International Journal of Data Science and Analysis
2021; 7(1): 13-19
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20210701.14
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Elon Musk's Twitter and Its Correlation with Tesla's Stock Market

Daniel Pyeong Kang Kim[1], Jongwhee Lee[2], Jungwoo Lee[3], Jeanne Suh[4]

[1]Stony Brook School, New York, USA

[2]Yongsan International School of Seoul, Seoul, South Korea

[3]Peddie School, New Jersey, USA

[4]Saint Paul Preparatory, Seoul, South Korea

Email address:

jaydenjongwheelee@gmail.com (Jongwhee L.)

[*]Corresponding author

To cite this article:

Daniel Pyeong Kang Kim, Jongwhee Lee, Jungwoo Lee, Jeanne Suh. Elon Musk's Twitter and Its Correlation with Tesla's Stock Market. *International Journal of Data Science and Analysis*. Vol. 7, No. 1, 2021, pp. 13-19. doi: 10.11648/j.ijdsa.20210701.14

Received: February 28, 2021; **Accepted**: March 16, 2021; **Published**: March 26, 2021

Abstract: Over the past few years, Twitter has rapidly grown into a prominent social media platform, and various research papers have attempted to prove the relationship between the stocks and the tweets made on Twitter. The purpose of this research paper is to investigate the specific connection between Elon Musk's twitter and the stock value of Tesla. The primary form of analysis used was Exploratory Data Analysis to be able to more easily distinguish patterns within our dataset, which was preprocessed to exclude any stopwords. Utilizing various graphs and Machine Learning algorithms such as Logistic Regression and Support Vector Machine, we wrote this research paper that respectively analyzes the change in the close price of Tesla's stock and Elon Musk's Twitter engagement, including tweets, likes, and retweets dating from the start of 2015 up until July of 2020. Furthermore, the article illustrates the contents of Elon Musk's tweets and allows a deeper understanding of other correlations that may exist through the use of Machine Learning to perform Sentiment Analysis. This was achieved by categorizing Elon's tweets into three different tones (positive, negative, and neutral) and seeing how the underlying mood would correspondingly affect Tesla's stock value. The combination of such techniques and factors allowed for a conclusive result in which a distinct correlation was apparent: an increase in the number of tweets/engagement would lead to an increase in the closing price of Tesla, as well as vice versa.

Keywords: Data Science, Elon Musk, Stock Market, Machine Learning, Exploratory Data Analysis, Tesla

1. Introduction

Renowned for his remarkable assets, Elon Musk became the richest person on earth in 2020 [1, 2]. Elon Musk developed x.com, which later became eBay, in 1999, followed by Space X in 2002 and Tesla in 2003 [3, 4]. Despite all the acclaims that Elon Musk receives for his work, his most well-known undertaking is his success with Tesla. Tesla, a company that mainly produces electric vehicles and energy storage products, is said to be the fastest-growing car brand in the world [5]. However, Tesla was surprisingly not founded by Musk himself. It was instead, first established by Martin Eberhard and Marc Tarpenning in 2003, selecting their company's name after a Serbian-American inventor, Nikola Tesla. Musk emerged as a chairman in 2004, after contributing more than $30 million to the new venture. Since then, Tesla has been regarded as a company with unlimited potential in terms of technological advancements with its stocks growing explosively under Elon Musk's supervision. Even though Elon Musk's success is impressive, he has been frequently criticized for his frequent use of Twitter which can be deemed as inappropriate and unprofessional. However, new research shows that when utilized correctly, Twitter can be a powerful marketing tool.

In recent years, with the rise of Twitter as a prominent social media platform, correlations between Twitter tweets and the stock returns have become evident. For example, in 2019, Elodie Michaux conducted research that identified a

correlation between stock returns and tweets, though she was unable to find any sort of correlation regarding tweet count and stock trading volume proving that the stock prices can be affected by tweets. In addition, ScienceDirect, a world-renowned science magazine, shares multiple accounts of research that produced similar results, indicating that Twitter coverage and positive mood in tweets are inevitably going to boost the stock prices [6, 7]. Through the lens of such research in various search engines and repositories, it becomes apparent that an undeniable connection between the stock market and Twitter exists.

Therefore, the purpose of this research paper is to expound upon this topic, narrowing the relevance down to Elon Musk's tweets and the stock returns of Tesla, Inc. Furthermore, with the introduction of Sentiment Analysis, the extraction of subjective emotion out of text and words, our research is able to grasp the hidden emotions behind tweets and replies, then correlating them to the stock price of Tesla as well. In essence, through this research, we aim to use Elon Musk's Twitter activity and content from the past five years (2015-2020) to unearth some form of parallel, or inverse, relationship between what Elon Musk himself is saying on social media and how that affects the stock prices of the largest electric car company in the world.

Section 2 will take a deep dive into what Exploratory Data Analysis, along with how it has been used in our specific research.

Next, Section 3 will discuss our work in preprocessing, the filtering of unnecessary information in the original dataset to suit the information you need.

Last but not least, Section 4 is an in-depth explanation of Machine Learning, its specific algorithms, Sentiment Analysis, and how they all play a role in our research paper.

2. Exploratory Data Analysis (EDA)

EDA, or Exploratory Data Analysis, is an analytical approach that groups the output of a dataset according to a specific set of characteristics. In other words, this type of analysis picks out patterns within the data and categorizes it accordingly. There are four different ways of Exploratory Data Analysis: multivariate non-graphical, multivariate graphical, univariate non-graphical, and univariate graphical [8]. EDA is used primarily to ensure that the data used does not contain any anomalies or outliers, allowing scientists to work with a more reliable and proper set of data.

In this section, we will examine multiple figures and graphs that quantify Elon Musk's twitter account, be it through the number of tweets, likes, replies, retweets, etc. Doing so will allow for easier identification of links and patterns between twitter and Tesla's stock price.

2.1. Close Data

Figure 1 is a line graph illustrating the change in the close of Tesla stock from 2010 to 2020. The price of the stock increased gradually until 2019, when the price began to escalate at a faster pace. It is also the volume of the stock that increased, typically in the year 2020. Such data overall indicates that the value of Tesla and the interest they received had definitely increased throughout the years.

Figure 1. *Closing Price of Tesla Stock.*

2.2. Twitter Engagement Statistics

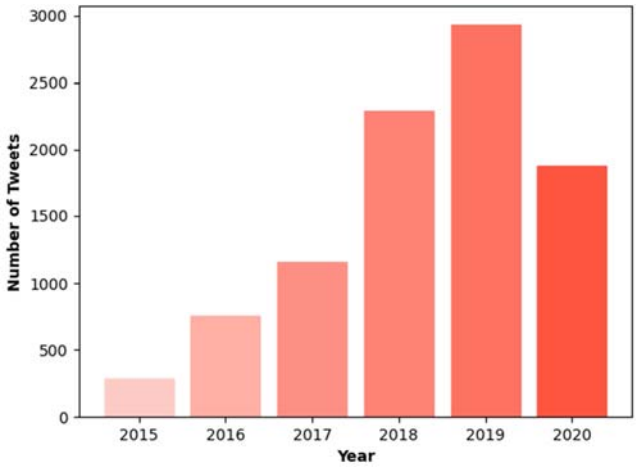

Figure 2. *Number of Tweets by Elon Musk.*

In Figure 2, the columns represent the number of tweets made by Elon Musk each year from 2015 to 2020. Overall, the graph shows an upward trend. The dip in 2020 can be explained by the fact that the dataset only accounted for tweets in 2020 up until July; despite this "setback", the number of tweets is higher than that from 2017 (just three years before), and it seems to have been on track to surpass the tweet count from the previous year. Based on this result, we can predict that 2020 will also join the upward trend made by the past years.

In Figure 3, the graph illustrates Elon Musk's twitter engagement over the years. Engagement in this scenario includes replies, retweets, and likes on Twitter. His engagement early on, from 2015 to 2017, seems to be limited. To be specific, he was not replying to as many comments, retweeting as many posts, or receiving as many likes on his tweets. However, as 2018 approached, his engagement,

especially his likes, skyrocketed. His likes seem to be highly correlated with the number of tweets he posted, as shown in Figure 2.

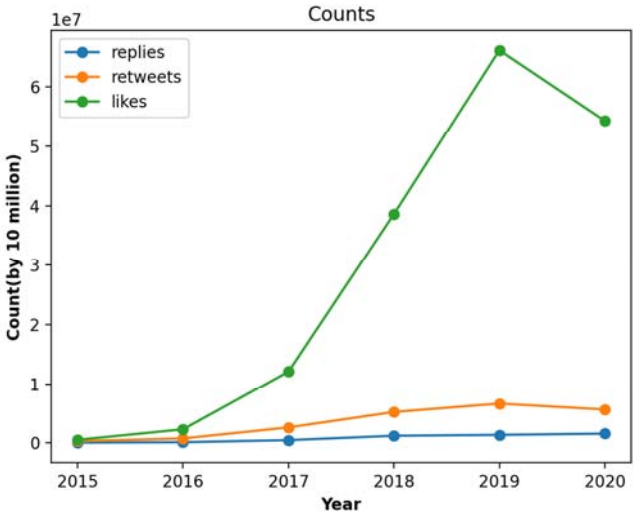

Figure 3. *Graph of Elon Musk's twitter engagement.*

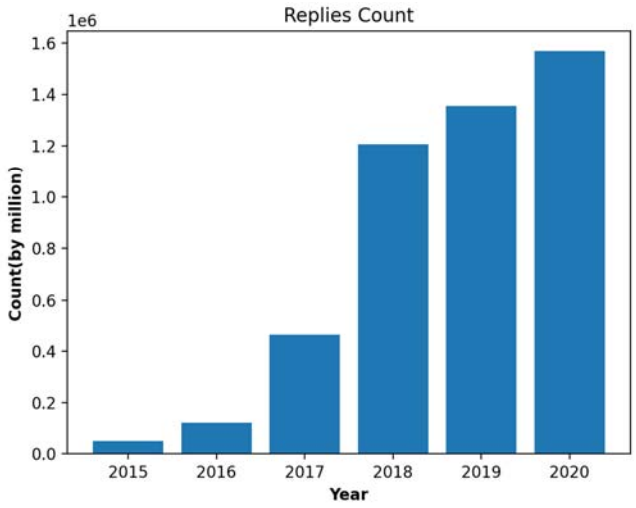

Figure 4. *Elon Musk's replies count illustrated.*

Figure 4 zooms in on the "replies" line from Figure 3, visualizing it in the form of a bar graph. While it may not seem as though there was any change at all in the previous figure, Figure 4 clearly depicts the rise in Elon Musk's Twitter reply count around 2018. This, once again, corresponds with the sudden rise in both tweets and likes of Elon Musk's after 2018, strengthening the correlation in Twitter engagement around this point in time.

In Figure 5, to explicitly display the frequency of words employed by Elon Musk, we used a word cloud, which is conventionally utilized to facilitate visualization of key words. In the word cloud above, it can be seen that the word SpaceX is used the most, 36 times; followed by NASA, 18 times; Space Station, 15 times; TeslaMotors, 10 times; and Hyperloop, 8 times. Other than these keywords, the frequency of words used seems relatively similar.

Figure 5. *Elon Musk's reply count.*

2.3. Further Analysis of Twitter and Stock Market

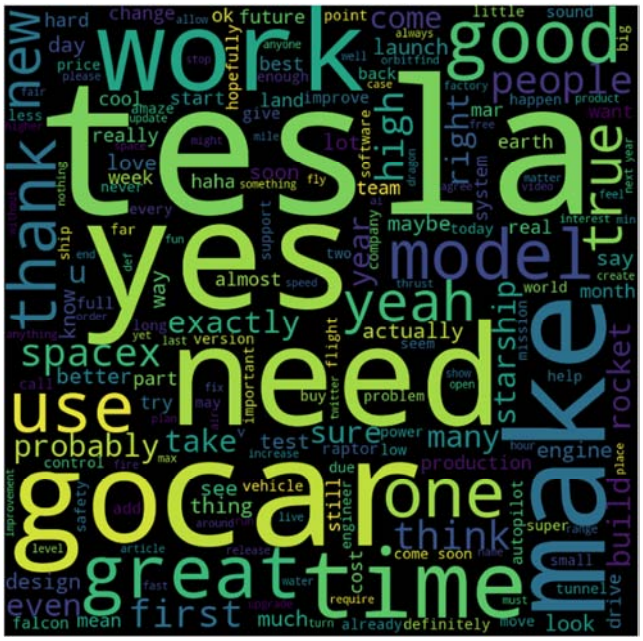

Figure 6. *Word Cloud with Musk's tweets.*

As mentioned earlier, Figure 6 is what is referred to as a Word Cloud. Using Python, we've gathered all tweets that Elon Musk posted from 2015 to 2020 and neatly organized these tweets into a dictionary in Python. With this information, it is now possible to preprocess the data. Since there is a dictionary of all the tweets Elon Musk posted over the years, we can identify any nouns, verbs, adjectives, and adverbs that Elon Musk frequently uses. This Word Cloud is useful since it not only gives us a clue about Elon Musk's favorite words, but it also gives us an insight into Elon Musk's personality. Understanding his audience and the type of content he posts on his Twitter allows us to assess data with increased accuracy and find other correlations.

2.4. Relationships Between Tweets & Stocks

As shown in Figures 7 and 9, there seems to be a high correlation between the number of tweets Elon Musk posted and his engagement during July of 2020. Especially after the

high amounts of tweets he posted from the start of July to July 5th, there was a noticeable spike in his engagement all around. Now, in contrast, how this affects the price of his Tesla shares still remains unclear for the most part. The relationship between the increasing of the share price and engagement isn't strong enough to draw a conclusive connection between them. However, it has been noted that every time there is a small dip in the prices of Tesla shares worldwide, the number of tweets and the engagement also experience a dip. With that being said, the vice versa is not as consistent: a dip in engagement and tweets doesn't always lead to a dip in the closing price. Nonetheless, it seems that after such dips, an excessive number of tweets with positive engagement effectively "pulls the stocks out of the dip", though the correlation is most likely insignificant. This form of distinguished correlation shows that a tenuous link between engagement and stock shares, while the link between engagement and tweet count is quite parallel.

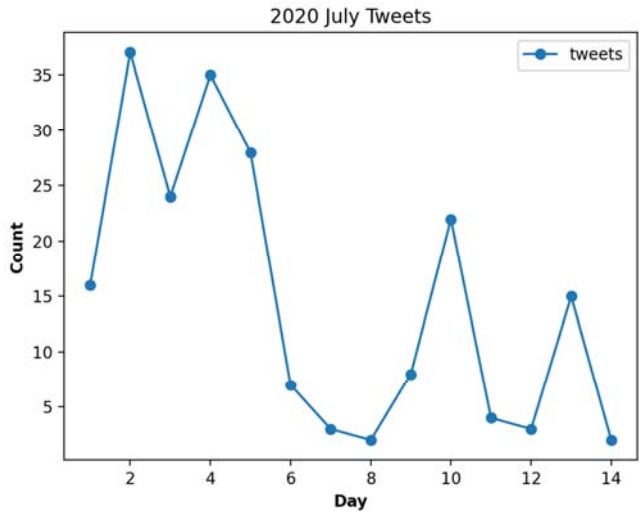

Figure 7. *July 2020 Tweet count.*

Figure 8. *July 2020 Closing Price.*

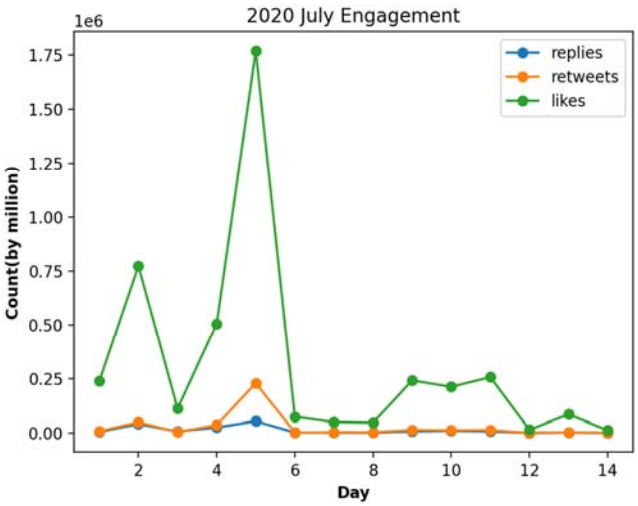

Figure 9. *July 2020 Engagement.*

2.5. Additional Supplementary Findings

With regards to our previous findings, we extrapolated the results to gain further insight into this correlation.

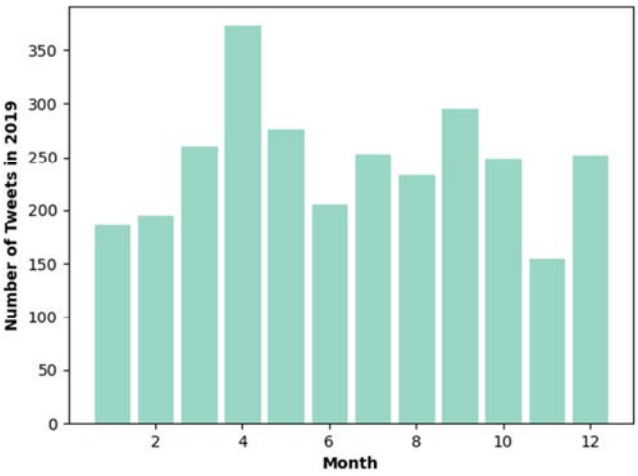

Figure 10. *Number of Tweets per Month in 2019.*

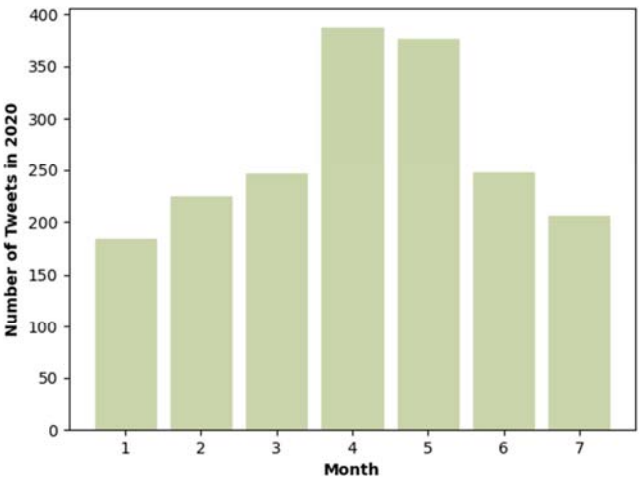

Figure 11. *Number of Tweets per Month in 2020.*

Figure 10 represents the number of tweets made by Elon

Musk every month typically in the year 2019. April has the highest number of tweets and November has the lowest number.

Figure 11 represents the number of tweets made by Elon Musk per month until July of 2020. April and May have the highest number of tweets.

As shown by both Figures 9 and 10, Musk's twitter shows a sharp growth sometime around April, the fourth month. As such, it can be concluded that, using the previous results, the closing prices of Tesla stocks will most likely rise in the month of April, as well as the days leading up to it/after it. With that being said, the reason why April shows such a high number of tweets is not clear.

3. Preprocessing

Preprocessing was a technique employed to help break down the dataset, by removing terms known as "stopwords". Simply put, stopwords are a set of commonplace words that add no real depth to the message or meaning of a sentence (ex. for, at, to). The removal of these terms was necessary to focus on the more important words, ones that may aid in finding a link between Elon Musk's twitter and the Tesla stock market.

First off, we imported a module known as NLTK (Natural Language Toolkit), which had a built-in stopword dictionary. This allowed for easy downloading of all existing stopwords into our program.

Next, we converted all existing tweet words into lower-case, since a word with a capitalized first letter was interpreted differently from a word that was full lower-case by the program.

This was then followed by a series of tokenizing within which all stopwords were effectively removed from the tweet data. *Tokenize* in Python refers to the "splitting up [of] a larger body of text into smaller lines [and] words [...]." [9].

Afterwards, we utilized a function called *lemmatize*, which remodeled words into their base form (ex. tried → try). This way, finding the repeat count for a certain word became easier to organize. Quite similarly, *stemming* was utilized as well, which, like *lemmatize*, is used to reduce certain words into a base form. These two functions differ in the way they carry out this reduction, however. While *lemmatize* dismantles a word into the corresponding basic form, *stemming* does not necessarily do the same. Instead, it crudely cuts off part of the word by following a specific set of stemming rules. For instance, while the word "ponies" would be transformed into the word pony by *lemmatize*, it would alternatively be turned into poni by *stemming*, following the rule IES → I [14]. Such subtle variations may cause errors and flaws within the data, so each function must be used accordingly and appropriately.

The task of preprocessing ended off by applying and compiling all these newfound changes into a new csv file: *elonmusk_preprocessing_data.csv*.

4. Machine Learning

Machine Learning (ML) was a major form of computing used in this research. In simple terms, ML refers to the computing method in which a processor improves in its accuracy through experience and trial-and-error, paralleling the way a human brain may think. Machine Learning is often used among data scientists to build software that can improve easily through the constant feeding of datasets and algorithms.

4.1. Machine Learning Classification Algorithms

This section covers various well-known algorithms that are often implemented within Machine Learning. These algorithms are essentially the basis of the code, as the machine 'trains" its intelligence by using the inputted algorithm. Different algorithms are built for different types of Machine Learning; for instance, one algorithm may be suited for classification while one may be suited for prediction. Different methods will present different accuracies, so to figure out a method that is going to analyse the sentiment with the highest accuracy, we have to be aware of the mechanics behind these algorithms to identify which algorithms to use next.

4.2. Sentiment Analysis

Sentiment Analysis, otherwise referred to as Opinion Mining, is a field of study dealing with humanities and linguistics, possibly coupled with data science and artificial intelligence. The main premise of Sentiment Analysis is to take a set of linguistic data (which in this case is Elon Musk's Twitter) and identify the emotion and subjective feel behind those words. Machine Learning is one way to conduct Sentiment Analysis; as a program reviews large datasets of text, it can learn to distinguish human emotions. Sentiment Analysis was used because we hypothesized that emotions in the tweets can also correlate with the stock prices of Tesla.

Sentiment Analysis is used everywhere in our daily lives. Advertisements, campaigns, and commercials all involve Sentiment Analysis one way or another, therefore carrying pathos with it. Our research made substantial use of Sentiment Analysis by taking the subjective state behind Elon Musk's tweets/replies and applying them to Tesla's stock, uncovering new links between the former and the latter.

Each tweet on Twitter was categorized according to its predicted emotional language into three groups: positive, neutral, and negative. This was done using scikit-learn software via Python. This machine learning was conducted based on preprocessed tweets, which made up the text column of the csv file, and the results were hand-inputted into the type column of the csv file. Through the use of such techniques, we were able to link Musk's emotions online to the prospective stock market fluctuations of Tesla and have a deeper look into our hypothesis

4.3 Sentiment Analysis Model Section

To create a dataset of tweets so the machine can learn from it, we proceeded to manually categorize about 1700 tweets into three groups: Neutral, Positive, and Negative. As the name implies, each category referred to the underlying tone of Elon Musk's tweets. If his tweet was overall happy or

lighthearted, for instance, it would be categorized as Positive. If it had a gloomy or angry tone, it would be considered Negative. The tweets that would be categorized as neutral would be tweets regarding his companies or tweets meant to inform the general public about subjects of his choosing.

Figure 12 is a pie chart we created using Matplotlib in Python to visualize Elon Musk's overall attitude in his tweets. In the pie chart above, we can see that Elon Musk's tweets are dominantly neutral, almost covering up to half of his tweets. Positive tweets easily outnumber negative tweets and through this data, we can get a better grasp on his personality.

To categorize the rest of his tweets, we applied Machine Learning algorithms to automate the process. As previously mentioned, scikit-learn was used to import and utilize each algorithm. The TfidfVectorizer feature was implemented to "transform" a certain word or phrase into a number rating, which would then be directly used to quantify and compare the three sentiments: Positive, Negative, or Neutral [15]. Next, different algorithms were imported and used, using the ttsplit, model. fit, model. predict, and accuracy_score function. The data went through an 80:20 split, in which 80% of the data was used for training and the other 20% was used for testing. Simply put, the program would attempt to correctly categorize 20% of the data (without having seen them) using the other 80% of examples (having seen them).

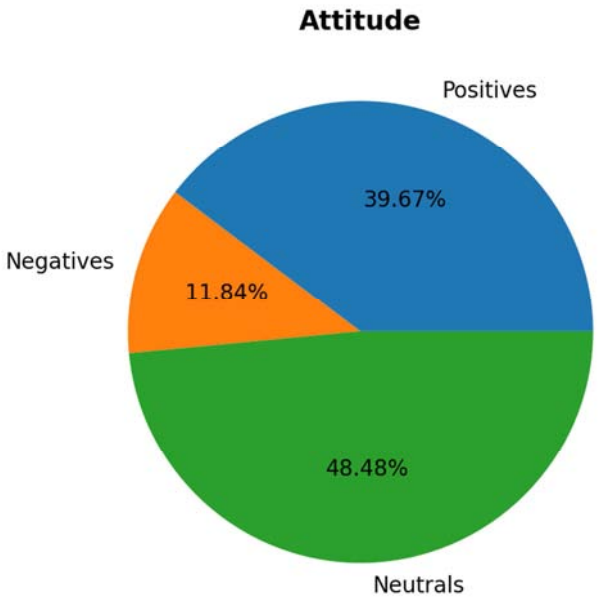

Figure 12. *Sentimental analysis on Musk.*

Here is an explanation as to how some of the more popular algorithms work. Logistic Regression is a probability-based algorithm that relies on the logistic growth model of mathematics. Depending on how close the input is to 0 or 1, the algorithm replicates its answer accordingly [10]. Decision Tree, another well-known algorithm, utilizes a strategy resembling a flowchart; two or more options are given, and depending on the option chosen, a series of new choices are presented until a final option is reached, which is going to be the output [11]. *Naive Bayes* [12], like logistic regression, is a probability-based algorithm, which relies on the Bayes Theorem. In simple terms, this algorithm takes the likelihoods and variables of the dataset and inserts them directly into Bayes' equation [13].

A total of six algorithms were used: Logistic Regression, Decision Tree, Naive Bayes, Random Forest, Gradient Boosting Classifier, and Support Vector Machine. Using the functions mentioned above, the algorithm was run ten times; the final accuracy was the mean of all 10. The accuracies for each were approximately 68.0%, 63.2%, 66.9%, 68.3%, 64.1%, and 95.6% respectively. Therefore, in the end, the Support Vector Machine algorithm was used due to its highest display of accuracy.

Algorithm	Accuracy(%)
Decision Tree	63.2
Gradient Boosting Classifier	64.1
Logistic Regression	68.0
Naive Bayes	66.9
Random Forest	68.3
Support Vector Machine	95.6

Figure 13. *Observed accuracy.*

Figure 13 is a table that consists of different algorithms and their accuracies. From this figure, we can safely say that the Support Vector Machine trumps all the other algorithms in terms of accuracy. Though the Support Vector Machine takes more time returning outputs, it has the highest accuracy among the others. Therefore, this algorithm will be mainly used to identify emotions behind Elon Musk's tweets.

4.4. Data on 2020/21 Tesla and Twitter

When putting our data of 2020 through Sentiment Analysis (albeit only the first seven months), we found that 33.9% of the data was positive, 11.5% was negative, and 52.1% was neutral on average. The remaining 2.5% consisted of unclassifiable tweets. Even though the data was spread out throughout each month, the positive distribution formed a U-shaped curve while the neutral formed an upside-down U (∩). As for the negative, there was a rough representation of a U-shape as well but it was not distinct enough to categorize it as such.

A look at the closing price of Tesla during the first nine months of 2020 revealed an exponential graph with an upward trend; knowing the trends of positive and neutral tweets as mentioned above, a correlation could be drawn.

We could organize the data of 2020 July's tweets into 60 neutral comments, 94 positive comments, and 32 negative comments. We noticed the correlation between the stock price and the tweets made on Twitter in which the number of

positive tweets increased with the increase in stock price.

Next, we performed Sentiment Analysis on a small portion of Elon Musk's tweets in 2021 to make sure our dataset was as fresh and relevant as possible. This included tweets from January and the first few days of February.

From a dataset consisting of 58 tweets, 21 of them were classified as positive, while the remaining 37 were negative. Furthermore, these estimates were verified and double-checked by hand to make sure that our dataset was as accurate as possible.

Subsequently, we checked the Tesla stock in 2021. From January 1st to February 1st, the stock market increased by a factor of roughly 20%, just like the previous year. We could then compare this with other compiled data to perhaps detect a correlation.

5. Conclusion

In sum, our research was able to utilize a multitude of functions to attain our primary objective of finding a distinct relationship between Elon Musk's Twitter and Tesla's stock value.

To obtain this goal, we exploited a variety of implements to extract certain data.

By scrutinizing the frequency of replies or tweets, and the stock value varying through time, we found that in the short run, the number of tweets Elon Musk posted and his engagement marginally correlates with the stock price of Tesla. However, looking at the data, in the long run, the correlation becomes more apparent. In other words, the correlation can be more evidently viewed when the tweets were analyzed by months or years rather than days. By analyzing all the data, we established that fluctuations in the closing price of Tesla had a direct, parallel correlation to Musk's engagement.

Our research is valuable to the public because it shows that Elon Musk's Twitter engagement is a clear indicator of stock value growth. Though not completely accurate, the insight that the tweets bring is worthwhile to note before taking any actions.

However, what we acknowledge in our research is that it lacks comprehensive information. In other words, it is not definitive that the correlation discovered in Tesla applies to other companies in general. This shortcoming, therefore, needs further investigation for our research to enhance practicality.

References

[1] Pendleton, Devon. "Elon Musk Overtakes Bill Gates to Grab World's Second-Richest Ranking." *Bloomberg.com*, Bloomberg, 24 Nov. 2020.

[2] "Elon Musk Biography." *Biography.com*, A & amp; E Networks Television, 17 Nov. 2020.

[3] Ross, Sean. "Elon Musk's Best Investments." *Investopedia*, Investopedia, 8 Sept. 2020.

[4] O. Vynakov, E. V. Sa, and A. I Skryn "Modern Electric Cars of Tesla Motors Company", Automation Technological and Business Processes.

[5] Schreiber, Barbara A. "Tesla, Inc." Encyclopædia Britannica, *Encyclopædia Britannica*, Inc., www.britannica.com/topic/Tesla-Motors.

[6] Tahir M. Nisar and Man Yeung, "Twitter as a tool for forecasting stock market movements: A short-window event study", *Science Direct*, Feb 2018.

[7] Emanuele Teti et al., "The relationship between twitter and stock prices. Evidence from the US technology industry", *Science Direct*, Nov 2019.

[8] "Exploratory Data Analysis." *Carnegie Mellon University Statistics & Data Science*, www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf.

[9] "Python - Tokenization." *Tutorialspoint*, www.tutorialspoint.com/python_text_processing/python_tokenization.htm.

[10] Pant, Ayush. "Introduction to Logistic Regression." *Medium*, Towards Data Science, 22 Jan. 2019.

[11] H. Patel and P. Pra "Study and Analysis of Decision Tree Based Classification Algorithm, International Journal of Computer Sciences and Engineering, Oct 2018.

[12] F. Qin, X. Tan, Z. Cheng "Application and research of multi_label Naïve Bayes Classifier, Proceedings of the 10th World Congress on Intelligent Control and Automation, July 2012.

[13] "A. Wibawa, A. Kurn, D. Murti, R. Adi "Naïve Bayes Classifier for Journal Quartile Classification, June, 2019.

[14] T. Koren, K. Jarv, J. Lau, M. Juh "Stemming and Lemmatiztion in the clustering of Finnish text documents", January 2004.

[15] S. Qaiser and R. Ali "Text Minig: Use of TF-IDF to Examine the Relevance of Words to Documents", July 2018.