STOCK MARKET PRICE PREDICTION USING SENTIMENT ANALYSIS OF STOCK ANALYSIS HEADLINES

Jyotika Kundu
Research Scholar
IMSAR, Maharshi Dayanand University
Rohtak, Haryana
E-mail: kundujyotika@gmail.com

Anshula Garg
Assistant Professor
Delhi School of Journalism
University of Delhi, Delhi
E-mail: anshu.anshula.garg@gmail.com

Aarti
Assistant Professor
IMSAR, Maharshi Dayanand University
Rohtak, Haryana
E-mail: rtichahal@gmail.com

ABSTRACT

Adequate analysis is required before investing in the stock market. Researchers remain interested in finding out the factors that can accurately predict the tomorrow's stock market price. This study examined the impact of sentiment analysis using stock analysis headlines and to compare two popular tools of prediction namely TextBlob and VADER sentiment analysis. The method used was sentiment analysis using two tools VADER and TextBlob (using Python). The headlines are used for analysis in this study from the relevant period suits to research. The Sentiment analysis is gaining popularity in stock market prediction. However, according to the results sentiment analysis should not be used alone for predicting stock prices. Further, it is observed that VADER sentiment analysis performs better than TextBlob. Using VADER, sentiment analysis explains 3.168% of the total variation in stock prices whereas TextBlob could explain only 0.16% which indicates the effectiveness of VADER over TextBlob. This research focused on analysing the impact of stock analysis headlines on tomorrow's stock prices whereas it can have an immediate impact too. Future researchers can focus on this aspect. Very few researches compared the two sentiment analysis tools to find out their accuracy. This is crucial as it can impact their decision.

Keywords: Machine Learning, Sentiment Analysis, Opinion Mining, TextBlob, VADER

INTRODUCTION

Researchers and analysts always remain interested in finding the factors that affect stock market prices. It is observed that predicting the stock market prices is hard due to its volatile nature and it depends on some factors like political and economic situations, international transactions, and change in leadership, investor sentiment and many more. Prediction of stock market prices based only on historical data or information in text form appears insufficient. Some existing studies revealed that there is a strong correlation between the movement of stock prices and the publication of news articles. Various sentiment analysis studies are done on various levels by using the algorithms like support vector machines, naive Bayes regression, and deep learning (Mohan *et al*., 2019).

Stock Market

It is a collection of markets and exchanges where activities like buying, selling, and issuance of shares of publicly-held companies happen. All these financial activities are conducted by institutionalized formal exchanges or over-the-counter (OTC) marketplaces (Chen, 2021). Stock markets have rewarding kind of nature, so they appeal to retail investors and general public, but their working culture is complex and difficult (Pavan *et al*., 2021).

Factors Affecting Stock Market Prices

Mitchell (2020) defines the following factors that have profound effect on stock prices

• **Government and Capital Banks:** The monetary and fiscal policies can cause fluctuations in the stock market. The decision to increase or decrease interest rate and increase or decrease in government spending plays a huge role in deciding the price movement of stocks.

• **International Transactions:** More movement of money outside India can impact the economy negatively whereas when India exports and money is coming to India can impact the economy and hence, the stock prices.

• **Speculation and Expectation:** Politicians, investors and consumers have different opinion regarding the economy performance. These views can be analysed via the sentiment analysis to get the fair idea about the performance of economy.

• **Supply and Demand:** Demand and supply play a huge role in price changes. If demand is more, supply is less than the price will rise and vice-versa.

These factors can create both short-term and long-term fluctuation in the markets. Government news regarding changes in tax policy, any news related to a company or any change in Federal Reserve can drastically impact the short-term trends.

Stock Market Price Prediction

Artificial intelligence (AI) and data mining techniques are helpful to do the stock market data analysis. Volatile nature of stock prices is affected by gains and losses of certain companies where news articles are also the most important factor that influences the stock market prices (Kirange and Deshmukh, 2016). As per Market Future Report (2021), sentiment analytics market is expected to reach $ 6 billion by 2023.

Prediction of stock market prices is very difficult and so many factors like physical vs. psychological, rational and irrational behavior, etc. are also involved in this process. All these are combined for volatility of share prices. But through machine learning technique some patterns can be unearthed and can be used to do accurate predictions (Singh, 2018).

Machine Learning

This technique was first time introduced by Arthur Samuel in 1959 (Machine Learning Tutorial, n.d.). Machine learning provides the capability to computers to learn without being explicitly programmed. As defined by name it makes the computers more similar to humans: The ability to learn. This technique is used actively nowadays at many more places (Machine Learning, 2019). It is called as the subset of AI, it mainly involves the development of algorithms. This technique makes the computer learn from data and past experiences on its own. Machine learning has combined the computer science and statistics for the creation of some predictive models (Machine Learning Tutorial, n.d.).

Sentiment Analysis

A decade ago, in research it was concluded that between sentiment and future prices there was no significant correlation. It can be due to use of social media in large no. and more amount of data can be drawn. Sentiment analysis is a process to evaluate text and score in three departments like negative, neutral and positive. Vader, an incredible library, is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. This tool can analyze slang, emojis, initialism and acronyms. It appears easy to do analysis before saving any scrapped data from FinViz and the tweets (Maruya-Li, 2020).

Sentiment analysis helps to do opinion mining and text analysis and emotion. AI is used for natural language processing to do analysis of conversations and determination of emotional tone behind the words and know the attitudes and opinions. Humans do this instinctively- without thinking (Meg, 2021).

Need for Sentiment Analysis

It is needed to monitor the reputation of a brand and improve the experience of customers, to manage the crisis, competitive benchmarking, to find best influencers and customer services (ibid).

Tools of Sentiment Analysis

There are various tools for sentiment analysis like HubSpot's Service Hub, Net Promoter Score, RapidMiner, NCSU Tweet Visualizer, Meaning Cloud, Social Mention, TextBlob, VADER, Rosette, and Social Searcher (ibid).

LITERATURE REVIEW

Nemes & Kiss (2021) in their study have detected that there is a significant difference in correlation metrics. When talked about TextBlob, "there is a significant effect on the values of stock exchange in terms of opening, closing, highest and lowest values" (p.392). NLTK-Vader Lexicon also provides similar results but slightly reduced efficiency than the previous model. Further, this study described that economic news headlines impacted the stock market values and significant differences are observed between multiple analytical tools. On the other hand, Carvajal, Schmitt and Johnson (2020) stressed that VADER is a better performer than TextBlob.

Another study based on movie reviews by Bonta and Janardhan (2019) argued that VADER performs better than TextBlob.

Houston *et al.* (2019) in their research study have reported that a machine learning approach is defined for training and prediction of bimolecular thermal rate constants in a temperature range at larger level. Gaussian process regression technique is used to evaluate the differences.

Chiong *et al.* (2018) for prediction of financial market through news disclosures have presented a sentiment analysis-based machine learning approach. Results have shown that to extract useful features from financial news, sentiment analysis is better way as it helps to reduce the feature dimensions in significant manner.

Puri & Prasad (2018) in their research study have exclaimed that there is a correlation between technical and sentiment analysis, sentiment selection significantly helps to improve accuracy while creating technical analysis based financial

forecasting system. Applied machine learning technique is able to extract relevant information from input data. Scherbina, & Schlusche (2015) have documented that economically-linked stocks have the ability to predict each other's returns. It is challenging to showcase economic challenges between firms and these linkages can only be identified through coverage of media.

Medhat *et al.* (2014) in their research paper have clarified that there is an open field of research which is observed in various algorithm enhancements. Even though there are lack of resources, interest of researchers in more languages including English is increasing. Wordnet is most commonly used and it is used by other languages too other than English. In sentiment analysis task resources are built, for many natural languages these are needed. In sentiment analysis, recently information from micro-blogs, blogs, forums and news source is used widely. In this way information from media has a great role to express feelings and opinions of peoples about a certain topic or product. There is a need of deeper analysis of SNS as data source, likely in reviews of IMDB. Sometimes context of text and user preferences becomes important.

Liu (2012) has analysed that sentiment analysis is highly challenging in technical point of view. Processing of natural language is not an easy task but popular way of doing research as mostly is believed on machine learning. There are many machine learning algorithms that produce results that are not understandable to humans.

Lanne (2002) has concluded that there is no predictability of stock returns in stationary test and asymptotically controls the size also in finite samples. T-test suggests the predictability but it is not valid asymptotically and rejects the finite samples. In T-test power properties can be evaluated using Monte Carlo simulation methods.

Ferson and Korajczyk (1995) have revealed the fact that with the help of models, predictability in large fractions can be explained for investment horizons. It is resulted that for shorter-horizon returns, statistical tests reject the constant-beta assumptions but not for long-horizons returns. Models that are using five economic models and principal components act as risk factors. These have same ability of explanation to describe the predictability in portfolio returns.

In the light of above discussion, we conclude that there are few researches that compared widely used VADER and TextBlob approaches and predict correlation using them.

The objectives are:

1) to compare the performance of TextBlob and VADER sentiment analysis

2) to find correlation between sentiment analysis scores and tomorrow's stock price values

METHODOLOGY

Data Frame of Analysis Headlines

There are many ways to approach data frame building. One way is to collect the news headlines manually which takes a lot of time for its collection. Further, it needs accuracy in collecting the relevant headlines that can impact the stock price. Other possibility is to collect the headlines from automated sources that are highly reliable, up-to-date, and accurate. Keeping this in mind, automated process was chosen for collection of data. Seeking Alpha was selected because it provides comprehensive coverage of news. It is also worth noting that while predicting the stock market prices, analysis headlines play the most important role in deciding its sentiment. This is the reason why Seeking Alpha was picked as it provides two different sections namely news headlines and analysis headlines. For example, the title of the news headline on Oct 6, 2021 is "Hot Stocks: EV stocks fall; VLCN pops in debut; IGT sets high; TAK suspends trials; short-seller targets DNA". This news is hard to interpret whether this is positive, neutral or negative. This can mislead the results whereas when analysis headlines are considered that provides the fair analysis of the news which can be described as positive, negative or neutral.

The American electric vehicle and clean energy-based company Tesla Inc. was chosen for the study, as this company best suits to the relevancy to the research study. The period of data was from 27-03-2021 to 07-10-2021. Total headlines were 106.

Manual scraping of articles was performed by entering the headlines and the dates in numbers (version 11.0). For writing code in python, Google Colab was used.

Figure 1: Initial Title and Date Data after Uploading the .csv file

Source: Yahoo Finance

```
import pandas as pd

from google.colab import files

uploaded=files.upload()

df1 = pd.read_csv("Tesla_newsOct11.csv", encoding='UTF-8', header=None)

df1.columns = ['Title', 'Date']

df1
```

Listing 1: Part of Data Frame Builder

Listing 1 shows the process of data frame building. Firstly the headlines and dates stored in .csv file were uploaded using files.upload(). Pandas here allows various formats such as Microsoft Excel, SQL, Comma-separated values (csv), etc. Data frame (df1) was built using pd.read_csv. Figure 1 shows the initial data shows the title of the headlines and the date when it was published. Here, the date column is not well organised which requires further cleaning. The organised date and title data is shown in Listing 2 and Figure 2

Figure 2: Final Data for Analysis after Cleanup

Source: Yahoo Finance

```
import re

print(df1['Date'][1])

match        =        re.search(r'\w{3}\.\s\d{1,2}',
df1['Date'][1])

modifiedDate = match[0] + " 2021"

print(modifiedDate)
```

Listing 2: A Part of Proper Date Format for Data Analysis

This code was used to get the complete date in day, month and year format. As we know that by default the dates are in string format in python (Liew, 2020). For further analysis, it was converted to date time format. After these changes, Figure 2 shows the final data.

Company Specific Stock Prices

Stock market data was collected using 'Yahoo Finance' from 27-03-2021 to 07-10-2021. This gives the opportunity to analyse the stock market prices using headlines in the specific period. A part of this is shown in Figure 3.

Figure 3: Company Specific Stock Price Data
Source: Yahoo Finance

RESULTS

Sentiment Analysis Using VADER and TextBlob

There are various ways to perform sentiment analysis from structured approach to unstructured approach. According to Talpada, Halgamuge and Vinh (2019), lexicon based approaches (TextBlob, VADER etc.) perform better than any other tools. In the present case, we will compare TextBlob and VADER Lexicon based approach. The main motive is to study the analysis headlines of Tesla and to determine the sentiment values whether these are positive, negative or neutral.

In sentiment analysis, headline related to a company is labelled according to the sentiment score and the polarity associated with the headlines. With the help of the data, possible analysis and comparisons can be carried out. As a result, analysing and displaying the emotional impact of economic news headlines on stock market fluctuations, as well as determining how powerful these headlines may be without further content.

The other objective is to assess the strength and accuracy of TextBlob and VADER sentiment analysis techniques based on the context.

TextBlob Sentiment Analysis

TextBlob is a strong NLP (Natural Language Processing) library built upon NLTK (Natural Language Toolkit) which provides a user-friendly interface to the NLTK library (Nemes and Kiss, 2021). TextBlob provides the polarity of the news based on its dictionary from -1 to +1 and the sentiment type ranges from negative, neutral, and positive.

Figure 4 represents the collective sentiment analysis of headlines. Out of 106 analysis headlines, neutral statements were 67 (63.20%), positive were 31(29.25%) and finally, negative statements were 8 (7.55%).
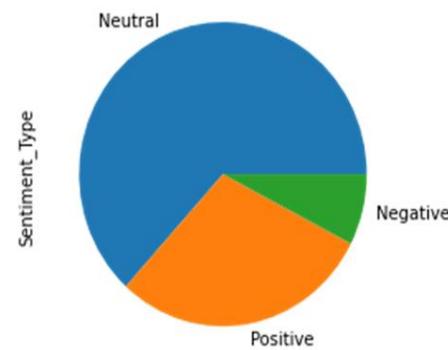
Figure 4: Aggregate Sentiment Analysis of Headlines
Source: Author's Own

Figure 5 shows the aggregate polarity values day wise. Here, aggregate means if there were more than one analysis headline on a particular day that could affect the stock prices then sum of all the polarity scores on that day is taken. Y-axis represents the day-wise polarity scores. Polarity scores greater than 0 shows positive news and less than 0 means negative news.
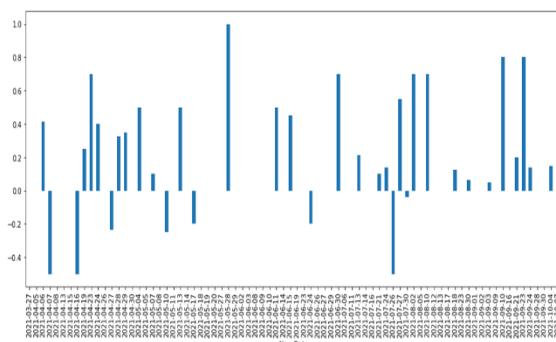
Figure 5: Day-wise Aggregate Polarity Values
(27-03-2021 to 07-10-2021)

Source: Yahoo Finance

It is observed that 63.20% (neutral) news are shifting to positive side which can misrepresent the real results. The blank spaces indicate that there was no news of significance or simply no news at all regarding stocks in those particular days. The talks about opening of Japan's markets for Tesla and Tesla's solar roof electric production are the key factors that lead to positivity in the polarity scores. In the recent times, its outstanding performance during the pandemic despite its semi-conductor shortages seem to create positive news.

VADER Sentiment Analysis
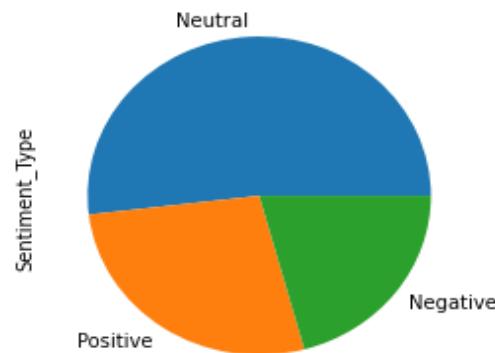
Figure 6: Aggregate Sentiment Analysis of Headlines

Source: Author's Own

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a rule based and lexicon based dictionary that is specifically designed for sentiments expressed on social media and texts from other platforms (Nemes and kiss, 2021). It also ranges between -1 and +1. There is also classified in three categories namely positive, negative and neutral.

Out of 106 headlines, Figure 6 shows 55 (51.87%) neutral headlines, 29(27.36%) positive headlines, and 22(20.75%) negative headlines.

Figure 7 shows the aggregate sentiment scores for each day. Here, unlike TextBlob not all neutral are shifting towards positive side. According to polarity, there are many headlines according to VADER dictionary which are moving in negative direction. Score greater than 0 indicates positive outlook towards the news headline on that day and score less than 0 indicates negative outlook.
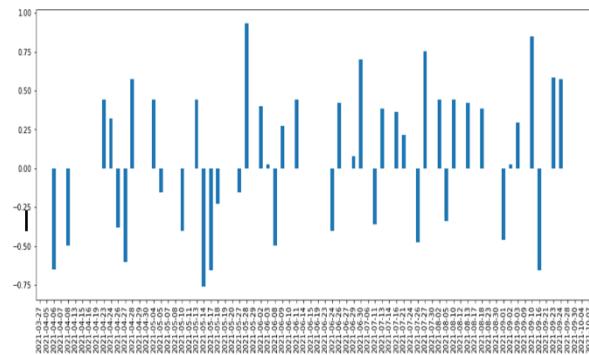
Figure 7: Day-wise Aggregate Sentiment Values
(27-03-2021 to 07-10-2021)

Source: Yahoo Finance

Sentiment and Stock Price Analysis

To find the correlation between the sentiment scores and adjusted closing price, returns were calculated using the formula current adjusted closing price-prior day's closing price/prior day's closing price. There are instances when there were no headlines in some days. So, only those adjusted closing price were taken for which there were headlines on that day. Stock market prices do get affected even after small changes in the market. In this study, effect of sentiment today on tomorrow's stock prices is observed. This is done due to two reasons 1) it is not possible to see every minute's changes, 2) there are cases when there are more than one headlines in a particular day so to calculate its impact collectively on tomorrow's stock prices seem suitable.

Sentiment and Stock Price Analysis Using TextBlob

Date	Returns	Polarity(1)
2021-07-06	-0.028458	0.70
2021-07-30	0.014542	0.55
2021-08-05	0.005219	0.70
2021-08-12	0.020387	0.70
2021-09-16	0.001535	0.80
2021-09-24	0.027533	0.80

Figure 8: Returns and Polarity Scores in TextBlob (>0.5 and <-0.5)

Source: Yahoo Finance

dfEodPrice4   =
dfEodPrice3[(dfEodPrice3['Polarity(1)'] > 0.5) |
(dfEodPrice3['Polarity(1)'] < -0.5)]

dfEodPrice4

dfEodPrice4['Returns'].corr(dfEodPrice4['Polarity(1)'])

Listing 3: Correlation Coding

Listing 3 shows the correlation relating coding and Figure 8 indicates its output. These few scores are considered for correlation calculation because either >0.5 or <-0.5 scores reveal significantly positive and negative headlines that can impact tomorrow's price. When the correlation between returns and polarity was performed (the coding is mentioned in Listing 3). The correlation coefficient in case of TextBlob came out to be 0.042616252826418614. This means that sentiment analysis explains only 0.16% of the total factors that are causing stock prices to move up and down.

Sentiment and Stock Price Analysis Using VADER Sentiment Analysis

Similar code was used for calculating the correlation between returns and sentiment score as mentioned in Listing 3. Figure 9 indicates the sentiment score using VADER sentiment analysis i.e. the headlines that can have significant impact on tomorrow's stock prices (>0.5 and < -0.5). When the correlation between sentiment scores and returns was performed, the correlation coefficient came out to be 0.17821111687478544.

This shows that the sentiment analysis explains 3.168% of the total factors that are causing movement in stock prices. This can be verified from Figure 7, the day wise sentiment scores are not moving in just one direction. Instead, according to its polarity, it is either going to positive or negative direction.

Date	Returns	Score(1)
2021-04-07	-0.029857	-0.6489
2021-04-28	-0.014672	-0.5994
2021-04-29	-0.025058	0.5719
2021-05-17	-0.021891	-0.7579
2021-05-18	0.001803	-0.6520
2021-07-06	-0.028458	0.7003
2021-07-30	0.014542	0.7513
2021-09-16	0.001535	0.8481
2021-09-21	0.012614	-0.6519
2021-09-24	0.027533	0.5859
2021-09-28	-0.017438	0.5719

Figure 9: Returns and Polarity Scores in VADER (>0.5 and <-0.5)

Source: Yahoo Finance

The findings suggest that VADER performs better than TextBlob. These results are in-line with the results presented by Bonta and Janardhan (2019) on movie reviews. Our results also match with Carvajal*et al.* (2020) regarding VADER and Text Blob performance. The overall correlation of return and sentiment scores (using both TextBlob and VADER) are significantly low to be used alone for prediction.

CONCLUSION

This study conducted the sentiment analysis with the aim to explore the correlation between sentiment scores and company's returns and to compare two sentiment analysis tools namely TextBlob and VADER. Sentiment analysis represents a very small percentage of all the factors that impact the stock market prices. VADER seems to perform better than TextBlob as it explains 3.168% and TextBlob explains only 0.16% when calculated on the same data. Sentiment analysis should not be used alone for estimating the future stock prices. It can be used with other strong techniques including the technical analysis of the stocks.

LIMITATIONS AND FUTURE WORK

Like any study, it too suffers from few limitations. The impact of analysis headlines was checked on tomorrow's stock price whereas it can have impact on that day too. So, future researchers can see its immediate impact too on the stock prices.

REFERENCES

Ahmed, N., Michelin, R. A., Xue, W., Ruj, S., Malaney, R., Kanhere, S. S., & Jha, S. K. (2020). A survey of COVID-19 contact tracing apps. *IEEE access*, 8, 134577-134601. https://doi.org/10.1109/ACCESS.2020.3010226

Bonta, V., & Janardhan, N. K. N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. Asian Journal of Computer Science and Technology, 8(S2), 1-6.

Carvajal, G., Schmitt, K., & Johnson, G. B. (2020). Sentiment Analysis on New York Times Articles Data. In Symposium on Undergraduate Research and Creative Expression [Symposium], Indiana.

Chen, J. (20, 03, 2021). Stock Market. Retrieved from Investopedia: https://www.investopedia.com/terms/s/stockmarket.asp

Chiong, R., Fan, Z., Hu, Z., Adam, M. T., Lutz, B., & Neumann, D. (2018, July). A sentiment analysis-based machine learning approach for financial market prediction via news disclosures. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (278-279).

https://doi.org/10.1145/3205651.3208231

Ferson, W. E., & Korajczyk, R. A. (1995). Do arbitrage pricing models explain the predictability of stock returns? Journal of Business, 68(2), 309-349. https://doi.org/10.1086/296806

Houston, P. L., Nandi, A., & Bowman, J. M. (2019). A machine learning approach for prediction of rate constants. The journal of physical chemistry letters, 10(17), 5250-5258.

Kirange, D. K., & Deshmukh, R. R. (2016). Sentiment Analysis of news headlines for stock price prediction. Composoft, An International Journal of Advanced Computer Technology, 5(3), 2080-2084.

Lanne, M. (2002). Testing the predictability of stock returns. Review of Economics and Statistics, 84(3), 407-415. https://doi.org/10.1162/003465302760556804

Liew, L. (2020). *Sentiment analysis with python- A beginner's guide*. Retrieved from https://algotrading101.com/learn/sentiment-analysis-python-guide/(accessed1 October 2021).

Liu, B. (2012). *Sentiment analysis and opinion mining* (Synthesis Lectures on Human Language Technologies, 5(1), 1–167). Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Machine Learning Tutorial. (n.d.). Retrieved from Java T Point: https://www.javatpoint.com/machine-learning

Machine Learning. (19, July 2021). Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/machine-learning/

Market Research Future. (2021, February). *Global Sentiment Analytics Market*. https://www.marketresearchfuture.com/reports/sentiment-analytics-market-4304

Maruya-Li, K. (2020, 10, 28). Stock Price Prediction Using Sentiment Analysis and Historical Stock Data. Retrieved from Medium.com: https://medium.com/swlh/stock-price-prediction-using-sentiment-analysis-and-historical-stock-data-587488db8576

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, 5(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Meg. (2021, 05 27). The best sentiment analysis tools. Retrieved from Talkwalker: https://www.talkwalker.com/blog/best-sentiment-analysis-tools

Mitchell, C. (2020, October 16). *4 Factors That Shape Market Trends*. Investopedia. Retrieved October 19, 2021, from https://www.investopedia.com/articles/trading/09/what-factors-create-trends.asp

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019, April). Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)* (pp. 205–208). IEEE. https://doi.org/10.1109/BigDataService.2019.00042

Nemes, L., & Kiss, A. (2021). Prediction of stock values changes using sentiment analysis of stock news headlines. Journal of Information and Telecommunication, 1-20. https://doi.org/10.1080/24751839.2021.1898957

Pavan Krishna, S. F. (2021, March). Stock Market Prediction Using Sentimental Analysis- A Vader Approach. International Journal of Advanced Research in Engineering and Technology (IJARET), 311-321.

Retrieved from Academia: https://www. academia.edu/50698309/Stock_Market_Pr ediction_Using_Sentimental_Analysis_A_ Vader_Approach

Puri, N., Agarwal, A., & Prasad, P. (2018). A Hybrid Technique for Stock Market Prediction using Machine Learning Approach Sentiment Analysis. International Journal of Pure and Applied Mathematics, 120(6), 11707-11720. https://doi.org/10.12732/ijpam. v120i6.24357

Scherbina, A., & Schlusche, B. (2015). Economic linkages inferred from news stories and the predictability of stock returns. *SSRN Electronic Journal*. https://doi.org/10. 2139/ssrn.2363436

Singh, A. (2018, 10 25). Stock Prices Prediction Using Machine Learning and Deep Learning techniques (with python codes). Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/20 18/10/predicting-stock-price-machine-learningnd-deep-learning-techniques-python/

Talpada, H., Halgamuge, M. N., & Vinh, N. T. Q. (2019, October). An analysis on use of deep learning and lexical-semantic based sentiment analysis method on twitter data to understand the demographic trend of telemedicine. In 2019 11th International Conference on Knowledge and Systems Engineering (KSE) (1-9). IEEE.