International Journal of Data Science and Analysis
2021; 7(3): 76-81
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20210703.14
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Scientific Data Analysis: Employing Sentimental Analysis to Prove Correlation Between Social Media and Electric Vehicles in Modern Society

Sungjoon Cho

Cheongshim International Academy, Gyeonggi, Republic of Korea

Email address:
seongjun2615@gmail.com

To cite this article:
Sungjoon Cho. Scientific Data Analysis: Employing Sentimental Analysis to Prove Correlation Between Social Media and Electric Vehicles in Modern Society. *International Journal of Data Science and Analysis*. Vol. 7, No. 3, 2021, pp. 76-81. doi: 10.11648/j.ijdsa.20210703.14

Received: April 26, 2021; **Accepted**: May 12, 2021; **Published**: May 31, 2021

Abstract: In recent decades, the development of technology has brought several changes in the global society. Enhanced communication methods enabled rapid dissemination of information, impacting peoples' decision making and consumption. Moreover, indiscreet production and resource consumption caused environmental damage, hence leading to the advent of electric vehicles in the automotive industry. This research paper delves into the influence of social media on market share and stock prices of electric vehicle manufacturers. Social media plays a significant role in conveying information and therefore influencing consumption. To conduct research, we gathered data – tweets, news articles, EV stock prices, EV market shares, air quality of major cities – to prove correlation between social media and EV stock prices. Market data were mainly used for analysis and prediction, and information regarding air quality was used to explain how electric vehicles could gather huge momentum. We analyzed how electric vehicle market shares have changed in 10 years, and how individual manufacturers, such as Tesla, General Motors, and Hyundai, increased production and sales over time, using data analysis and visualization. By comparing these data with media coverage of electric vehicles using sentimental analysis, we could figure out how social media could impact sales and stock prices of automotive producers. The main driving force of the meteoric rise of electric vehicles was favorable media coverage of electric vehicles. Data collection was done by effective Python tools that could significantly reduce time.

Keywords: Electric Vehicles, Sentimental Analysis, Machine Learning, Data Analysis

1. Introduction

In recent decades, the global society increasingly has started to pay more attention to the environment and sustainable development [1]. Environmental destruction and global warming occurring at unprecedented pace raised awareness amongst people, leading to collective efforts to preserve nature and turn to renewable energy. Internal combustion engine (ICE) vehicles using fossil fuels especially have proven to be detrimental to the environment, emitting air pollutants and creating noise pollution. Hence, electric vehicles have risen as alternatives to ICE-vehicles and are considered to be the future of automobiles. Such meteoric rise of electric vehicles has caused stock prices of electric vehicles manufacturers to soar. Tesla's stock prices skyrocketed in the last decade, making Tesla not only a leader in electric vehicles production but also one of the most influential, admired, and famous corporations in the world.

Sentiment analysis, also known as opinion mining, is a natural language processing technique that is frequently used to identify and extract subjective information in diverse source materials [2], [3]. In other words, it involves the process of determining whether a piece of writing contains positive, negative, or neutral emotions. Sentiment analysis allows businesses to garner public opinion to understand public awareness of the brand and monitor brand or product reputation. Nowadays, due to the advancement of technology, it has become easier than ever to express one's own thoughts through social media. Therefore, companies can easily gain customer feedback and make adjustments to meet customer needs. [4], [5] Throughout the research, sentiment analysis was used to determine the overall tone of news articles

featuring electric vehicles and Elon Musk's tweets.

This paper will analyze how the electric vehicle market has over time. In doing so, Exploratory Data Analysis (EDA) was used to interpret numerical data - electric vehicle market shares, and stock prices of major electric vehicle manufacturers. Based on current market shares and trends, predictions on how electric vehicles will continue to garner attention in the upcoming future were made.

Nowadays, politicians, entrepreneurs, celebrities, and sports athletes commonly use social media to communicate with the public. This research paper employs sentiment analysis to analyze tweets posted by Elon Musk, the CEO of Tesla, and news reports covering electric vehicles. The results combined with changes in stock prices of electric vehicle producers could be employed to deduce how electric vehicles would gain momentum and impact the economy in the future.

2. Data Preparation

2.1. Data Collection - Web Crawling

Data collection was done by a process known as web crawling. We gathered news articles featuring electric vehicles from 2020 to 2021, and tweets posted by Tesla CEO Elon Musk, since these data would be used for sentiment analysis to prove correlation between social media and stock prices of EV manufacturers. Web crawling allows users to "scrape" the web and gather plenty of information without necessarily having to search them one by one. In other words, users can parse HTML and XML pages simply by typing in the keyword. Using the 'Beautiful Soup' package from Python, it was possible to gather information, namely headlines for news articles regarding electric vehicles. Data were separated by time span - first half of the year, and the second half. After that, gathered data were preprocessed for simplification.

2.2. Preprocessing

Preprocessing is an important step of transforming data before the user feeds it to the algorithm. In other words, it is the process of converting raw data into a clean and categorized data set so that the computer can understand the data. In general, data preprocessing is divided into two categories: selecting data objects for analysis, and creating or changing the attributes. Data is often incomplete, inconsistent, and prone to error. Data preprocessing addresses such problems and allows further processing. [6] Throughout the research, preprocessing was used mainly to convert between lower/upper case and remove parts using the 'strip' function in python. This process was necessary as the inclusion of unnecessary data could alter the results.

In deep learning, to understand the meaning of text, it is important that we split the text into smaller parts. Tokenization refers to the process of splitting text into smaller units with semantic meaning, tokens. Tokenization can occur at the word level, character level, and the subword level. [7]

Stop words, such as "the", "a", "an", are words that a search engine has been programmed to ignore; they do not contain significance in meaning. Since these words are unnecessary, they can be removed from occupying space in the database so that users leave the only important parts. [8] Data was converted into a more categorized data set by removing such stop words.

Preprocessing, including tokenization of the tweets and removal of stopwords from Elon Musk's tweets, was done through lines of codes using the Python language.

Lemmatization and stemming, often referred to as Text Normalization, are the techniques of finding the original form from derivationally related words [9]. In other words, they aim to obtain the root form from inflected words. They are similar, but are executed through different mechanisms.

Stemming uses heuristics - rule of thumb. To be more specific, it removes suffixes or prefixes from a word [10]. The Porter Stemmer is the most effective and commonly used type for English stemming. Since stemming automatically removes the ends of words, results can be words that do not exist.

Lemmatization, on the other hand, uses the dictionary to find the base form of words. It entails morphological analysis of words to return the dictionary form of words, known as the 'lemma.'

Both stemming and lemmatization were incorporated in the research. These tools were used to organize Elon Musk's tweets and leave only the necessary parts, the meaningful parts that would be used for sentiment analysis. A new set of data, coined 'new_data', was created from the original data through preprocessing. (stemming, lemmatization, tokenization, removal of stop words).

3. Electric Vehicles (EV) in Today's Society

Exploratory Data Analysis, EDA, is used to analyze and investigate data sets through visualization. Histograms, Box plot, and Scatter plot are used for visualization. Through EDA, it is possible to spot anomalies, find patterns, and check assumptions [11], [12]. Throughout the research, EDA was primarily used for visualization and analysis on stock prices of different electric vehicle companies around the world - Tesla, General Motors, and Hyundai.

3.1. Electric Vehicle Manufacturers Stock Price

This section focuses on how stock prices of major electric vehicle manufacturers have changed over time to illustrate the impact of electric vehicles on the global economy.

As shown in figure 1, stock prices of Hyundai, a multinational automobile manufacturer in Korea, were visualized using the python language. Hyundai stock prices started to increase significantly in 2020. On January 1, 2017, Hyundai stocks were transacted at 146,000 KRW. That value increased to 271,000 KRW by January 20, 2021. Hyundai,

which mostly produces ICE automobiles, recently began to develop and invest in electric vehicles. As electric vehicles started to gain more and more popularity in the global market, Hyundai experienced exponentially growing stock prices as a result.

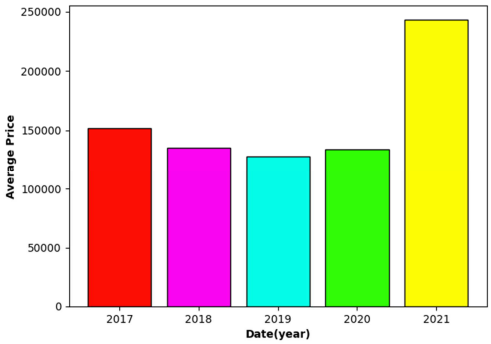

Figure 1. *Hyundai Stock Prices.*

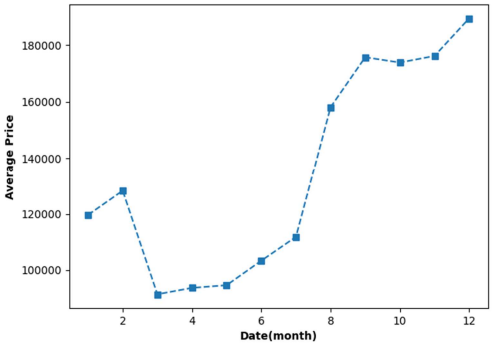

Figure 2. *2020 Hyundai Stock Price Changes.*

As mentioned above, Hyundai stocks started to increase significantly in value around 2020. Average prices decreased from 130,500 KRW on January 27th to 67,200 KRW on March 20th, as demonstrated in figure 2. However, a strong rally began in May; stock prices boosted, reaching 196500 KRW on December 4th of 2020.

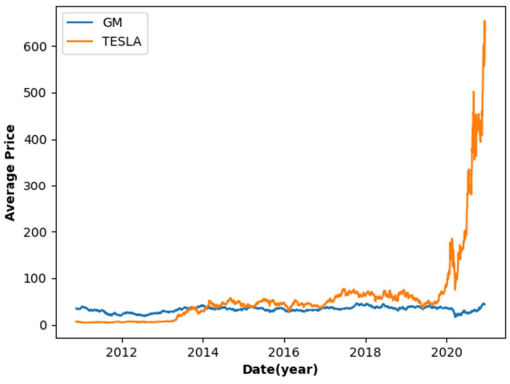

Figure 3. *Tesla, GM stock prices (2012 ~ 2020).*

As shown in figure 3, EDA was employed to visualize data sets for stock prices of automobile manufacturers. Stock prices for Tesla were compared with those of General Motors (GM). From 2012 to 2020, GM stock prices remained stagnant - prices were around 19~27 USD in 2012, and 17~37 USD in 2020. Although there were slight ups and downs, prices remained largely unchanged. In the meanwhile, Tesla stock prices increased overall. Tesla stocks were priced at 5~7 USD in 2012, which continued to increase over time. The figure above shows that prices increased exponentially around 2020. Tesla stocks skyrocketed in 2020, being transacted at 653 USD in December 2020. The results show how influential Tesla is in the global EV market, and how electric vehicles have gathered momentum over time. Tesla was arguably one of the earliest companies to mass produce electric vehicles, while GM joined the EV market relatively late. GM stock prices are projected to grow in the near future.

3.2. Electric Vehicle Sales & Market Share

This section illustrates how electric vehicle market shares have increased over time.

Table 1. *Electric vehicle market share (2010 ~ 2020).*

Market Share (Year)
0.014339942 (2010)
0.072637591 (2011)
0.173683548 (2012)
0.278237441 (2013)
0.424335462 (2014)
0.679583389 (2015)
0.904920529 (2016)
1.382505747 (2107)
2.289726036 (2018)
2.498946714 (2019)
3.179357465 (2020)

The table 1 above shows how electric vehicle market shares have changed from 2010 to 2020. In 2010, there were few electric vehicles being produced and sold. However, as people started to care more about the environment and sustainable development, electric vehicles rose as an alternative to traditional vehicles using fossil fuels. Therefore, electric vehicle sales gained momentum and stock prices for electric vehicle manufacturers increased as well. Toyota, one of the largest car manufacturers in the global market, produced only 912 electric vehicles in March 2012. In 2019, Toyota produced up to 42,345 electric vehicles. In short, increasing public awareness of global warming and environmental pollution created a new trend in the global automobile market; electric vehicles started to receive more attention than ever before, resulting in unprecedented increase in electric vehicle production and stock prices for car manufacturers - especially Tesla.

3.3. Air Quality

As mentioned earlier, ICE-vehicles emit air pollutants that contribute to overall air pollution and global warming. More environmental destruction would ultimately lead to more production and sales of electric vehicles. Air quality of large cities was analyzed to exhibit the relationship between air pollution and electric vehicle sales.

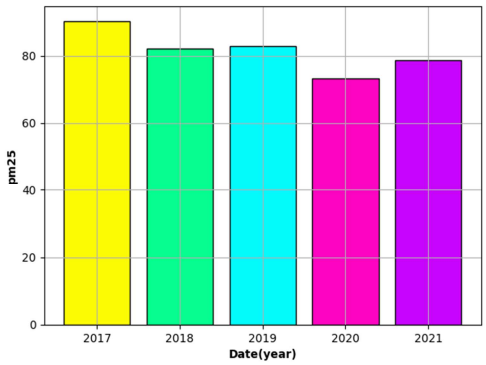

Figure 4. *2020 Seoul Air Quality.*

As shown above, figure 4 visualizes how Seoul's air quality changed over the past few years. The data shows how concentration of pm (particulate matter) 2.5 has changed over time. On average, the air quality index of Seoul based on pm 2.5 concentration has approximately been 70, being marked as "moderate" by the US-EPA 2016 standard - a measurement of air quality established by the US National Ambient Air Quality Standards. Moderate levels of air pollution (51-100) indicates that some pollutants may pose health concerns for people sensitive to air pollution and warns children and adults with respiratory disease to limit prolonged outdoor exertion. [13]

3.4. Elonmusk's Tweets

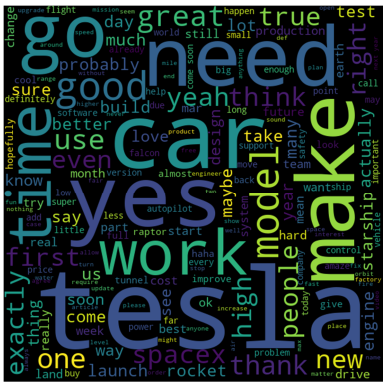

Figure 5. *Visualization of Elon Musk's tweets.*

Once again, using the Python language, it was possible to visually interpret Elon Musk's tweets after preprocessing. Based on Figure 2, it is possible to see that Elon Musk frequently uses words such as 'Tesla', 'yes', 'great', and 'work'. In general, he uses words that contain positive connotations and are related to his own business. Such positive posts played an important role in inculcating positive images about electric vehicles, eventually leading to more investment in the EV market and purchasing of EV manufacturer stocks.

4. EV Market Share Prediction

4.1. Algorithms

(a) Gradient Boosting

Gradient Boosting is a machine learning algorithm that makes use of decision trees. It is used for classification and determining the relationship between the dependent and the independent variable [13]. It builds predictive models by the adding of subsequent models to the ensemble, wherein the subsequent ones improve the performance of prior models.

(b) Logistic Regression

Logistic Regression is a statistical model used for calculating probabilities. It uses a logistic function for binary classification problems [14].

(c) Random Forest

Random Forest, as the name suggests, is a type of learning algorithm that builds a forest composed of decision trees [15]. Random Forest is also used for classification and regression. It operates by the combination of learning models to increase the overall result; decision trees based on sub-samples of the dataset and averaging are used to enhance accuracy.

4.2. Modified Sentimental Analysis

Sentiment analysis, or opinion mining, allows users to determine whether a piece of writing contains positive, negative, or neutral tone. Generally, sentiment analysis by definition focuses on the emotion; if a news article contains harsh, critical expressions, the overall sentiment would be 'negative.' From all the data collected, 80% was used to train the computer. The remaining 20% of data was used for testing. The results, however, were inaccurate, so we decided to analyze the sentiment of news articles from the perspective of an electric vehicle - from a more economic perspective. For instance, legislation that subsidizes electric vehicle manufacturers would definitely be considered positive.

4.3. Result

Using web crawling, data - news headlines featuring electric vehicles - were gathered and preprocessed using the python engine. Then, we categorized data based on chronology - EV news articles from 2020 and 2021. After that, we applied modified sentimental analysis from the perspective of an electric vehicle to figure out the overall tone. News headlines were labeled as either negative, positive, or neutral. The results were visualized in the form of bar graphs.

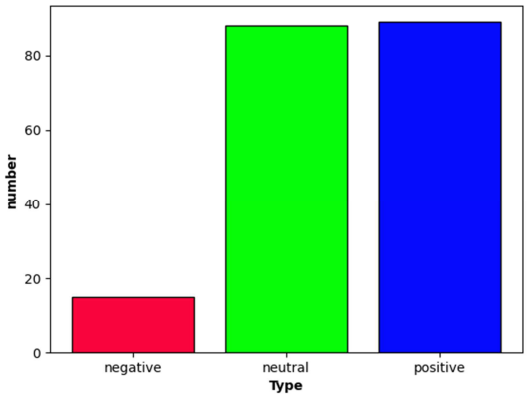

Figure 6. *Modified Sentimental Analysis of EV news articles in 2020.*

As demonstrated in figure 6, 7.81% of news articles were classified as 'negative', meaning that they conveyed a negative and pessimistic tone. 45.8% of the data was considered 'neutral', and 46.3% considered to be 'positive.' In general, an overwhelming portion of news articles featured balanced and positive attributes of electric vehicles, accounting for the increase in stock prices of ev manufacturers.

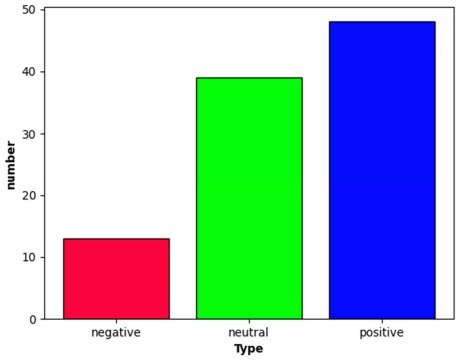

Figure 7. *Modified Sentimental Analysis of EV news articles in 2021.*

Figure 7 shows the results of modified sentimental analysis of EV news articles written in 2021. The results were quite similar to those of 2020; only 13% were considered to be negative. Figure 8, 9 shows visual interpretation of news articles reported in 2021, using the Word Cloud creator from Python. Positive articles included words such as "fund", "infrastructure", "enthusiasm", and "new dominion". Positive news articles mainly featured how governmental action such as the creation of new legislation, tax cuts, and funding could fuel further growth of the electric vehicle industry. Also, they reported on new research plans announced by EV manufacturers and how increasing EV infrastructures could entice more people to purchase electric vehicles. On the other hand, negative news articles were concerned about the imperfection of EV technology - battery fire sparks and general malfunctioning. Such articles frequently used the terms "threat", "miss", "fall", and "recall" to show that EV technology had flaws, discouraging consumers to choose EVs over Internal Combustion Engine vehicles.

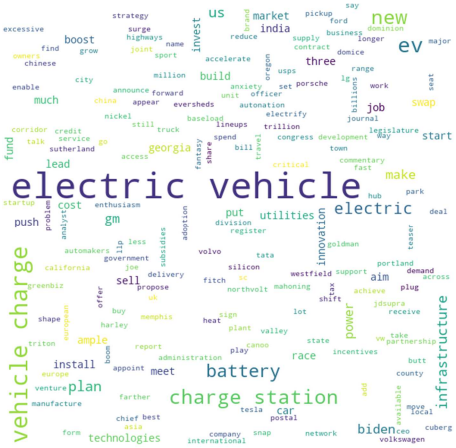

Figure 8. *Visualization of 'Positive' EV News Articles.*

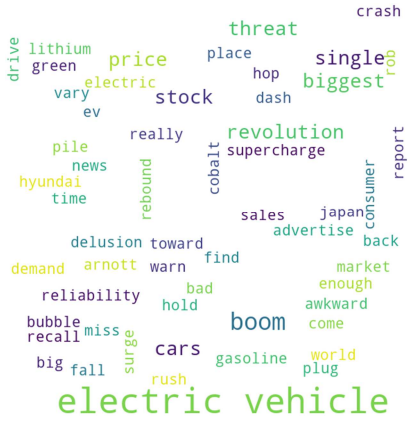

Figure 9. *Visualization of 'Negative' EV News Articles.*

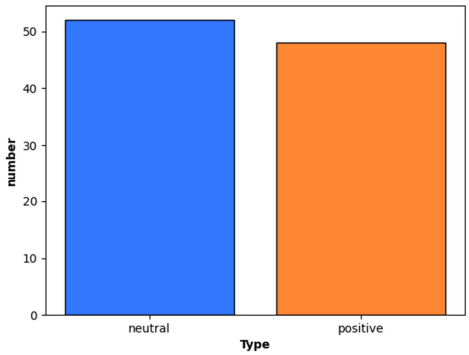

Figure 10. *Sentimental Analysis of EV news articles - Gradient Boosting.*

After analyzing news articles from the perspective of an electric vehicle, we used a computer model using gradient boosting to compare the results. Data collection was done by web crawling. The results showed that 52% of all data was neutral, and the remaining 48% contained a positive tone; none of the data was classified as negative. However, based on the perspective of an electric vehicle, 4 of the news articles headlines could be considered negative. Another computer model using the random forest algorithm showed different results. The amount of neutral data was increased to 63%, and positive data was reduced to 37%. Overall, media coverage of electric vehicles were friendly and supportive in recent years, accounting for the skyrocketing increase of stock prices and sales of EV producers.

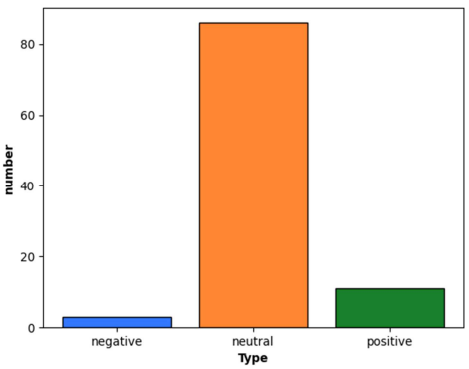

Figure 11. *Sentimental Analysis of Hyundai news articles - Gradient Boosting.*

Figure 9 demonstrates the results of sentimental analysis using the computer model; we initially used the gradient boosting algorithm to yield results. 86% was neutral, 11% positive, and only 3% considered negative. Nevertheless, from an electric vehicle's point of view, 6 additional data could be classified as negative. Articles that were classified negative mostly featured recall of vehicles due to battery fires. Using the random forest algorithm instead of gradient boosting, we could yield different results. Based on the random forest algorithm, none of the data was negative. Instead, the amount of positive data was increased to 24%, and neutral data reduced to 76%. As illustrated in the negative articles, some people were skeptical and cautionary of safety issues. Nevertheless, social media in general portrayed electric vehicles as innovative, practical, and eco-friendly. Hence, more and more people invested in EV manufacturers like Tesla and Hyundai, allowing EV companies to increase sales and revenue.

5. Conclusion

Based on the findings, it would be possible to conclude that social media played a significant role in the rise of EV stock prices in recent years. Favorable news coverage and tweets by influencers, as shown by sentimental analysis, formed positive public opinion about electric vehicles and led to the escalation of EV stock prices. Public awareness about environmental preservation also contributed to the rise of stock prices. Social media has heavily influenced the way people think and behave, and led to market transformations. It will continue to impact our lives in the upcoming future.

References

[1] D. Romero, and A. Molina, "Towards a sustainable development maturity model for green virtual enterprise breeding enviornments," Proceedings of the 19th IFAC World Congress, 2014.

[2] D. Den, "Analyzing Scientific Papers Based on Sentiment Analysis", Cairo University, January 2016.

[3] B. Norm, E. Lett, and C. Villegas, "Sentiment analysis and opinion mining applied to scientific paper reviews", Intelligent Data Analysis, February 2019.

[4] S. Gupta, "Sentiment Analysis: Concept, Analysis and Applications," Toward Data Science, 2018.

[5] "Sentiment Analysis Explained", Lexalytics.

[6] V. Agar, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis", International Journal of Computer Applications, 2015.

[7] A. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification", International Journal of Computer Science and Information Security, June 2018.

[8] W. Wilbur, "The automatic identification of stop words", Journal of Information Science", 1992.

[9] L. Skork, "Application of Lemmatization and Summarization Methods in Topic Identification Module for Large Scale Language Modeling Data Filtering", 15th International Conference, 2012.

[10] A. Jivani "A Comparative Study of Stemming Algorithms", The Maharaja Sayajirao University of Baroda, 2011.

[11] C. Yu, "E, xploratory data analysis in the context of data mining and resampling", International Journal of Psychological Research, June 2010.

[12] C. Moral, A. An, R. Imbert, J. Ramirez "A survey of stemming algorithms in information retrieval", Information Research, March 2014.

[13] S. Mal, R. Har, and A. Kun, "XGBoost - A Deep dive into Gradient Boosting (Introduction Documentation)", Medium, Feburary 2017.

[14] J. Peng, K. Lee, and G. Ing, "An Introduction to Logistic Regression Analysis and Reporting", The Journal of Educaitonal Research, Septemper, 2002.

[15] J. Ali, R. Khan, N. Ahmad, I. Maq "Random Forests and Decision Trees", International Journal of Computer Science Issues, September, 2012.