Proposal

Student
Name: Stefan Winter
Student number: 2067606

Thesis Supervisor
Dr. Peter Hendrix

School of Humanities and Digital Sciences
Tilburg University
October 2021

Project Definition

The internet has enabled humankind to access information, exchange ideas and become part of a community. Of course, that also applies to online message boards. Ever since the internet found mainstream adoption, people joined those message boards to discuss trading the stock market. Most recently the Reddit forum wallstreetbets attracted a lot of interest and now counts over 10 million members. In this subreddit, members talk about various investment ideas. However, most of those ideas are of speculative nature with members trying to get rich quick, usually by using risky derivatives like stock options. While the wallstreetbets community undoubtedly minted many millionaires, there are also numerous people who lost their life savings.

Even though the Reddit subforum was already founded in 2012, it got most of its media attention in 2021 due to a short-squeeze of the GameStop (GME) stock that drove the stock price up several hundred percent. However, it was not the rapid price appreciation that amazed market participants. Instead, it was the unprecedented decentralized and coordinated buying of Gamestop shares by members of the wallstreetbets community that attracted attention. (Anand & Pathak, 2021) Interestingly, the story repeated itself when forum members sent other stocks, such as AMC Entertainment and BlackBerry to the moon.
Organizing the mass-coordinated buying of stock, however, requires that enough participants share the same sentiment. Some research shows that social media sentiment has a particularly strong impact on uninformed traders (Danbolt, Siganos, & Vagenas-Nanos, 2015). Interestingly, finance scholars did not consider Reddit as a platform that could have such a big impact on the financial markets. As a result, the site has mostly been ignored in their research (Long, Lucey, & Yarovaya, 2021).

Hence, this Master thesis will focus on the *"meme stock"* driven investor sentiment of the wallstreetbets subreddit. By performing sentiment analysis on the aforementioned forum, it is assumed to be able to predict daily changes in the stock price of selected securities. Being able to accurately measure the sentiment ensures more efficient, and hence, less volatile markets. Furthermore, this thesis will analyze how to best incorporate domain-specific language, which is commonly used on wallstreetbets, into the sentiment analysis.

This thesis will try to answer the following research questions:

Main Research Question:
Can sentiment analysis of the WallStreetBets Reddit-forum be used to predict daily changes in the stock prices of selected securities?

Sub Research Questions:
1: Which sentiment analysis approach performs best on predefined key performance indicators?

2: How can the domain-specific language of the Reddit forum WallStreetBets best be incorporated into sentiment analysis?

3: Which machine learning algorithm delivers the best predictive performance for changes in daily stock prices of selected securities based on the sentiment analysis performed earlier?

Motivation

Scientific Contribution
Since the wallstreetbets subreddit has become very popular just recently, there is little academic research about the impact of the community on financial markets so far. This thesis not only tries to shine some light on those new and influential market participants, but also tries to put forward some methods that work best to perform sentiment analysis on the forum. The obtained sentiment will then

be used to explore if it is possible to predict daily changes in stock prices. As a result, this thesis hopefully contributes some new insights to the currently very limited research on wallstreetbets-specific sentiment analysis.
Researchers, such as Talamás (2021), specifically propose future work on "inclusion of features derived from alternative manipulation of the data like sentiment analysis could lead to new insights". Even though there is some research about sentiment analysis on wallstreetbets, that research does not use state of the art algorithms to perform sentiment analysis. This thesis will implement a variety of sentiment analysis algorithms, compare their strengths and weaknesses for the task at hand and finally decide which algorithm performs best.
In addition, this thesis will try to shine some light on the impact of domain-specific language which is used on the wallstreetbets subforum. Since there are various methods for handling domain-specific language, this thesis will explore both unsupervised and supervised machine learning algorithms. The results of that exploration will then be juxtaposed and implemented into the sentiment analysis.

Societal Relevance
As mentioned earlier, the "to the moon" movement had a tremendous impact on the lives of individuals, both to the positive and negative. Besides that, however, many investment funds have also been negatively impacted by the recent short-squeezes. While it might seem noble to root for individuals who try to force large funds out of their positions at big losses, it is easy to forget that many of those funds manage money for charitable endowments, pensions and others. Furthermore, such disruptions to the financial markets can harm its stability, thus causing spillover effects which can also negatively impact the lives of many people (Lyócsa, Baumöhl, & Vyrost, 2021).

Background

Gauging sentiment of online forums to predict movements in stock prices has been a research subject for many years now. Das & Chen (2007) did a study on the Yahoo! message board, which was amongst the first ones on the internet for investors to exchange ideas. In their paper, they show that the relationship of stock price to sentiment is significant and that market activity is related to activity of the message boards.

Research Foundation
The impact of the wallstreetbets subreddit set an unforeseeable precedent. It wasn't expected that message boards can have such an enormous influence on certain stocks. Since the wallstreetbets meme-stock movement is a relatively recent phenomenon, there is very little research on the impact of wallstreetbets on individual stocks.
Long, Lucey, & Yarovaya (2021) try to establish a foundation for future research of sentiment analysis derived from Reddit on the stock market in what they believe to be the first paper on that topic. They try to uncover if specific emotions, such as *"Angry, Fear, Happy, Sad and Surprise"*, of comments on Reddit posts impacts intraday returns of a specific stock. They conclude that the impact of tone, as well as the number of comments do have an impact on returns. However, they show that the number of comments are not related to sentiment. Instead, it is the number of comments that is posted within an hour that has the biggest effect on one minute changes in the stock price. Furthermore, the paper shows that the emotions *Sad, Anger* and *Surprise* have a significant impact on the gamestop 1-minute stock price. The *Happy* sentiment does not show a significant impact on 1-minute price changes, however, a causality test showed a link between the *Happy* sentiment and intraday returns of the GME stock. In addition, the paper shows, that sentiment only impacts intraday returns if a thread has more than 2000 comments. Hence, the authors confirm that Reddit sentiment has an impact on the stock market. They also argue that any asset that is targeted by a large crowd from wallstreetbets can become a subject of excessive volatility, without being driven by any fundamental reasons.

Project Structure

The research by Jemai, Hayouni, & Baccar (2021) proposes a system, according to which a sentiment analysis project should be structured. The first phase is the *data collection* phase. In that phase, data is to be obtained from a source. In the second phase, the *preprocessing* phase, the text is cleaned up. As a result, it will be easier to feed the text into a machine learning algorithm. In this phase, several steps are taken. One of the steps is *data tokenization*. This is a popular technique, in which a body of text is broken down into several sentences and each sentence into a list of words. Another step in the preprocessing phase is to delete stop words, such as *is, the, a* and other common words. Furthermore, special characters such as @ and urls should be removed. Additionally, it is proposed to change the text to lowercase. As the final step, they propose *lemmatization*. By doing that, the structure of a word is analyzed and then converted into its normalized form.

The next step is *data preparation*. In that step the data is prepared for sentiment analysis by converting the tokens into a dictionary. The dictionary is then split up into train and test sets.

In the final *classification* phase machine learning algorithms can be used to learn from the training data.

In addition to the proposed steps, the paper also touches on related work by peers. For example, they briefly explain the work carried out by (Parveen & Pandey, 2016). In that work they showed that preprocessing data with emoticons, leads to more accurate results than preprocessing data without emoticons.

Domain Specific Words

Since the wallstreetbets community uses many domain-specific words, those words also need to be accounted for. Since these typically are words with strong positive or negative sentiments and quite often used, it is very important to identify these words' polarity for determining the semantic orientation. One way to handle domain-specific words, is by having a dictionary that is customized for those words. This dictionary can then be searched for finding and scoring the sentiment of the word (Asghar, 2014). Other research deviates from the aforementioned dictionary based approach. Instead, they examine how deep learning methods can be used to automatically detect and identify domain-specific words from sentences. By doing so it is assumed that the algorithm can not only detect whether domain-specific words are used (sentence-level detection), but also to identify the exact position of the term in the sentence (token-level identification). Hence, it is possible to detect new meanings of words in an already existing dictionary. In addition, this approach also allows to classify newly created words, that do not yet exist in a dictionary. This can be achieved by having models that formulate domain-specific word detection as a sequence-labelling task. It is shown in experiments that the flexibility of a part of speech feature performs best in detecting domain-specific words. That is because domain-specific words often entail a structured part of speech transformation of existing syntactic uses of words. Novel domain-specific tokens can be learnt by understanding the contextual structure within a sentence. Those out-of-vocabulary tokens can be learnt in the hidden layers of LSTMs (Hochreiter & Schmidhuber, 1997). The model can be improved, by applying a character-based convolutional neural network to encode the spelling of words (Pei, Sun, & Xu, 2019).

Gupta, et al. (2019) introduce SLANGZY, an algorithm that uses a mathematical "slang" factor to better judge social-media word definitions found in the Urban Dictionary, which is the largest crowd-sources slang dictionary available on the internet. The research shows that SLANGZY succeeds in normalizing the unstructured meanings of internet jargon in the Urban Dictionary. Hence, the algorithm can be provide more accurate meanings of non-standard words.

Dataset

This Master thesis relies on two data sources. First, to perform sentiment analysis, posts from wallstreetbets need to be obtained. Second, in order to predict the stock prices of selected securities, the stock prices need to be accessed. Both datasets will be explained in the following section.

Reddit Data

While Reddit does offer an official API, the API is most useful for streaming data. There are some strict limitations on loading historical data. As a result, the official API is not the best choice for this thesis. However, pushshift.io provides a solution for the strict limits. Pushshift is maintained by the /r/datasets mod team. The FAQ on the pushshift subreddit states, that pushshift data is best used to:

- Analyze large quantities of Reddit data
- Grab data for a specific date range in the past
- Search for comments
- Aggregate data

Pushshift copies data from Reddit at the time it is posted. Since Pushshift uses the document-based database Elastic, it is extremely fast to query data. However, currently Pushshift does not regularly update certain metadata, such as scores, edits to a submission's text or comments. Hence, there might be some minor inconsistencies of what is shown on Reddit and what is in the database. The scores, for example can easily be accessed via the official reddit API and, if needed, joined with the data obtained from Pushshift. Based on the data verification I performed, the number of comments only deviates by a marginally small amount. It is hypothesized that the small difference can be explained by forum moderators deleting spam. Those spam comments are assumed to not have a big impact the thesis anyways, which is why the small difference in the number of comments do not need to be addressed.

To access the Pushshift API, I used an API wrapper called PMAW. Since requests are I/O-bound, PMAW is multithreaded. Hence requests can be run asynchronously which allows the data to be loaded much faster.

When making the API request, the most important parameters are the following:

- subreddit: Name of the subreddit
- q: The search term based on which the subreddit is queried
- before: The starting date of the query
- after: The end date of the query

The query returns 89 columns. Most of which, however, can be dropped since they either aren't useful or contain no data. The most important columns are the number of comments, the title of the post and the content of the post. Emoticons are also included in the content text.

Stock Prices

The stock market data is obtained from yahoo finance, using a package called yfinance. The data can be downloaded by providing the ticker symbol, the start date and the end date for the query. The query returns the Open, High, Low, Close, Adjusted Close prices as well as the trading volume for every trading day. For weekends, as well as public holidays, no data is returned.

Algorithms & Software

All algorithms will be implemented in *python*. To do so, I will import several modules including numpy, pandas, keras, sklearn, sktime, matplotlib, nltk, transformers and pmaw. Furthermore, the thesis will be written using Latex. Visual Studio Code will be used as the IDE for the python code and as the editor for Latex. Furthermore, all code, papers and other files that are less than 100 mb are maintained in a public git repository (https://github.com/StefanWinterToo/Master-Thesis).

The next section will briefly explain which algorithms will be used for sentiment analysis and for the prediction of daily changes in stock prices.

Sentiment Analysis

For the sentiment analysis task the following three machine learning methods will be used:

Naïve Bayes (NB): NB is a probabilistic supervised machine learning algorithm. The relatively simple algorithm works probabilistic, meaning that it assigns the probability of belonging to a class based on given features (Jemai, Hayouni, & Baccar, 2021).
Since text naturally has many dimensionalities, which can be handled very well by NB, this algorithm established itself as one of the standards for sentiment analysis. In this thesis Multinomial Naïve Bayes will be used for text classification. That is due to the strength of the model to handle larger vocabulary sizes. Furthermore, the algorithm is relatively easy to implement, can be used for real-time applications and is highly scalable. The downside, however, is that the prediction accuracy of the algorithm oftentimes is lower than other sentiment analysis techniques (Song, Kim, Lee, Kim, & Youn, 2017).

Support Vector Machines (SVM): SVMs can be used for both regression and classification problems. Classification is done by finding a hyper-plane with the biggest margin, meaning it looks for the greatest distance to the nearest sample points (Jemai, Hayouni, & Baccar, 2021). SVMs fit the hyper-plane by using spatial transformations, also known as kernel functions. Kernels can be linear, RBF or others. The radial basis function (RBF) kernel is best used for non-linear problems and is a general-purpose kernel that is often used in pattern recognition problems. The linear kernel, on the other hand, is typically used when there are only two classes present. A good example for that might be positive and negative sentiment (Alves, Baptista, Firmino, de Oliveira, & de Paiva, 2014).

Long Short Term Memory (LSTM): LSTMs are becoming increasingly popular for sentiment classification. LSTMs are built on a recurrent neural network architecture (RNN). In an RNN the neurons are connected to themselves through time. As a result, the input from a time instance t_i will also be used as an input for the next time instance t_{i+1}.That leads to the problem of vanishing gradients. LSTMS are designed to overcome that problem.
The LSTM architecture does so via its four constituents: A memory cell which can remember a lot of information from previous states, an input gate which controls the inputs into the neurons, an output gate with an activation function and lastly a forget gate which resets the neuron (Priyantina & Sarno, 2019).

Bidirectional Encoder Representations from Transformers (BERT): BERT is a relatively new machine learning algorithm developed by Google in 2018 and mainly designed for natural language processing. BERT is pretrained on the English Wikipedia and BooksCorpus. Because of the pretraining users won't need as much computing power to achieve good results, even if the dataset is relatively small (Devlin, Chang, Lee, & Toutanova, 2019). The BERT github page even states that "Most NLP researchers will never need to pre-train their own model from scratch" (google-research, 2020).

Predict Stock Prices

ARIMA established itself as a standard when it comes to time-series forecasting. It is capable of capturing various temporal structures in time-series data. An AR-I-MA model consists of three parts. The Auto Regressive (AR) part means that the model looks at the dependent relationship of an observation and some pre-defined observations that have a time lag. The Integrated (I) part makes the time-series stationary, which is an essential part in time-series analysis. The Moving Average (MA) analyses the relationship between the observations and the residual errors (Brownlee, 2020).

Multiple Linear Regression will be used to predict stock prices, because it is able to determine the linear relationship between a dependent and n independent variables. Linear Regression uses an Intercept β_0 and a slope β_i for each independent variable x_i to predict the dependent variable y.

Regression analysis typically involves numerical input data which may consist of a wide range of values (Siew & Nordin, 2012).

LSTMs will also be used to predict daily changes in stock prices. That is because of its strengths in analyzing connections among time-series data by using LSTM's memory function. Other feed-forwards neural networks, as a comparison, cannot handle the complex time correlation between information. Furthermore, there is a lot of literature that proves the suitability of LSTM for time-series analysis (Jin, Yang, & Liu, 2020).

Evaluation Methods

Ground Truth

Since there is no ground truth dataset available for wallstreetbets and manual annotation is too time consuming, I will derive the labels from SentiWordNet and VADER (Valence Aware Dictionary and sEntiment Reasoner). SentiWordNet contains the associated sentiment for words. VADER is a sentiment analysis tool that is specifically designed for social media. The ground truth will then be derived by taking the intersection of the two methods.

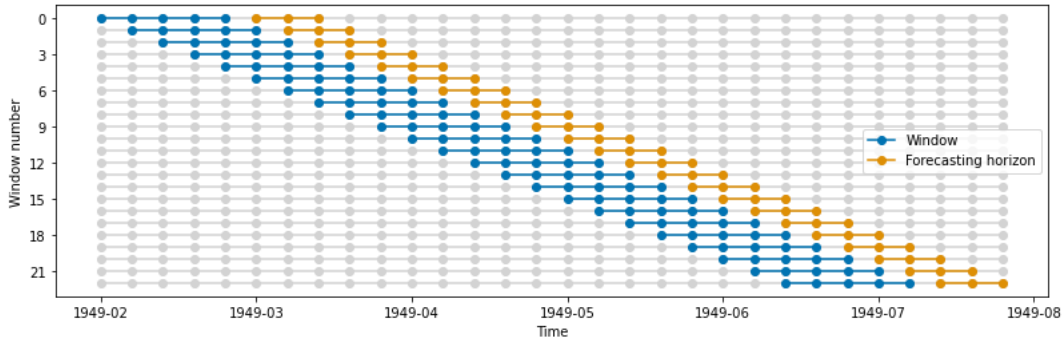
Sentiment Analysis

Typically, accuracy, precision, recall and the F-score are used as evaluation metrics to assess the performance of a sentiment analysis model.
Accuracy is the percentage of correctly predicted observations over all observations. However, accuracy should only be used if the classes in the data are balanced.
Precision expresses the proportion of how many classes were classified as positive, that actually are positive.
Recall refers to the percentage of total relevant results that were correctly classified. Hence, it is a good metric to see if the model was able to find all relevant instances in a dataset.
F-score is a metric that combines precision and recall (Garcia, 2020).

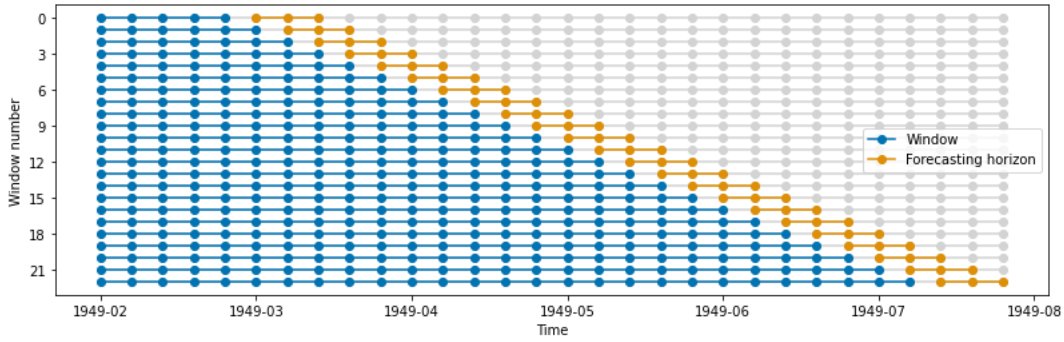
Due to simplicity of Naïve Bayes, it will be used as the baseline.

Predict Stock Prices

The data-source for this task is the aforementioned data obtained from yfinance. However, since stock prices are time-series data one needs to be careful on how to split the dataset. If, for example, a standard k-fold cross validation method is used to split the data, the window at t_0 will be used to forecast. However, that window occurs before the training data. For other windows on the other hand, the model would be able to peek ahead into the future.
Scikit-learn provides window-splitters that work similar to cross validation, but account for the influence of time. Two examples are the SlidingWindowSplitter and the ExpandingWindowSplitter. In the SlidingWindowSplitter the training window includes a window of predefined size *n* and the forecasting horizon is a size after the training window.



The ExpandingWindowSplitter generates folds across a sliding window over time. The size of the training set, however, grows over time. As a result, each subsequent fold retains the full history up to the current point. The size of the testing set remains constant (Amidon, 2021).



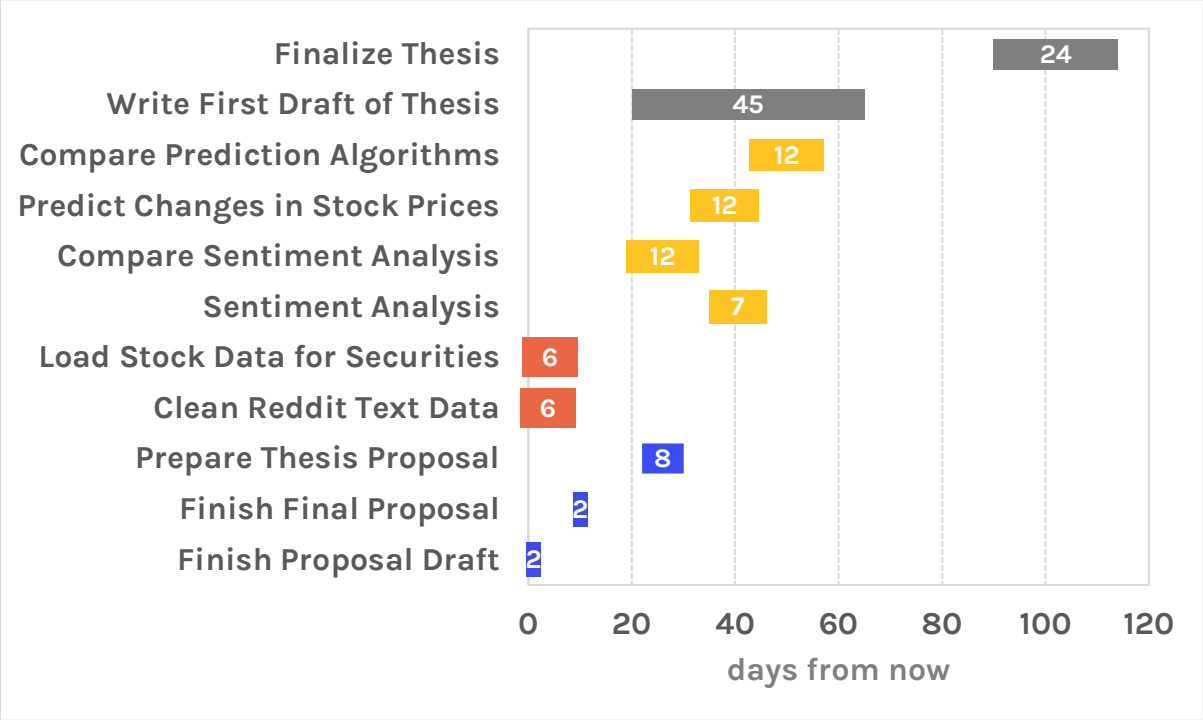
Because predicting time-series data is a regression problem the evaluation methods mean absolute error (MAE) and root mean square error (RMSE) will be used (Jin, Yang, & Liu, 2020).

Since the LSTM is far more complex, the results of the linear regression will be used as the baseline.

Milestones and Plan

The thesis can be completed in time if the following schedule is adhered to.
Note: The scaling of the bars is not 100% accurate.

Task	Category	Days From Now	Days To Complete Milestone	Start Date	End Date
Finish Proposal Draft	Proposal	0	2	29.09.21	01.10.21
Finish Final Proposal	Proposal	9	2	08.10.21	10.10.21
Prepare Thesis Proposal Presentation	Proposal	22	8	21.10.21	29.10.21
Clean Reddit Text Data	Data	0	6	29.09.21	05.10.21
Load Stock Data for Securities	Data	0	6	29.09.21	05.10.21
Sentiment Analysis	Algorithm & Evaluation	37	7	05.11.21	12.11.21
Compare Sentiment Analysis Algorithms	Algorithm & Evaluation	20	12	19.10.21	31.10.21
Predict Changes in Stock Prices	Algorithm & Evaluation	32	12	31.10.21	12.11.21
Compare Prediction Algorithms	Algorithm & Evaluation	44	12	12.11.21	24.11.21
Write First Draft of Thesis	Writing	20	45	19.10.21	03.12.21
Finalize Thesis	Writing	90	24	28.12.21	21.01.22



References

Anand, A., & Pathak, J. (2021, June). WallStreetBets Against Wall Street: The Role of Reddit in the GameStop Short Squeeze. *Indian Institute of Management Bangalore Research Paper Series*.

Danbolt, J., Siganos, A., & Vagenas-Nanos, E. (2015). Investor sentiment and bidder announcement abnormal returns. *Journal of Corporate Finance*, 164-179.

Lyócsa, Š., Baumöhl, E., & Vyrost, T. (2021). YOLO trading: Riding with the herd during the GameStop episode. *Finance Research Letters*.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 1375-1388.

Jemai, F., Hayouni, M., & Baccar, S. (2021). Sentiment Analysis Using Machine Learning Algorithms. *International Wireless Communications and Mobile Computing*, 775-779.

Parveen, H., & Pandey, S. (2016). Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm. *2nd International Conference on Applied and Theoretical Computing and Communication Technology* (pp. 416-419). Bangalore: IEEE.

Asghar, M. Z. (2014). Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet. *Life Science Journal*, 66-72.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 1735-1780.

Pei, Z., Sun, Z., & Xu, Y. (2019). Slang detection and identification. *Proceedings of the 23rd Conference on Computational Natural Language Learning* (pp. 881-889). Hong Kong: Association for Computational Linguistics.

Gupta, A., Teneja, S. B., Malik, G., Vij, S., Tayal, D. K., & Jain, A. (2019). SLANGZY: a fuzzy logic-based algorithm for English slang meaning selection. *Progress in Artificial Intelligence*, 111-121.

Song, J., Kim, K. T., Lee, B., Kim, S., & Youn, H. Y. (2017). A novel classification approach based on Naïve Bayes for Twitter sentiment analysis. *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL.*, 2996-3011.

Alves, A. L., Baptista, C. d., Firmino, A. A., de Oliveira, M. G., & de Paiva, A. C. (2014). A Comparison of SVM Versus Naive-Bayes Techniques for Sentiment Analysis in Tweets: A Case Study with the 2013 FIFA Confederations Cup. *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, 123-130.

Priyantina, R. A., & Sarno, R. (2019). Sentiment Analysis of Hotel Reviews Using Latent Dirichlet Allocation, Semantic Similarity and LSTM. *International Journal of Intelligent Engineering and Systems*, 142-155.

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, 9713-9729.

Garcia, E. (2020, Feb. 21). *how-is-sentiment-analysis-used-in-the-real-world*. Retrieved from super.ai: https://super.ai/blog/how-is-sentiment-analysis-used-in-the-real-world

Siew, H. L., & Nordin, J. (2012). Regression techniques for the prediction of stock price trend. *International Conference on Statistics in Science, Business and Engineering (ICSSBE)* (pp. 1-5). IEEE Xplore.

Long, C., Lucey, B. M., & Yarovaya, L. (2021). 'I Just Like the Stock' versus 'Fear and Loathing on Main Street' : The Role of Reddit Sentiment in the GameStop Short Squeeze. *SSRN Electronic Journal*.

Amidon, A. (2021, July). *Don't Use K-fold Validation for Time Series Forecasting*. Retrieved from towardsdatascience: https://towardsdatascience.com/dont-use-k-fold-validation-for-time-series-forecasting-30b724aaea64

Talamás, J. A. (2021). Social media Effects on the market: Reddit Data analysis on Stocks. *10.13140/RG.2.2.24180.88960*.

google-research. (2020, March 11). *bert*. Retrieved from Github: https://github.com/google-research/bert

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Brownlee, J. (2020, December 10). *How to Create an ARIMA Model for Time Series Forecasting in Python*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/