NLP-Technical Report

Building an Application to give Financial Trading Advice based on "r/Wallstreetbets."

Group H

May 2021

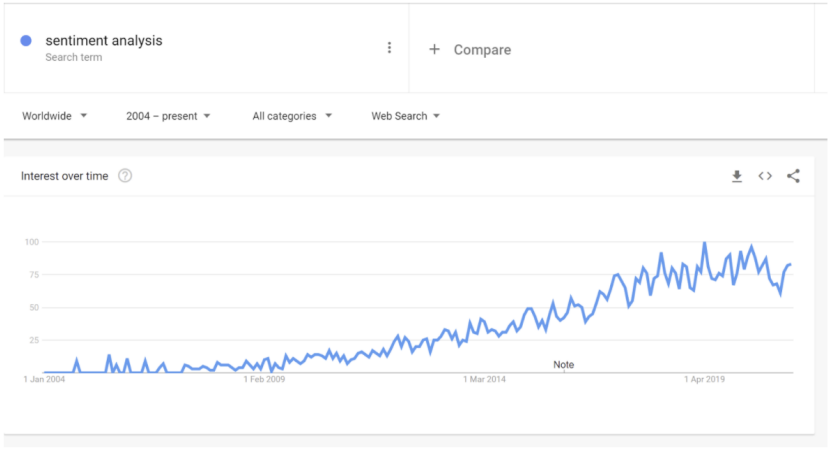
This report is divided into three main sections. The first section gives a review of sentiment analysis, elaborates on the thread "Wallstreetbets" on the social media platform Reddit and the FinBert model. The second section presents our application which uses Reddit posts regarding stocks as input data and advice based on sentiment analysis whether to buy or sell stocks. The last section presents the results of our application.

1 Literature Review-Sentimient Analysis

Our opinion and our view of the world are a central part of our everyday life. People are in constant exchange with each other about their opinions regarding specific products, news, and trends. Through this interactive exchange, we influence each other and frequently change our views (Liu, 2012). In the age of Big Data, purchasing decisions are increasingly influenced by the sentiments of others that can be found online. The opinions of others on the internet are mainly unfiltered and are therefore seen as a reliable, unbiased source of information. With the surge in data created on the internet, interest in mining and studying user sentiment has also increased (Hajiali 2019). Text classification and analytics include sentiment analysis, which senses the subjectivity or polarity of texts. Modern sentiment analysis only became popular in the mid-2000s and is primarily focused on online product reviews. Since then, sentiment analysis has been applied to various other fields, including consumer segmentation, campaign scheduling, and financial market forecasting (Mäntylä et al. 2018; Hajiali 2019). Fig.1 shows the increase in popularity of sentiment analysis as a search term on Google since 2004. Mäntylä et al. show in their paper how modern sentiment analysis has received a 50-fold growth within only ten years between 2005 and 2016 (Mäntylä et al. 2018).

In their paper, Vinodhini and Chandrasekaran describe sentiment analysis as a form of natural language processing used to monitor public opinion on a specific product or subject. Sentiment analysis, also known as opinion mining, entails creating a framework to capture and analyze product opinions expressed

Figure 1: "Google Trends (www.google.com/trends) data showing the relative popularity of search string "sentiment analysis"



in blog posts, comments, reviews, or tweets (Vinodhini and Chandrasekaran 2015). There are several challenges present in the field of sentiment analysis. The majority of conventional text processing is based on the assumption that minor variations between two pieces of text have no impact on the meaning. However, in sentiment analysis, a single word will completely alter the context of a sentence. Another difficulty is that a term that is considered positive in one context can be regarded as negative in another. The majority of reviews will include both positive and negative feedback, especially in more personal formats like Twitter or blogs, which are convenient for a person to understand, but more difficult for a machine to parse. Moreover, it is often even difficult for humans to understand what someone thinks based on a short text because it frequently requires some additional context, for example when someone says "I like this movie more than the last one", you need to know how the person liked the previous movie (Vinodhini and Chandrasekaran 2015).

Sentiment Analysis (SA) or opinion mining refers to the use of natural language processing (NLP), computational linguistics, and text analysis to identify and extract subjective states and information. Nowadays, sentiment Analysis is applied to survey responses, online reviews, and social media. Thus, SA objectives are: to classify a text between positive, negative, or neutral, estimate the subjectivity of the text, aspect extraction, sarcasm detection, and entity recognition (Madhoushi et al, 2015). There are many techniques to analyze sentiments. These are divided between Machine Learning approaches (supervised, unsupervised, deep learning), and Lexicon approach (Qazi et al, 2017)

1.1 Lexicon approach

The lexicon approach is the oldest and simplest method. It relies on a lexicon that assigns to each word a polarity score representing whether it has a positive, neutral, or negative connotation. The document is classified by counting the sentiment that predominates the most (Strine, 2019).

Among the most popular dictionaries are the Binng lexicon, the AFINN dictionary, the NRC dictionary, and the MPQA lexicon. All four lexicons are based on single words or unigrams. The simplest dictionary is the Bing lexicon. It includes 6,788 words which are categorized in a binary fashion: positive or negative. The negative words predominate in this dictionary, representing 70,4% of the total of words. The NRC lexicon includes 14,182 words. It classifies the words in two categories: positive or negative and recognizes emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, trust. In contrast, words in the AFINN lexicon are assigned to a range that goes from -5 to 5. In this range, -5 represents negative scores and 5 represents positive sentiment. The MPQA dictionary besides assigning sentiment also includes a subjectivity measurement. (Strine, 2019).

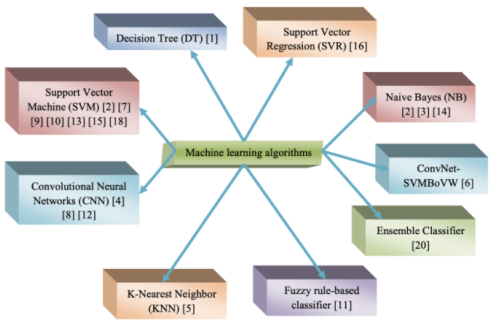
The evident disadvantage of the dictionary methods is that they only represent a small fraction of the existent words, and cannot be applied to context-specific topics. According to the Oxford English dictionary, there are 600,000 words. In this regard, the Binning lexicon represents the only 3.95% of the English words, meanwhile, the NRC includes 8.27% of the total English words. Another disadvantage of the dictionary methods is that they ignore context and the analysis of unigrams could generate a biased analysis (Strine, 2019). If a dictionary does not count how words are used, it will be more likely to misclassify them. A dictionary method will not be able to recognize sarcasm. Hence, it is important to take into consideration the context and not only to count the polarity of a word.

1.2 Machine Learning approach

The Machine Learning approach to Sentiment Analysis is more practical than the lexicon approach since it can be implemented automatically and can handle a large amount of online data (Madhoushi et al, 2015). As we can see, in Figure 2 this approach includes Naïve Bayes, Support Vector Machine (SVM), K-Nearest Neighborhood, and Neural Networks, among others.

These algorithms can be categorized into three types: supervised, unsupervised, and semi-supervised learning methods. Within the supervised machine learning algorithms, Naïve Bayes and SVM are the most common and effective. However, they can be sensitive to both the quality and quantity of the training data (Madhoushi et al, 2015). Nevertheless, Qazi et al (2017) argues that SVM outperforms Naïve Bayes's performance while classifying. Under the unsupervised learning scope, algorithms such as LDA are used for sentiment analysis. Their main limitation is that they need a large amount of data to be trained, otherwise they will produce incoherent classifications.

Figure 2: "Machine Learning Algorithms used in SA"



Additionally, deep learning algorithms such as Recursive Neural Nets, Recurrent Neural Nets, and Long Short Term Memory Networks have been useful for Sentiment Analysis. These deep learning algorithms are capable of identifying predictive features from subtle linguistic attributes. Thus, they can identify sarcasm, humor, and complex sentiments better than any other algorithm (Kumar et al, 2020). In this regard according to Stine (2019), Deep Neural Nets provide the most common accurate classifiers for sentiment analysis. However, to perform properly they require large amounts of data. Therefore, in a small amount of data, their performance do not differ from traditional machine learning algorithms such as SVM or logistic regression.

1.3 Sentiment Analysis in the context of Stock Market

For a long time, academics have been interested in predicting stock market values. Stock markets are difficult to forecast due to their high volatility, which is influenced by a variety of political and economic influences and changes in leadership, market sentiment, and a variety of other factors. Price movement is guided by positive and negative sentiment, which also provides trading and investing options for successful traders and long-term buyers. According to existing sentiment research reports, there is a clear connection between the movement of stock markets and the circulation of news stories (Kala, 2020). Aside from historical prices, the new stock market is influenced by public sentiment. One of the essential variables that affect the stock price is the general social attitude towards that company. Significant quantities of mood data are now accessible thanks to the rise of online social networks, and the predictive performance of the models can be improved by combining knowledge from social media with historical prices. Social network sentiment analysis is often complicated because the text is typically brief, with several misspellings, unusual grammar structures, and so on (Nguyen and Shirai 2015). The algorithms used for sentiment analysis in this context are typically Support Vector Machines, Naive Bayes Regression,

and deep learning algorithms. The accuracy of the deep learning algorithms depends strongly on the amount of training data (Kala, 2020). There are already some successful publications in the field of sentiment analysis and social media concerning stock buying/selling decisions. Mittal and Goel, for example, intending to exchange have published a research paper in which they predicted the changes in the closing values of the Dow Jones Industrial Index based on Twitter sentiment analysis. Using a Self Organizing Fuzzy Neural Network, they achieved an accuracy of 75.56% in their predictions (Mittal and Goel 2009). In an article published in the newspaper "Medium," the best model that could be archived to predict Nifty Index Data was an LSTM model including the news sentiment from 5 years (Kala 2020).

1.4 WallStreetBets

WallStreetBets is a subreddit dedicated to witty memes and high-stakes options trading (Boylston et al. 2020). The subreddit now has over 10 million subscribers as of May 6, 2021. Jaime Rogozinski founded the subreddit of the social media platform Reddit in 2012 to exchange high-risk investment and trading ideas.

The topic of sentiment analysis concerning investment recommendations is currently the subject of particular public interest, as the Gamestop stock has been subject to a so-called "short squeeze" caused mainly by the forum's users, which made shares go up to almost $ 500 before going down to around $ 50 just a couple of days later (Bradley et al. 2021). People in the subforum "WallStreetBets" observed that many hedge funds bought options with which they tried to make profits from the decline of the Gamestop. The platform's users banded together and collectively surged, causing hedge funds to suffer losses in the billions of dollars. We can say that many users bought a stock not because of any underlying trend or press but because it was being bought by other people of the forum (Financial Times 2021; The Economist 2021).

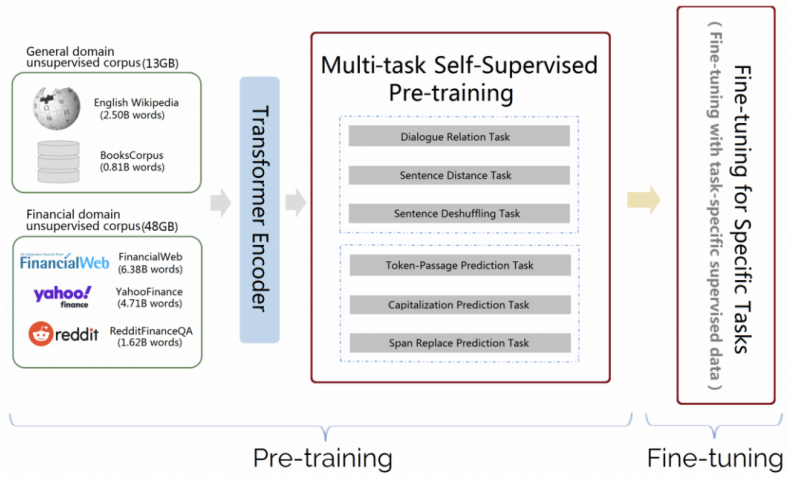
These "investment crowds" on Reddit and other social media have become a driving force influencing the financial markets, and they are sometimes even seen as a threat to functioning markets. This shows that even regulators have no power to control these types of agreements on social media.

1.5 Bert and FinBert

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a deep learning model for natural language processing (NLP) created and published by Google in 2018. BERT is designed to pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers (Akshay Prakash. 2019). Meaning, it looks at the words before and after entities and context pre-trained on Wikipedia to provide a richer understanding of language. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create powerful models for a wide range of tasks, such as question answering, language inference and

sentiment analysis (Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. 2018). In 2019, Prosus AI presented FinBERT (BERT for Financial Text Mining) that is a domain specific language model pre-trained on large-scale financial corpora. FinBERT's purpose was to achieve better domain adaptation by exposing the model to financial jargon. As shown in Figure 3, the FinBERT model is built based on the standard BERT architecture (Devlin. 2019). It is based on the two-stage, first pre-training and then using fine-tuning approach. However, different from BERT, FinBERT is simultaneously trained by six constructed pre-training tasks on general corpora and financial domain which allows the model to better capture semantic information and language knowledge (Araci, D. 2019).

Figure 3: "FinBert"



2 The Application

2.1 The Idea

The phenomenon of "social contagion" that drives up the value of certain stocks is very likely to happen again (Financial Times 2021). The relevance of sentiment analysis and Reddit is illustrated by the constant discussion of the "WallStreetBets" forum in connection with the furor over the Gamestop shares in the news. The "WallStreetBets" forum users have developed their own slang, which makes it not easy for everyone to fully understand the platform. The Economist described this slang even as a "barrier to entry" for new members (The Economist, 2021). In this regard, our team decided that we want to dig deeper into the language used in the forum and build an application that tells its users which stocks to buy and/or sell based on performing a sentiment analysis on the subreddit "WallStreetBets".

2.2 Preprocessing

To start our data preprocessing, we first ingested data from the Subreddit WallStreetBets using the pushshift.io Reddit API, which is the main Reddit API used for extracting data from the forum using Python. It would have been nice to ingest data from more days, but since our computational power was limited and since the hype around the thread is relatively new, we went for one year of data. As a second step, we extracted the cashtags of the individual posts. The cashtags uniquely identify each stock that is discussed in the thread.
Furthermore, we did some basic data cleaning. We created and transformed a timestamp for each post and dropped some null values. Some cashtags needed to be cleaned so that no special characters were contained, and we dropped a majority of the columns that were not required for our model.
Consequently, we imported the VADER (Valence Aware Dictionary for Sentiment Reasoning) sentiment lexicon. This dictionary is an open-source, rule-based sentiment analysis tool specifically trained using sentiments expressed in social media. It not only extracts if a text is positive or negative but also the intensity of emotion. VADER is based on a dictionary that maps lexical characteristics to emotion intensities, referred to as sentiment ratings. Then, a text's sentiment score can be calculated by including the strength of each word in the text (Beri 2020).

As mentioned in the literature review, WSB is characterized by the increased use of slang. This is why, as a next step, we manually added 240 words that are commonly used in WallStreetBets and their corresponding sentiment score to our lexicon. At first, we tried creating our own small slang dictionary, but luckily we could contact the author of an article published on www.medium.com who attempted to build a similar model. He was very friendly to provide us his slang dictionary (Mjysong, 2020). For further reference, please find the article in the source list at the end of this report.
Subsequently, we also changed the level of positive/negative sentiment for certain emojis included in the VADER sentiment lexicon. Although the lexicon consists of emojis, the corresponding sentiment scores were not always correct nor content specific. This is again because the people on WallStreetBets have created their own slang; for example, a moon has a very positive sentiment in the context of the forum, while it usually is kind of neutral.

Then, we applied our lexicon approach to the comments we gathered from Reddit. We defined a comment as positive if its compound score is bigger than 0.1, as neutral if the score is between -0.1 and 0.1, and as negative if the score is smaller than -0.1. The scores attributed to these comments were the data that we will use in the next step to train our FinBert model. Taking advantage of the powerful capabilities of the pre-trained FinBERT model, we proceeded to fine-tune the model to perform sentiment classification on the Subreddit WallStreetBets by training the model further with the labelled data obtained using the lexicon approach.

2.3 Backtesting

We were not able to test or model in a traditional way due to the lack of labeled data serving as a "source of truth," so we decided to perform backtesting to assess how well our model would have done ex-post: "How much profit would we have made if we had used our model for trading stocks?".

After evaluating the sentiment predictions performed by our fine-tuned FinBERT we realised that the accuracy and quality was lower than when using only the lexicon approach. We attributed this to several reasons, first being that the FinBERT model was not able to capture the meaning and sentiment of emojis. This is because of the data tokenization necessary to train the model. FinBERT will tokenize words that it understands and has been pre-trained on, but will not tokenize accurately emojis as it has not been pre-trained on emojis. On the other hand, in the lexicon approach we were able to manually assign a sentiment to each emoji and keyword that appeared frequently. We believe this was the biggest contributor to a better performance because the language contained in the Subreddit was mostly informal and often vague in message and words, meaning sentiment on stocks was often conveyed through emojis and keywords (e.g "to the moon") over the use of elaborated financial language which is the type of text on which FinBERT is pre-trained on. Additionally, we had limited amounts of data to train the FinBERT model as the computational power we had access to was quite limited and the labelling of the data was very time consuming. Therefore, we decided to proceed with the sentiment classification done by the lexicon approach.

We took our data set and removed the dates that were not trading days using a package called "pandas_market_calendars." Then, we looked at how the positive, negative, and neutral comments developed over time. We realized that there are some "peak times" when the positive comments exploded and the negative and neutral ones. We can say that a stock is hyped when there are many comments in general, not only when there are positive comments. We then scraped the prices of those stocks that were hyped on Reddit according to our model. Then, under the assumption that we have a portfolio value of 1000€, we decided to buy a stock if there were more positive comments about a stock than the day before, and we decided to sell a stock if there were fewer positive comments about a stock than on the previous day. The author of the medium article used a ratio on positive vs. negative comments to determine whether to buy or sell a stock, but since we saw that it is not only about positive or negative comments but about "hyped stocks" in general, we decided to only focus on positive comments.

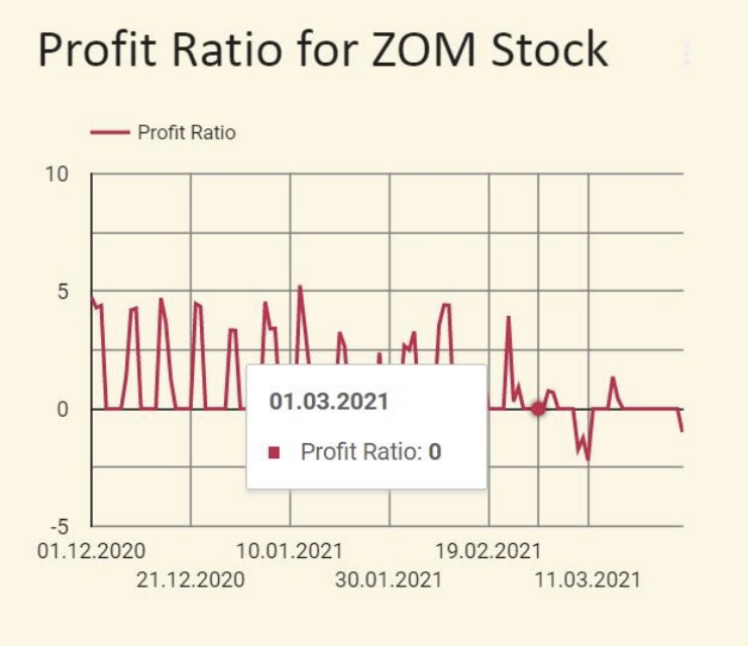
As a next step, we calculated our profit. One problem of our model was that, for example for the GameStop stock, we were able to make a lot of profit at first, but our model did not take into account the "overheating" of the market. Using our model, we still bought stocks when the price of the stock was already very

high. This is why there were many losses for the GameStop stock in our model, but we were able to make a significant profit for other stocks. Initially, we wanted to also look at some SP 500 stocks. Still, after researching the nature of WallStreetBets, we realized that it does not make sense because the forum deals with high-risk stocks and is instead about gambling than rational, traditional investing.

3 Results

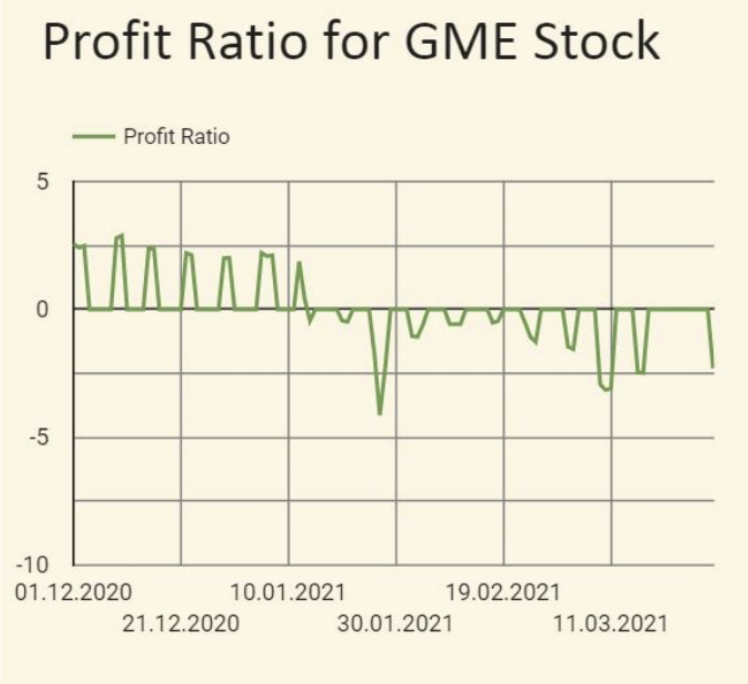
To sum up what we have found, we tried to summarize the results of our analysis in a visually interactive dashboard. Throughout our approach, we discovered that not necessarily the more advanced approach works better (i.e. FinBert) and that with the more manual approach (Lexicon) we were actually able to achieve more accurate results. Why does this happen? In WallStreetBets posts, we can identified a highly unique language including a lot of emojis. Adding these manually to a lexicon helped us to nearly perfectly capture the emotions. Furthermore, we proved that our investment strategy might not only make profits but in overall it can give us a good starting point for a successful investment strategy (see Figure 4). One problem, which would need to be tackled in future is to capture the overheating of the market (see Figure 5).

Figure 4: "Profit ratio for ZOM stock resulting from backtesting our model"



Furthermore, the overall sentiment analysis could be extended by not only including WallStreetBets but also other sources such as Twitter or any other social media platform. This way, we would be able to train our model with more data and maybe by means of that FinBert would start performing better. Finally, for a better overview of the market the dashboard could be transformed into a real time one.

Figure 5: "Profit ratio for GME stock resulting from backtesting our model"



4 References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. https://arxiv.org/pdf/1908.10063.pdf

Akshay Prakash. (2019). BERT: Bidirectional Encoder Representations from Transformers.https://medium.com/swlh/bert-bidirectional-encoder-representations-from-transformers-c1ba3ef5e2f4

Beri, A. (2020, May 27). SENTIMENTAL ANALYSIS USING VADER - Towards Data Science. Medium. https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

Bradley, D., Hanousek Jr., J., Jame, R., Xiao, Z. (2021). Place Your Bets? The Market Consequences of Investment Advice on Reddit's Wallstreetbets. SSRN Electronic Journal, 1–34. https://doi.org/10.2139/ssrn.3806065

Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/pdf/1810.04805.pdf

Hajiali, M. (2020). Big data and sentiment analysis: A comprehensive and systematic literature review. Concurrency and Computation: Practice and Experience, 32(14), 1–21. https://doi.org/10.1002/cpe.5671

Mäntylä, M. V., Graziotin, D., Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

D.Vidhya G. Sivaselvan, V.Vennila PG Scholar, Asst. Prof, Department of Computer Science and Engineering, K.S.R. College of Engineering, Tiruchencode (2015) Techniques of Opinion Mining and Sentiment Analysis: A Survey. International Journal of Communication and Computer Technologies, 3 (2), 72-78. doi:10.31838/ijccts/03.02.05

Georgia Institute of Technology, Boylston, C. J., Palacios, B. E., Tassev, P. T. (2021, January). WallStreetBets: Positions or Ban. https://arxiv.org/ftp/arxiv/papers/2101/2101.12110.pdf

Kala, S. (2020, August 16). Stock Market Prediction using News Sentiments-I - Shagun Kala. Medium. https://medium.com/@kala.shagun/stock-market-prediction-using-news-sentiments-f9101e5ee1f4

Liu, Z., Huang, D., Huang, K., Li, Z., Zhao, J. (2020). Finbert: A pre-trained financial language representation model for financial text mining. https://www.ijcai.org/proceedings/2020/0622.pdf

Liu, B. (2012). Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies). Morgan Claypool Publishers.

Kumar, A., Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. Concurrency Computation , 32(1), 1–29. https://doi.org/10.1002/cpe.5107

Mjysong. (2020, May 2). Momentum Trading off Sentiment from r/wallstreetbets. Medium. https://medium.com/@mjysong/momentum-trading-off-sentiment-from-r-wallstreetbets-149c19c7538d

Madhoushi, Z., Hamdan, A. R., Zainudin, S. (2015). Sentiment analysis techniques in recent works.

Proceedings of the 2015 Science and Information Conference, SAI 2015, March, 288–291. https://doi.org/10.1109/SAI.2015.7237157

How herd behaviour drives action on r/WallStreetBets. (2021, February 10). Financial Times. https://www.ft.com/content/971df303-726a-4bdf-93eb-9a9e848f7109

Nguyen, T. H., Shirai, K. (2015). Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 1354–1364. https://www.aclweb.org/anthology/P15-1131.pdf

Qazi, A., Raj, R. G., Hardaker, G., Standing, C. (2017). A systematic literature review on opinion types and sentiment analysis techniques: Tasks and challenges. In Internet Research (Vol. 27, Issue 3). https://doi.org/10.1108/IntR-04-2016-0086

Stanford University, Goel, Mittal. (2009). Stock Prediction Using Twitter Sentiment Analysis.
http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsing TwitterSentimentAnalysis.pdf

Stine, Robert A., Sentiment Analysis (March 2019). Annual Review of Statistics and Its Application, Vol. 6, Issue 1, pp. 287-308, 2019, Available at SSRN: https://ssrn.com/abstract=3382106 or http://dx.doi.org/10.1146/annurev-statistics-030718-105242

The Economist. (2021, February 5). How WallStreetBets works.
https://www.economist.com/finance-and-economics/2021/02/06/how-wallstreetbets-works