy sparse matrix, which is a high-level representation of each word contained within the tweet dataset and its importance weighting. This by itself is not immediately useful to the project, but can be utilised to create an array representation which quantifies each tweet and enumerates each word in the corpus, numerically representing each tweet in the context of all the words utilised throughout the dataset. This can then be used as an input into machine learning models which will then attempt to predict the price change relating to each tweet.

### Days Ahead Modelled

In order to test this project’s second hypothesis regarding the longer-term predictive capacity of the tweets in question, this paper re-reviewed its previous steps in data pre-processing, specifically with regards to the function created to sort tweets which occurred on weekend and national holidays. This paper needed to create a new parent function which encapsulated this but dealt better with looking multiple days ahead. Initially, this function simply added days to the input day. However, as it stood, this original function had difficulties with managing looking several days ahead when these days included the weekend. Currently, it counted the weekend as part of days ahead which was unacceptable for the purpose of this paper. This paper iteratively developed solutions and ultimately created a function which accepted the number of days ahead desired as an input, and moved ahead one day at a time, and checked each day for whether it was a weekend or public holiday. It then utilised several lambda functions to apply this function to create a new column for each variation appropriate, and iteratively merged dataframes to obtain the stock price movement for each day required.

### Language

Twitter is a global company, and therefore has users from across the world, speaking and using different languages with different scripts. Each language will have unique phrases and meanings which are not shared by others, which may pose some difficulty for a machine learning model. As people over the world may invest and otherwise influence Apple’s stock price, it is relevant to attempt to examine these differences and this project will attempt to create separate models by language to review its impact on performance. Initially, this project created a simple count to illustrate the distribution of tweets among the languages utilised. This resulted in a list of 64 languages, each with their respective number of tweets. However, the vast majority of these were numbers low enough to cast doubt on whether any model and resultant RMSE score generated would be reliable, relative to the population number. Therefore, this paper elected to investigate only tweets in languages numbering at least 100,000 over the year examined. The table below illustrates those languages, which accounts for 83.37% of the tweet population, a large proportion from only 5 of the 64 total languages.

Table 1: Tweet Language Distribution (For languages totalling over 100,000)

|  |  |
| --- | --- |
| Language | Tweet Count |
| English | 1,962,786 |
| Japanese | 346,167 |
| Korean | 216,566 |
| Spanish | 162,353 |
| French | 136,451 |

### Implementation

For the MLP hyperparameter optimisation, there were multiple options available. The two most optimal were found to be Gridsearch and Randomsearch, both functions from Sklearn. Gridsearch is a system by which a set of trials are created by forming and iteratively testing every possible combination of the allowed hyperparameter values. Randomsearch on the other hand, is an approach by which each hyperparameter is treated as a range from the lowest to highest allowed values, and takes a random value from each hyperparameter range and tests it. This is then repeated as many times as requested, and returns the optimal hyperparameter values obtained.

While Gridsearch is generally more thorough, it is extremely time intensive as it requires the model to iterate for each possible hyperparameter configuration. Furthermore, it suffers from what is known as the ‘curse of dimensionality’, where the number of combinations increases exponentially with number of hyperparameters (Bellman, 1961), further amplifying this concern. Contrary to this, Randomsearch does not run every configuration, but instead runs a number of randomised configurations according to the researcher’s prior set hyperparameter distribution. This has been found to be far less time and computationally intensive, as well as producing strong and arguably superior results (Bergstra & Bengio, 2012). The Randomsearch approach was utilised for the hyperparameter optimisation of all other models reviewed as part of this project due to its strengths as discussed above.

For hyperparameters where Randomsearch was not appropriate, a Gridsearch-style iterative approach was taken to iterate through potential options like activation function, neuron number, and hidden layer number. This process also includes cross-validation which provides further confidence on model optimisation by splitting the data into 5 folds and training and validating the model on each individual fold.

* + Testing/Evaluation

# Experiential Results

## MLP Optimisation Results

Through a series of iterative trials, the optimum neuron number was found to be 160. Initially, this paper attempted a basic test of neurons numbering between 20 and 100 with intervals of 20, which returned 100 as the most optimum. From here, following Heaton’s suggestions, this paper attempted a test of 3,896 and 11,688, which were 2/3’s of the input layer plus the output layer, and 2 times the input layer respectively. The optimum RMSE score was still found to be 100 of those trialled, so this paper then attempted values between 100 and 200 with intervals of 20. The optimum value was found to be 160 due to this paper’s assumption of a linear relationship between neuron number and RMSE. The highest performing number of hidden layers was found to be one.

The optimum learning rate was optimised with Randomsearch as mentioned earlier and found to be 0.0016. This is a notably low learning rate considering the upper bound for values tested was 1, and will ensure the model can gradually optimise without overshooting the global RMSE minimum. Momentum was found to be optimal at 0.94 which is relatively high, ensuring the model does not get stuck in local minima, and the optimal activation function recognised as ReLU. Optimum epoch number was also optimised as part of this process and found to be 20.

With regards to the test set results of the trial of the same day price with all languages, this paper’s trial of multiple regression returned an RMSE score of 2.06. Ridge regression returned an initial unoptimised RMSE score of 1.9896, and optimisation of this model’s alpha value to be 2.68 returned an only slightly marginally superior RMSE value of 1.9892. A basic Lasso regression model initially returned 2.05, and optimisation of the alpha hyperparameter at 0.001 returned a relatively superior value of 1.98. Elastic net regression initially returned a RMSE of 2.05, and optimisation of the hyperparameters alpha and l1-ratio returned a value of 2.047. However, as we can see from the above results, we can understand that all models examined had difficulty with predicting the stock price movement on the same day of the tweets within the holdout set, suggesting for this parameter, significant overfitting as the training and test RMSE scores were typically far lower. However, all models generally performed far better when reviewing one and more days ahead.

The following results were generated which illustrates the performance of each trialled model on each language and days ahead reviewed. Including the training and test RMSE scores allows this paper to have a deeper understanding of the results, particularly where overfitting may have occurred.

## RESULTS

Across each language subset, multiple regression did not perform particularly well with the partial exception of French, but still returned typically the worst RMSE scores of all models on the holdout set.

Another phenomenon noted is that generally each model performed better on the holdout set than the training and test set, with some exceptions. One notable exception is Korean which all models performed worse on with the holdout set, compared to the training and test sets.

Furthermore, we can see from the results that MLP typically performed better than all other models with regards to the training and test sets but did not improve to the degree some other models like Elastic Net did on the holdout set.

Overall for the models trialled by this paper, the most superior with consistently the lowest RMSE score on the holdout set was found to be Elastic Net. In particular, for each language subset as well as the total dataset, the optimal value was found predicting two days in advance, with the best RMSE value recorded by this project as 1.307 for Elastic Net when using only Japanese tweets.

Further Ahead Studies

For the aspect of this project which reviewed modelling stock price movements based on Twitter data historic relative to the price, six different iterations of the optimum MLP model was reviewed, the initial model evaluating the stock price on the day of tweets, then from one to five days ahead of the tweets in question.

Table 2: Table illustrating the RMSE scores of the Days Ahead Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Days Ahead | 0 | 1 | 2 | 3 | 4 | 5 |
| RMSE | 1.992 | 1.853 | 1.955 | 1.931 | 1.934 | 1.985 |

As can be seen from the table above, the model with the most promising results was the iteration modelling two days ahead of the tweets. One can also see that of the six model iterations trialled, all were superior to the initial model predicting the price of the same day.

With regards to this project’s hypotheses as restated below, this paper suggests for the 1st hypothesis which examines whether sentiment derived from Twitter has any provable correlation to Apple’s stock price movement and whether this can be competently modelled using this project’s selected models, that initially while the RMSE score for the same day testing was relatively low for the training and test datasets, the models generally performed very poorly on the holdout set, and would therefore suggest that the null hypothesis (H0) is accepted. However, as mentioned earlier, these two hypotheses are intrinsically linked, and the examination into modelling a number of days ahead was notably more successful than modelling the stock price using tweets from the same day. This suggests that null hypothesis for the 2nd hypothesis should be rejected, and therefore suggests the null hypothesis for the 1st should also not be accepted.

**1st Hypothesis:**

**H0**: Sentiment derived from Twitter has no provable correlation to Apple’s stock price movement and cannot be competently modelled using this project’s selected model.

**H1**: Sentiment derived from Twitter has a provable correlation to Apple’s stock price movement and can be competently modelled using this project’s selected model.

*Conclusion: Reject H0*

**2nd Hypothesis:**

**H0**: Sentiment derived from Twitter cannot be utilised to predict Apple’s future stock price movements with any acceptable degree of performance using this project’s selected model.

**H1**: Sentiment derived from Twitter can be utilised to predict Apple’s future stock price movements with an acceptable degree of performance using this project’s selected model.

*Conclusion: Reject H0*

# Discussion and Critical Evaluation of Findings

This section of the project will be used to discuss and evaluate this project’s findings within the scope of this field’s contemporary literature as reviewed earlier.

An important note to consider with this project’s discussion and evaluation is that the vast majority of all the literature in this field do not consider this subject with the frame of a regression problem, but rather a binary classification problem, separating between bullish and bearish sentiments. This paper discussed earlier the advantages and disadvantages of utilising a regression approach as opposed to classification, and while classification does have benefits, this paper feels a more optimal approach with regards to the business problem that this project attempts to solve is a regression analysis, which may provide greater benefits to this project’s potential beneficiaries. With regards to this section, direct comparison is hampered slightly with regards to this as this paper cannot effectively compare metrics due to the fundamental differences between how classification and regression models are evaluated. For example, comparing accuracy to a reasonably scientific degree is not possible, and therefore this paper will review and discuss its models’ results relative to each other, and the shared concepts explored and discussed in literature.

## MLP

With regards to MLP optimisation, this paper notes that in order to better manage the significantly high upper bound neuron number this paper was assessing, it optimised based upon the assumption that neuron number to RMSE performance behaved as a linear relationship. While this allowed this paper to mitigate the otherwise extremely lengthy training times required, this paper reflects that this relationship, while unlikely to be highly deviated from linear, may follow a moderately different relationship, suggesting that the optimum neuron number may be different to the conclusion of this project. While all things considered, this paper deems this to be unlikely, the possibility remains and is therefore fairly acknowledged. Furthermore, this paper notes that the optimal value of 160 neurons does not fit Heaton’s suggestions on optimal neuron numbers, given that Heaton suggested between 3,896 and 11,688 for this model given its dimensions. This paper suggests that a possible reason why the optimal neuron number was found to be outside the range suggested by literature is because in this project, each input neuron is represented by a unique word within the dataset’s corpus. This paper suggests that while many words are very common and standalone, it could be that many words come together as common phrases to form meaning which is contained within a hidden layer neuron. There is an issue with fully understanding this phenomenon however, as while an MLP is a supervised neural network, it does operate as a ‘black box’, essentially meaning that it is very difficult to deconstruct and comprehend its active processes beyond theory.

The learning rate was extremely low, only 0.0006 higher than the lower bound of values tested for this hyperparameter. This suggests the minimum for RMSE occurs early with this model and a higher learning rate likely leads to overfitting unconducive to the project’s requirements. The epoch number was 20, the lowest of the three values tested. This results also fits this project’s suggestion that this model and dataset has an early RMSE minimum and further epochs may lead to overtraining. However, given the context of the situation and that Early Stopping has been applied, the epoch number is less relevant as Early Stopping would conclude model training whenever productive reductions in validation loss ceased to appear. It is notable that Early Stopping appeared to intervene in almost every iteration run by Randomsearch, typically stopping model training before epochs trained reached 10, suggesting that this paper might even have benefitted from trialling an even lower epoch number, although the suggestion is moot given Early Stopping is the safeguard for overtraining with this hyperparameter.

The optimal number of hidden layers was found to be only one, in a range of one to ten tested. There appeared to be a largely negative linear relationship between hidden layer number and RMSE. This paper suggests that this may be because adding additional layers may provide too much capacity for learning where it is not required, overfitting the data and reducing RMSE. This, alongside the low learning rate, suggests this model adapts very quickly to the dataset and therefore reaches an early RMSE minimum. However, this paper notes again that neural networks like MLP are black boxes and therefore this is difficult to categorically prove.

Given that momentum was found to be 0.94, high within the optional range of 0-1, this paper suggests this model has a moderately high requirement for momentum, perhaps balancing the extremely low learning rate from earlier. A higher momentum allows the model to bypass local minima to better obtain the global minimum, and this momentum is sufficiently high to do so while also sustaining sufficient learning. The optimal activation function of ReLU fit this project’s prior expectations regarding activation function performance. ReLU is a very popular modern function which suits this paper’s model and business problem. The vanishing gradients problem is less of a concern for this project’s model as the optimal number of hidden layers found was only one, suggesting that any problems relating to a deep neural network’s potential difficulties with backpropagation is unlikely to apply to this project. Furthermore, as mentioned earlier, ReLU trains up to six times faster than Tanh (Krizhevsky, et al., 2012), which is a concern for this project. Due to the volume of data which a year’s worth of tweets relating to Apple contains, computational time and effort is a notable concern. Furthermore, looking ahead to the business use case, the model in question may require retraining on a fairly common basis to ensure the neural network has the most relevant and recent data to understand and build into predictions. This is even more relevant for the stock market field where the circumstances are continually changing and evolving, increasing the rate at which a model will likely lose relevance and performance.

## Dataset Performance Comparison

A point this paper found curious was that the RMSE score for the holdout set generally notably lower than the training and test sets for all of the days ahead models produced. The reasoning for this is difficult to deduce, but can only be suggested as that the insights produced from the training dataset better fits the holdout set than it does the test set itself. What this might translate as is that the holdout set covers a specific period in which the market context may be moving in a particular direction. With the usage of contextual knowledge, this paper suggests this period was generally positive for stock price movements and this may have had a positive impact on the resultant RMSE scores.

The exception to this was Korean, which was found to produce worse RMSE scores on the holdout set compared to the train and test set. This phenomenon is slightly more in line with what this paper may have expected, and suggests on this particular language subset, the models overfit the dataset and produced comparatively worse results on the holdout set.

## Model Comparison Discussion

A point this paper found curious was the relative lack of competitive RMSE performance from the MLP model. This paper did expect MLP to have a stronger performance than the other, simpler models but at times the opposite appeared to be the case. While Ridge and Lasso did generally perform better than MLP, Elastic Net provided the highest performance. Elastic Net was the most complex of the other models, utilising both L1-ratio and alpha hyperparameters, allowing the strengths of Lasso and Ridge to be combined. This suggests that while MLP is more complex, or rather has the potential to be more complex with neuron number, hidden layer number and several other optimisable hyperparameters, Elastic Net provides the optimal capacity to learn, fit, and predict on the holdout dataset. Furthermore, Elastic Net performed better particularly on the Japanese tweet subset, suggesting that despite the strengths that is intrinsic to including all languages within the dataset, that is to say the increased volume of data, Elastic Net was better able to utilise outputs from TF-IDF when only analysing a corpus containing Japanese tweets. However, it is also worth noting that Elastic Net performed relatively poorly on the training and test datasets with regards to RMSE, which may raise some concerns over the context of the holdout set. Given that the training and test sets are from a wide range of dates (one year) as opposed to the two months covered by the holdout set, it could be suggested that the holdout set does not cover as wide enough range of dates for equal consideration. However, due to the constantly evolving context surrounding stock prices, this paper suggests that equal consideration such as is suggested above is not realistic and the holdout set examination approach performed in this paper is scientifically founded; the holdout set for many papers will fit this project’s profile.

As partially expected from literature, the results from the multiple regression model were relatively poor compared to the other models trialled, particularly on the language subsets, suggesting multiple regression benefits significantly from an increased volume of data, despite language differences. Given the context of the alternative models trialled by this paper, this performance reinforces the suggestion that Oliveira, Cortez, and Areal may not have utilised the optimal model when they determined that sentiment is not a strong predictor for stock market returns. However, it is important to note that Oliveira, Cortez, and Areal’s dataset was sourced from Stocktwits, whereas this paper’s was sourced from Twitter. While these sites are both public social media sites, there are some key differences between them which may have also influenced these results. One possible disadvantage of a model trained in Stocktwits data is the volume of users. As discussed earlier, Twitter has 206.67 times the number of monthly active users of Stocktwits at 310 million, compared to 1.5 million respectively. The larger volume of monthly active users on Twitter may represent a more representative proportion of the population who invest or otherwise influence stock market prices. However, it could be suggested that that despite the smaller population utilising Stocktwits, given its specialisation, users of Stocktwits could said to be likely more interested or knowledgeable with regards to stock price movement or generic company information. This could potentially translate to more relevant textual information which may otherwise be more strongly correlated with stock price movements. Alternatively, this paper also notes that it only focused on the Apple stock price, whereas Oliveira, Cortez, and Areal considered six major stocks, where it may be that Apple’s stock price may be more easily modelled than other stocks.

It is also relevant to consider Oliveira, Cortez, and Areal’s later paper on a similar topic but on Twitter instead, they also found similar findings that there is little evidence from their sentiment indicators that they are linked to stock market returns (Oliveira, et al., 2015). However, this paper would suggest several potential areas where disparities between this paper and theirs may have occurred. Firstly, Oliveira, Cortez, and Areal used seven lexical resources to assess tweet sentiment, including SentiWordNet and Harvard General Inquirer, whereas this paper elected to use only TF-IDF. It could be potentially suggested that their usage of multiple lexical resources together may provide an edge over TF-IDF, but further examination would be required for a meaningful conclusion on this point. Furthermore, tools like SentiWordNet is an opinion lexicon which associates a directory of words with either positive, neutral, or negative sentiment. Not only does this approach lend itself specifically to classification problems as opposed to regression, but this paper believes this approach does not sufficiently account for phrases, or combinations of words which, together, form a different meaning to that of its individual parts.

Length of the period examined also differed, with Oliveira, Cortez, and Areal considering only 32 days whereas this paper considered a full year, where it could be suggested that the additional length of time may have improved the model’s overall performance. Lastly, Oliveira, Cortez, and Areal elected to use a multiple regression model, the same approach as their last paper on the subject. This paper suggests that usage of a model with more depth like MLP or Elastic Net may have provided different insights. Additionally, it is notable that this paper trialled Ridge and Lasso Regression as well, and found a generally stronger RMSE for all these models over multiple regression, suggesting that multiple regression may not be the optimal model to select for this particular subject. Overall, this paper does suggest that further research utilises a review of a number of days ahead perhaps using models other than multiple regression, as predictions based on tweets for the same day were typically far poorer than looking further ahead.

## Examining the Days Ahead Evaluation

As mentioned earlier, Coyne, Madiraju, and Coelho previously reviewed how there may be a delay between the impact of social media sentiment and stock price movement. This was also examined by Oh and Sheng, who suggest that this delay may, in part, be caused by investors holding onto prior beliefs for a period before adjusting to new news and the consequences of such. This paper also found that modelling for a number of days ahead yielded promising results. As mentioned earlier in the results section, the optimal period was found to be two days ahead. However, it is notable that the models’ RMSE performances were superior on all days ahead tested, when compared to on the day, with the same day performing the worst for each model on the holdout set. This paper suggests this is likely also because on the day, any news or sentiment conveyed has not necessarily had sufficient time to diffuse within the public and thereby impact the stock price.

Chart, line chart

Description automatically generated

Figure 2: Graph showing the RMSE results of models predicting a number of days ahead

Pinch of salt for this stuff, not sure if prediction stuff becomes much harder when its looking at data from days it’s never seen before

## Profit Simulation Function: Additional Validation Testing

Considering this project’s business use case, this project elected to conduct a further small experiment with the optimum model and the holdout set to trial on a basic level what level of profit would this model potentially return if utilised. This was calculated by creating a function which would intake the model’s predictions, the real results, and the relevant dates for each. An important note to bear in mind is that in its current state, the model examines each tweet individually after training and provides a prediction for each tweet, not each day. Therefore, this function would iterate initially through a list of the unique dates contained within the test set, and for each unique date, it would iterate through the test set, predictions, and dates contained iteratively as a tuple using the zip() function. It would then check each tuple’s date for whether it matched this unique date, and then would note and sum the relevant prediction until a summation of all the predictions for a particular unique date had been reached. This was then saved as this day’s cumulative suggested investment, and would then by multiplied by the real value of the day, noting that this real value is the overall change between opening and closing of that day. This returns the value change of the day, which is then added to a profit metric. Further to this, if the summation of the day’s predictions matched the same sign as the real change, this was saved and accumulated across all the unique dates within the tested dataset. From the above function, this paper is able to deduce the estimated profit, sum invested, and profitability.

Text

Description automatically generated

Figure 3: Profit Simulation Function

Interestingly, the most optimum model of Elastic Net trained on Japanese tweets according to RMSE score actually returns a significant loss on the holdout set, with a profit margin of -44.1% over two months. In terms of the real business case of this project, this is an extremely poor result. This paper suggests that this may occur as RMSE only measures essentially the absolute error between each prediction and true value, and does not explicitly review profit or loss with regards to the context. Therefore, a low RMSE score may be returned even though the values may be regularly opposites with signage. Returning 0.5 instead of the real -0.5 for example would return a lower RMSE score than returning 3.5 instead of the real 0.5, while only the second option is profitable. This implies that RMSE is not the only consideration which is useful for this project and context. Furthermore, the test set returns a profit of 22.5% over a year. This suggests either extreme overfitting to the date range trained, or alternatively suggests that the holdout date period tested of two months may not be a representative comparison period with the test set which tests across the full 12 months of the initial dataset. Since the overall gain from the test set was 22.5%, it is possible there were periods where the model did not perform as expected, and then overperformed in other areas. This would follow with the contextual understanding of how tweets drive the model, as on certain days of high tweet volume and strong sentiment, the model may perform at a stronger level.

Furthermore, there are an array of caveats to these numbers. Firstly, one must consider how stock prices vary almost constantly when the market is open due to trading volume, particularly for highly traded stocks like Apple. Therefore, the close price is not necessarily the price an investor may be able to acquire a stock at, and there are further complexities within this area of implementation which may reduce or otherwise impact profit margin. Another issue not included in this profit simulation is spread, the margin a stockbroker charges on a trade. This would reduce the realised profit for each day, reducing profit margin, although this would depend on each broker and their specific rates. Furthermore, there is also the consideration of the context of the market. The stock market undergoes periods of overall rising or falling in terms of stock prices, known as bull or bear markets respectively. These markets may occur in a manner contrary to their context, for example the bull market from approximately March 2020 onwards when COVID-19 was appearing to have an adverse impact on many businesses, many encountered generally rising stock prices. Therefore, this paper notes that it is possible the results over the two months the holdout set covers may have been impacted by the overall passage of the market. From the graph below,

From all the models trialled by this paper, the RMSE-optimal iterations were taken and applied to the profit simulation function as additional validation and found that while MLP, Multiple Regression, and Lasso returned a positive profit margin for the test set, all models returned a negative profit margin for the holdout set. This suggests that there is possibly something inherent with the holdout set which means the models trialled are poor at returning profit for, possibly as part of the reasons discussed above, like overfitting to the training/test date range, or holdout date length.

One additional interesting point to add is that a more realistic estimation of profit would be to retrain the model for each new day of tweet data, so the model has the most up-to-date data. Currently, there may be topics and concepts impacting the Apple stock price in, for example, July 2021, which were not present in the training data period from May 2020-2021. Predicting the stock price under these circumstances may have returned sub-optimal results when compared to the actual use-case. However, as one can see from the below graph which illustrates the RMSE by day for the holdout set, the RMSE does not appear to increase in any meaningful pattern over the date range as one might expect if the above posited phenomenon did exist as suggested.

Chart, line chart

Description automatically generated

Figure 4: Graph Illustrated RMSE by Day in Holdout Set for Optimum Elastic Net Model

## Bullish and Bearish Sentiment Prediction Exploration

An interesting point this paper reviewed was the final evaluation, where, among the standard metrics reviewed, a conversion into a classification measure of bullish sentiment against bearish sentiment was performed. This was mainly an academic exercise to not only relate slightly more with the classification papers reviewed in literature, but also to review how well would this model perform at simply identifying whether the stock price would rise or fall. This was primarily applied to the optimum model as suggested by the lowest RMSE. Please note that for the following, the test set will be discussed, and then the holdout set.

Accuracy of a form can be determined using the sign matching criteria subsumed within the function above, and was interestingly noted at only 48.5% for the test set. An investment system based on this would return very poor results, and could be said to be worse than simply guessing. While this paper can infer from this that this model is frankly rather poor at determining simply whether sentiment is bullish or bearish, but it is more capable of understanding when the price is likely to rise or fall a great degree, and therefore capitalises by investing more in the appropriate direction, leading to a potential profit margin of 22% for the test set.

This paper suggests that this may be due to the sentiment analysis carried out, which may have clearer and more appropriate weights for stronger language which more clearly denotes sentiment, which may also have a similarly strong impact on the stock price. Similarly, when sentiment is relatively neutral and the stock price only moves only slightly up or down, it could be suggested that it is difficult from a core perspective for a sentiment analysis to accurately pick up on whether the increase will likely be positive or negative. As a result, it could be suggested that a more cogent investing strategy based on this model would perhaps not make a trade on days when the prediction is below a given threshold, so as to potentially reduce risk.

With regards to this project’s driving business problem, understanding when the price is likely to move sharply in either direction is very important and can be profited from. This also has noteworthy implications for further research into classification against regression tasks in this field, as will be discussed later, specifically about how TF-IDF performs within the scope of a regression problem with range to explore and exploit its capability to generate strong predictions across a range of possible results, maximising potential profit. Furthermore, this also holds implications for further classification-based methods in this field, potentially suggesting that classification methods which utilise a sentiment analysis tool like TF-IDF may benefit from not simply utilising a binary classification format with bullish and bearish sentiment, but also have a third category for neutral sentiment which encapsulates predictions for days when the stock price movement is relatively neutral and only slightly moves up or down.

However, when considering the holdout set and previously discussed, the accuracy in the format stated was only 44.4%, and the profit margin was -44.1% over two months. This is far worse than expected, and suggests that the above considerations and suggestions are suited only for the test set, and likely do not have the same implications for the holdout dataset.

## Experimentation and Discussion Caveats

It is notable that for the purposes of model optimisation, this paper utilised only a 15% sample of the full dataset so as to ensure the iterative optimisation undertaken did not take an unnecessarily long period of time, with the full dataset utilised after optimisation. However, when this sample was taken, this paper had obtained similar RMSE scores but a far superior profit margin of 57% on the holdout dataset, connoting a profit of 312% over 12 months if considered average. As discussed previously, once the full dataset was taken, the model’s profit performance became far poorer. This project underwent rigorous checking and self-validation and could find no procedural issues with how the experiment was carried out, and therefore concludes that the initial sample taken, using Sklearn’s ‘train\_test\_split’ function, contained tweets which were both less representative of the tweet population that may have been expected, and also more representative of the Apple stock price movements than the full tweet population.

# Evaluation, Reflections, and Conclusions

This final section will review the lessons learnt from this project’s work and suggestions for future research in the field to potentially build on findings from this paper. While these two categories have distinct similarities, they have been separated into subsections, although the lessons learnt will likely also have implications on future work in the field.

## Lessons Learnt

A lesson learnt for this project would be a better understanding of the context behind selecting regression or classification models. While regression was likely more fulfilling of the core business problem, it did make comparison to existing contemporary literature more difficult through a lack of shared performance metrics. Furthermore, while it is possible to use classification as this paper did to briefly explore bullish against bearing sentiment, it is not optimal to simply use a different model. As explored in some of the reviewed literature papers, there are tools for better understanding textual information which are far more suited to classification tasks than regression. For example, SentiWordNet is highly optimal for classification tasks and therefore was not utilised for this project, but may have rendered more optimal results if utilised for this project’s exploration into bullish and bearish sentiment, as opposed to TF-IDF.

## Future Research

### Data Source Evaluation

Future research may consider the sources both Stocktwits and Twitter to provide a fair comparative analysis of the predictive powers of both data sources given the advantages and disadvantages of both; Stocktwits is suggested to have the more overall informed userbase whereas Twitter may have a wider and therefore more representative userbase of the population investing in the stock market. Furthermore, future research may investigate replicating this project’s models and motivations with a larger number of companies and stocks than this project’s scope allowed. The purpose of this is to review how predictive performance may vary between companies, as one notable weakness of this project is that only one company is reviewed. While Apple is extremely highly valued as previously argued, from a business perspective, investing in one company is not feasible for a responsible investment portfolio and therefore reviewing additional companies may provide further predictive performance with regards to the stock market as a whole.

### Regression Compared with Classification

As mentioned earlier, future research may be further informed by this research’s work into classification against regression tasks in this field. While dependant on the field and overriding goal of the project, if the project’s aim is business dependent in a similar manner to this project, regression will likely provide highly interesting avenues of research, particularly considering the capabilities of TF-IDF and sentiment analysis conducted within the scope of regression problem, with the potential identification and exploitation of stronger language denoting importance and strong sentiment. Furthermore, if research is determined to use classification, this paper suggests trialling the inclusion of a third, neutral class for days of only minor movement. This would potentially not only increase accuracy, but also improve any subsequent business use-case, as a bullish prediction for example could relate to a meaningful movement of $4, or a small movement of $0.3.

### Performance Metrics

Another consideration for future research may be to look further into the potential performance metrics researchers may utilise to evaluate their models and their work. This paper exclusively utilised RMSE to evaluate performance, but as the profit simulation function created and utilised by this project illustrated, a lower RMSE score does not necessarily denote a higher performing model, given the context of stock market predictions. While error is useful, it does not fully encapsule the context where a small positive real price movement and a corresponding small negative prediction may have a low RMSE, but results in a tangible loss. This was a significant lesson learnt for this project, and therefore future work may consider using a similar profit simulation function, perhaps including things like average spread and so on, to create a superior illustration of model performances and results.

Overall, a significant avenue for future research to explore is to better understand the phenomenon this paper noted where the holdout dataset returned a lower RMSE score but performed far worse on the profit simulation function when compared to the test set. This paper has suggested reasons for why this may have occurred, with the limitations of RMSE as a performance metric discussed, as well as the differing timeframes, where the holdout set is a completely new set of data covering topics which may not have been present in the training dataset, whereas the test set covers the same date range and therefore may have been in a stronger position to make better predictions. However, future research can explore this in greater detail, with a view to further improve the implementation of this area’s body of work on new and current data.

### Implications from Contextual Market Conditions

As mentioned earlier, this paper was able to obtain clear improvements and comparisons between the RMSE scores of different regression models, but received very poor profitability on the holdout dataset. This paper previously noted that one potential factor may be that the contextual market conditions may differ between the holdout date range and the training date range. For example, the training date range may have contained particularly negative results from an overall bearish market, which then may have difficulty in predicting prices from a bullish market which may tend to more positive results. Therefore, future research may look to review how much data their models are trained on, and perhaps matching this to the contextual market conditions. For example, a model trained on a date range containing what is relatively exclusively a bullish market is more likely to be a better predictor for times of bullish sentiment as opposed to bearish. As one may see from the graphs below, the predictions for the test and holdout sets were on average, far lower than the real values, which could be suggested to be the result of a holdout date range which is more bullish than the overall data which the model has been trained on. The prediction values appear to generally be centred around -2.2 for the test set and -1 for the holdout set, which could be said to be bearish, whereas if the predictions were weighted to be higher to represent a more bullish or neutral period, the model’s profit margin would notably increase.

Chart, scatter chart

Description automatically generated

Figure 5: Graph Illustrating Real against Model Predictions for the Elastic Net Model Predicting Two Days Ahead on Test Set

Chart, line chart

Description automatically generated

Figure 6: Graph Illustrating Real against Model Predictions for the Elastic Net Model Predicting Two Days Ahead on Holdout Set

Future work may potentially choose to segment their data and create multiple models based upon market conditions, and apply them in various combinations to fit the data to return the optimal predictions.

### Tweet Quality and Sampling

As discussed earlier, this paper sampled 15% of the full dataset to optimise the models, and found a far superior profit margin resulting from models trained on only this dataset sample as opposed to the full dataset. Therefore, it is likely that some tweets like those contained within the sample taken are more representative of the Apple stock price movement than others. Consequently, this paper suggests future research may consider a method of evaluating tweets or perhaps tweet authors to selectively sample the full tweet dataset and better improve model performance. While the overall results of this paper did not confirm a profitable relationship between tweets and Apple’s stock price, intermediate results within this paper suggested that with a subset of the tweet dataset, a profitable model on holdout data was possible, and future research may be able to scientifically evaluate how to best select that data subset. A part of this research may also include a review of how data from Twitter and Stocktwits may compare in terms of predictive capabilities. These two data sources are overall the field’s focus points of research into public sentiment’s relationship regarding stock price movements, and an argument could be made to state that given one of this paper’s conclusions that the overall population of Twitter tweets was observed to perform poorer than a random subset taken, Stocktwits, a site based wholly around the stock market, may provide more informed and useful data from which to calculate sentiment. Another possibility is that the full dataset was simply too large for the models to optimally predict, and therefore further studies may look to review differing subsets of the overall dataset to better examine this factor.

# References