ASHESI UNIVERSITY

PREDICTING DIGITAL ENGAGEMENT FROM SOCIAL MEDIA

UNDERGRADUATE THESIS

B.Sc. Management Information Systems

Afua Fosua Ayiku

2022

ASHESI UNIVERSITY

PREDICTING DIGITAL ENGAGEMENT ON SOCIAL MEDIA

UNDERGRADUATE THESIS

Undergraduate Thesis submitted to the Department of Computer Science, Ashesi University, in partial fulfilment of the requirements for the award of the Bachelor of Science degree in Management Information Systems

Afua Fosua Ayiku

2022

DECLARATION

I hereby declare that this Undergraduate Thesis is the result of my own original work that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature:

……………………………………………………………………………………………………

Candidate's Name: Afua Fosua Ayiku

……………………………………………………………………………………………………

Date: May 13, 2022

……………………………………………………………………………………………………

I hereby declare that the preparation and presentation of this applied project were supervisedin accordance with the guidelines on supervision of applied projects laid down by Ashesi University.

Supervisor's Signature:

....

Supervisor's Name: Dr. Stephane Nwolley

....

Date: May 13, 2022

....

Acknowledgements

This project was especially testing as there were many doubtful nights nearing the time of submission. However, I would like to acknowledge God for providing me with strength, knowledge and understanding and success. To my parents and grandma who have supported me throughout my education thus far. To all the people who constantly checked up on me and provided help in any way possible. To my roommates who provided all their support and to Tobias Woode who was an emotional support and my sense of reason in moments of cloudiness; Thank you all.

For all the technical support given through the minds of Ernest Amenyedzi, Ekiyor Odoko and Derek Kweku and my supervisor Stephane Nwolley, I give a special thank you.

Abstract

The purpose of this research is to explore the adoption of a predictive machine learning model for business digital engagement. However, the focus aims to test the hypothesis of whether or not machine learning regression models can be used to effectively predict social media engagement metrics. The exploratory research this project undergoes consists of three data analysis experiments, filtered by factors such as the size of the data set and the inclusiveness of outlier data elements. At the start of the project, Instagram and Twitter were both considered as data sources, but revised data privacy policies prevented a data collection process within scope, for Instagram specifically. The following models: Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor and Artificial Neural Network was trained and tested on Twitter data. Of the three experiments conducted, the third experiment consisting of the larger data set and removed outliers proved to be the most effective. Though the predicted results are not accurate enough to be replicated across several edge scenarios.

Table of Contents

Chapter 1: Introduction

1.1 Background

Online Social Networks (OSN) has become one of the core fundamentals of an average person's day. Its usefulness and extent continue to grow, both on the positive spectrum and negative spectrum. Apart from people interacting with OSNs for leisure, the presence of businesses has also become a large part of this growing phenomenon. Due to the advantage of networks that these OSNs provide, enterprises have leveraged this asset to increase brand awareness. The growth of these networks has caused businesses to adopt social media to gain more customers, fuel brand awareness, address complaints and respond to enquiries. The integration of electronic commerce into traditional companies can be attributed to the benefits that arise from such an adoption, such as lower costs, wider reach, increased revenue, efficient customer service and advanced operational processes [1]. Just like electronic commerce, the resoundingly beneficial effects of machine learning have also been utilised in the traditional commerce space. The intersection of these technological advancements has caused an understanding of consumer behaviour by tailoring search pages to meet consumers' needs or predicting churn rates for businesses. The increase in online retailing therefore makes it imperative for companies to employ techniques that are tactful and 'decrease the cost of the sale network' [2]. Though OSNs do not necessarily offer the typical access to commerce that a website can offer, they still provide a valuable avenue to businesses as they can comfortably engage with their customers at no extra cost. Apart from the effectiveness of social media in driving sales, another significant advantage offered is the interactivity with customers. The genesis of such interaction is fostered through the creation of posts. [28] stated that 93% of marketers search for ‘increased exposure’ as the main goal for social media usage. Posts on social media networks deliver on this desired outcome as they can offer value in unique ways by attracting customers and initiating interaction and engagement [25]. Through these means, posts can serve as an advertising avenue.

Due to the heightening development of online retailing, there has been a surge in businesses that advertise their services or products online. Therefore, there is a need for businesses to create advertisements and content which drives consumers to the much-desired goals of the company, be it through an increased click-through rate, a direct sale, or an increase in knowledge. Using social network sites as a form of advertising can often be more effective than television advertisements. These forms of advertising are less intrusive because the consumer controls the consumption rate. Such forms of these advertisements offer a higher level of interaction, and these advertisements are served within a relaxed context [3]. The use of social media as an advertising space has increased more in recent years. Still, this increase has also led to dissatisfaction in consumers as most popular social media sites are populated with numerous adverts. The digital marketplace has three distinctive features that are equally beneficial and potentially damaging to businesses. The features include the growth of sources and media materials, the ease of access to these various media by consumers and the limited supply of human attention to consuming this massive number of media [10]. It is then clear that an enormous availability of media makes social media an overhaul of information [9]. Though there is a vast number of media, consumers have limited time to consume all of it [10]. Due to this phenomenon, it becomes imperative for businesses with a social media presence to optimise their approach to building a brand lest they risk being entrenched in the already existing mass of information.

This persistent problem gives rise to the research paper, which aims to predict the digital engagement of OSNs with machine learning. This paper considers two social media outlets: Twitter and Instagram, to thoroughly explore digital engagement. To achieve this via Twitter and or Instagram, past posts from businesses will be scraped from either platform to predict engagement (number of retweets or the number of likes) for future posts.

With the numerous social media platforms to choose from, businesses should understand how well a future post may potentially do. Ideally, a digital marketer will post content across various social media platforms like Facebook, Instagram, Twitter, or LinkedIn. It is rare to see a post on one platform which is not replicated on another one. However, due to the nuances associated with each social media platform, this paper focuses on only Twitter and Instagram as the test case.

To fulfil this research, social media posts of businesses in the consumer goods industry will be trained with varied regression machine learning models and tested to reveal the combination of post characteristics that give rise to higher digital engagement. Digital engagement in this context is related to the number of retweets of a post since is it one of the metrics used to measure the popularity of a post. Twitter has three features that categorizes engagement: likes, retweets (quoted tweets), and replies. The paper narrows down on retweets as a measure of engagement because retweets are a guaranteed method to increase the reach of a post. While the number of likes portrays a user's affinity to a post, Twitter's current algorithm does not automatically place a liked tweet on a follower's timeline. It may only do so if there are a reasonable number of likes from other mutual followers. On the other hand, Instagram has three features that categorise engagement: likes, saves, and comments. The paper narrows down on likes as the measure for engagement, as it can be viewed as an indicator of desirability.

One of the methods used to measure the performance of a business on social media is through digital engagement. Digital engagement can be viewed as ‘online behaviour resulting from a consumer’s thoughts, emotional connection, and intrinsic motivation to interact and cooperate with a brand or its community members’ [7]. Digital engagement on social media is measured through post metrics such as likes, shares, comments or saves. A higher amount of digital engagement is often associated with a more brand image since interaction with online consumers shows a higher likelihood of a business having greater brand awareness. Some certain features and factors are associated with each given post. Unique characteristics of the association include the time of the base, the type of post, the caption, or the image itself. These features exist with each post on Twitter, a social medium of focus for this research.

1.2 Twitter and Instagram as a Marketing Tool

Twitter is an online news and social networking site where people communicate in short messages called tweets [23]. Tweets have a 280-character limit and can incorporate other elements such as media in videos, images, and links. It is often called a microblogging website that is discoverable to people and or companies online. Twitter had some 330 million monthly active users as of 2019, and as of 2020, 166 million monetisable daily active users [22]. The platform has grown to be one of the biggest platforms for businesses to market their services and brands and engage with existing and potential customers, including Facebook and Instagram. Companies have adopted Twitter as part of their digital marketing strategy due to its ease in improving customer service experience and promoting new products to attract new customers [27].

Instagram users are more engaged than on other platforms, with higher engagement rates for businesses than on Twitter and Facebook [26]. It has been reported that 72% of Instagram users are ready to make a purchase decision after being influenced by Instagram [26]. This discovery reveals the potency of marketing on Instagram and contributes to why Instagram has been chosen as the focus of this research paper. This discovery shows the power of marketing on Twitter and Instagram. One of the contributing factors to why they have been chosen as the targeted social media networks for this research paper is consumer brand engagement and online presence.

It has been found that there is a positive correlation between consumer brand engagement and social commerce purchase intent. Though social commerce purchase intention varies across social networks, one thing that does remain is the effect of customer engagement on such measurement and expectation [6]. Consumer brand engagement gauges how interactive a consumer is with a brand. Such a factor is essential as it tells of a consumer's loyalty [7]. The use of brand engagement as a digital marketing tool is vital to understanding how well a business brand is fairing. Brand awareness has always formed a crucial role in consumers' purchase intention. The ubiquitous nature of social media capitalises on that factor, making mobile marketing a very effective tool for reaching potential customers and retaining existing ones. Social media content should continually be researched to create a more conducive environment for businesses trying to improve their online presence.

1.3 Research Hypothesis

With the growth of social media and its effects on business marketing strategies, it is becoming common for organisations to measure successful marketing efforts against the number of likes, retweets, views and comments under their social media posts. As such, being able to predict these metrics, and by extension, the success of these posts is becoming an invaluable benefit to various companies. Regarding predicting engagement such as retweets, most research has focused on retweets as a classification problem (i.e., will this tweet be retweeted or not). Fewer efforts have been geared towards a more focused data collection and a resultant regression analysis as opposed to classification. Most research has concentrated on the data collection process in a very randomised approach, implying that they did not look at specific accounts to gather their data. Though this has yielded acceptable results with these previously researched predictive models, there is little emphasis on companies and brands that have embraced the power of Twitter and even less emphasis on the viability of a regression model being used to achieve these results. The more random approach that is often adopted for data collection, encompasses a wide range of users' tweets being thrown into the predictive model. However, this research paper aims to skew the data collection towards past data from selected brands and companies with an evident online presence. Though this proposed model could potentially benefit Ghanaian brands on Twitter, many Ghanaian brands have not established a strong presence on Twitter, thus forcing the data collection process to align to brands with a more substantial company, community, engagement, and following - international brands. Due to this unique data collection process, the resultant accuracy level of the different models employed is yet to be determined.

The hypothesis that this thesis aims to explore is whether a regression model can be used to generate meaningful social media engagement predictions, although it is commonly addressed with a classification model. To test this hypothesis, there will be three experiments will be conducted to verify whether or not a regression model can bring forth meaningful predictions. The first experiment will conduct a regression analysis on a relatively small data set, inclusive of outliers. The second experiment will conduct a regression analysis on a significantly larger data set, also inclusive of outliers and the third mimics the second experiment but without the presence of outliers.

Chapter 2: Related Work

2.1 Predicting A Tweet's Popularity

Research related to tweet prediction is gaining ground as the usefulness of understanding such a feature is essential. Researchers in this field have adopted different approaches to solving the problem of prediction. One method employed uses different kinds of classifiers to determine the best model for predicting tweet popularity. One example is the research on Thai tweets, using a decision tree, neural network, and Naïve Bayes [24]. The research outcome showed the decision tree classifier yielding the best results with an accuracy rate of 98.8% when only tweet-related attributes were trained for the prediction. With the experiment run in three sets, user-related attributes, tweet-related attributes and a combination of all features, the research illuminated the efficacy of user-related attributes in predicting the popularity and unpopularity of a tweet.

Another research explores the popularity prediction for a single tweet using a heterogenous bass model [29]. It predicts the popularity of a post within its life cycle, given user features and tweet features. The adoption is the Bass Model, a model used to predict the performance of a product, which is tweaked to produce a Spatial-Temporal Heterogenous Bass model and Feature-Driven Heterogenous Bass model.

1. Spatial-Temporal Heterogenous Bass model

    This adjustment of a standard Bass model:

    $$f(t) = (p + qF(t))(1 - F(t))$$

    Where the influence of factors outside the population is p (innovators), the result of contact within the population is q (imitators), f(t) is the likelihood of purchasing time and (1-F(t)) is the probability of a user not adopting to a product, produces a newer equation to account for user features and single-tweet features.

2. Feature-Driven Heterogenous Bass model

    This model focuses on the outcome of different features based on the diversity of tweet prediction, mainly user features and single-tweet features. Both features resemble innovators and imitators in the equation on the Bass model.

Furthermore, the popularity of a tweet is weighted and measured with retweets, likes, and replies. Apart from this, the observed predictions are analysed quantitatively to determine the tweet's popularity. The perspectives adopted to predict the popularity of a tweet resembles a more realistic expression of tweet popularity in the real world.

[30] improves upon the problem of tweet predictions by predicting popularity in the context of only the tweet and in a context where the process of retweeting is well-known. The model follows a non-linear joint embedding network to map image and textual features. [30][8], employ a multimodal approach where the particulars of the textual and image features are analysed to produce encouraging results. In [30], prediction is based on the tweet language, image, and author specifications. The results postulate the replicability of the model on other social media networks. Like [31], components of a tweet that received higher virality included URLs, but unlike [30], [31] discovered the importance of hashtags in promoting the virality of a tweet. The prediction of retweets has been solved as a regression problem by [30]. However, their approach is deemed more complex as it amounted to millions of data points spanning over eighteen months and incorporating layers of an embedding model with a deep learning approach. This research seeks to understand the viability of simplifying this regression problem as its contribution to the academic community.

While some research predicts either the count of retweets or the probability of a retweet is based on user and tweet features, [41] predicts popularity based on dynamic time warping and sequential time series clustering. Dynamic Time Warping is a time series algorithm used to find similarities between two-time series with different lengths. Posts are grouped based on similar popularity and grouped again based on their temporal patterns. It relies solely on the length of the profile – observed over a period - with a collection of the total number of retweets and replies as the popularity count.

2.2 Predicting the Sentiment of a Tweet

The sentiment of a tweet often plays an essential factor in how well a tweet will be received. While more research has focused on the sentiment polarity of a tweet itself, [39] focuses on the sentiment polarity of a tweet reply. Replies are metrics used to measure the popularity of a tweet and gauging the sentiment of a reply as a further step in metrics measurement can be beneficial to businesses and individuals alike. The research's tweet classifier performs within range of both state-of-the-art and traditional machine learning methods. Apart from using sentiment in a tweet's reply, it is employed to understand a user's semantic-word presentations from extracted tweets of hashtags and outperforms traditional recommendation and popularity-based recommendation for users using hashtags [40]. To boost the quality of user engagement, such a model can be proposed to active users, predicated on hashtag-based events.

The different approaches of the topic's related research show the relevance of a tweet's semantics, user and tweet-related features, the static and dynamic setting for measuring tweet popularity and time features associated with a tweet.

2.3 Predicting Popularity on Instagram

The research on predicting social media engagement has been chiefly extended to the less commercial side of social media, such as predicting engagement in political conversations or the dissemination of news content [14, 15]. A large part of the idea of predicting engagement has been reduced and limited to the theoretical aspect, where training and testing of models have only been used on past data but never on data that is yet to occur. However, this factor does not discredit the need for Artificial Intelligence and machine learning within the commercial space on social media. The scope of this topic thus far has focused mainly on Facebook through other media like Twitter, YouTube, or Reddit. Notwithstanding the research that has been conducted to create models, data that has been retrieved has not focused entirely on Instagram. It is often in tandem with other social networks, and even if the data retrieved from there is the sole focus, it is for Higher Educational Institutions [16]. Since user engagement differs across social media [17], it is imperative to fully understand the inner working of one medium to create a sound output.

Research on predicting the popularity of a post has been primarily focused on deep learning-based models. In the past, models were focused on only one form of modality, such as text. However, in the more recent year, more researchers have focused their field of search on other modalities, like images. The introduction of image content has assisted in producing more accurate prediction results [32]. In this paper [32], the focus was on Flickr posts. Like Instagram, Flickr can be described as a site that emphasises shared photos, though it has its nuances.

[37] use Instagram exclusively as the platform to collect data to train and predict. The result of the implemented Deep Neural Network resulted in an 88% average accuracy. The

accuracy rate was confident enough for the model to be employed in practical commercial use. The fifteen input parameters that were included in the model included the filter applied, if the location was used for the post, the creation time of the post, week of the year and day of the week of the post, an hour of creation, URL of the image, number of tagged users, the caption, the length of the caption, the number of tags used and the list of the tags. This default input features formed part of the result of JavaScript Object Notation (JSON) extracted from Instagram's API. The number of likes, one of the metrics measured in the paper, were grouped into classes. For instance, likes between 0 and 25 were classified as Class 1 and so on. The authors adopted a Deep Neural Network, precisely a four-layer Stacked Auto-Encoder with four hidden layers. In this approach, the researchers deliberately used only one type of model to predict Instagram posts. The result of this paper was the prediction of likes, comments, and shares. Like in [37], [32] uses a deep learning approach, but [32] employs different models before concluding on the best model for the case of predicting popularity and digital engagement. [32] uses a random forest approach, convolutional neural networks, and transfer learning on images through an already existing pre-trained model for pictures but with tweaks – InceptionResnetV2. Each of the methods used in [32] is specific to the type of data in the posts.

Chapter 3: Methodology

From the stated problem associated with this research and other past work surrounding the prediction of engagement on social media, there is precedence for the basis of prediction, as it offers insight into the intricacies of increasing digital marketing. Whether the prediction of engagement is within a static or dynamic setting, the results provide enough understanding of this prediction problem in the context of Twitter. However, for the prediction of engagement on Instagram, pictures' context is crucial to understanding popularity, and so modelling will include such an attribute. Due to this, this paper's prediction of digital engagement is specified as a regression problem. The predictor variable is targeted as the number of engagements in the form of likes on Instagram and retweets on Twitter. The significant distinguishing factor of this research focuses on the relevance of post-related attributes, user-related attributes, sentiment analysis and picture analysis, and the nature of the type of data collected.

Although the understanding above was the paper's initial objective, the following sections reveal unforeseeable changes that affected the outcome of this research and brought about new insights concerning the hypothesis.

3.1 Data Privacy, User’s Information and Data Collection

Considering higher security measures and data privacy issues on the internet, social media networks such as Instagram have modified their data privacy to implement more stringent measures on data usage from third-party users. In the bid to collect data from the chosen social media network, the ethical concerns surrounding data collection had to be examined, which influenced the choice of social media to conduct the final regression analysis. As such, the approach for data collection had to be scrutinised using two methods – web scraping and Application Programming Interface (API). This paper is centred on adopting an

API for the data collection process. APIs programmatically grant access to information, compared to web scrapers that sometimes tread the grounds of unethical and illegal use and behaviour. Instagram’s updated Data Policy as of 4th January 2022, for example, states that:

> We are in the process of restricting developers’ data access even further to help prevent abuse. For example, we will remove developers' access to your Facebook and Instagram data if you haven't used their app in 3 months. We are changing Login to reduce the data that an app can request without app review to include only name, Instagram username and bio, profile photo, and email address in the next version. Requesting any other data will require our approval.

Such a process has already been implemented, making it harder for developers even using the company’s API to gain access to information like public metrics without the development of a fully functioning app approved through a review and implementation in Live Mode and stipulating requirements such as data deletion processes and privacy policy. This shows that some companies are taking more action to protect users’ publicised data even through their API.

The long-standing argument states that as long information is publicised openly on the internet, it is open to collection and interpretation [42]. While this paper does not exploit the freedom of posting online, some of these ethical concerns affected the chosen social medium network. With the introduction of the General Data Protection Regulation (GDPR), there are more defined guidelines for research on human subjects, such as the amount of data extracted within a timeframe and the method of extraction [43]. [43] states the importance of collecting publicised datasets, likely to be anonymised and not capable of causing harm. Given the above, the business accounts collected for this research are fully publicised and are in no way seeking to misrepresent, or slander gathered data on chosen entities.

3.1.1 Instagram: Scope Constraints, Stringent Data Retrieval

In 2022 January, Instagram updated its terms of usage, which affected its data usage as well. This, in turn, affected how developers can interact with their API, specifically their Graph API. For Instagram to grant permission for developers to gain insights into a public account's engagement, it requires an application to be in live mode instead of development mode for most APIs. For an application to be in live mode, Meta requires an App Review, including a fully functioning website with Data Deletion Instructions and a Privacy Policy URL. Up until early 2022, an App Review was not required to access metrics of users.

Continuing with Instagram's data would have pushed this research over its initial scope, as getting verification to use the Login Feature of the API would have required having developed a fully functioning website with specifications on how a user's data would be handled. As of the latter end of 2021, gaining access to the Login feature did not require having a developer operate in Live Mode, as operating in such mode indicates readiness for deployment. These changes occurred in the earlier parts of 2022, when this project commenced. However, the paper's focus is on assessing the viability of a regression model for social media prediction and not creating an app, especially not before any model was tested for the paper's objective.

3.2 Requirement Analysis of Data

Since the retrieval of public post metrics from Instagram was not achievable due to the scope creep, the focus from this point onward is shifted to Twitter alone. Granted that this paper aims to predict the amount of digital engagement in the form of retweets, the data acquired should aid in the prediction process as postulated by the researcher and deduced from previous work related to this topic. Since this work focuses on brand engagement from businesses, the data collected from Twitter is taken from twenty consumer goods companies abroad, as specified in the following table.

Skittles	KFC	TacoBell	Pringles	Trident Gum
Starbucks	Whole Foods	Burger King	PlayStation	Xbox
Netflix	DiGiorno	Jet Blue	Nike Store	Adidas
Microsoft	Nasa	Snickers	Reese's	Chanel
Versace	Slimjim	Sour Patch	Dolce&Gabbana	

Table 3.2: Accounts used for the Data Retrieval Process

The response variable is the retweet count of any given tweet from the case explored. However, since Twitter's introduction of quotes tweets in 2015, the count of this metric will be added to the retweet count because they function as very similar metrics. What differentiates them is the quoted tweets' ability to retweet and add some text, image, video, or emoji. The predictor variables constitute both user-related features such as the number of followers and tweet-related features such as the tweet's creation time and the tweet's sensitivity.

3.3 Dataset

The dataset for this paper was scraped and is gained from Twitter's API. The first set contains approximately 5000 rows, while the second step contains about 32,000 rows of data with 18 columns of the independent and dependent features.

The description below specifies the dataset acquired:

1. Tweet_ID: The unique ID associated with each tweet

2. Tweet_Text: Raw text of the collected tweet

3. Tweet_Analysis: Categorical variable representing the sentiment of the tweet represented by Positive, Neutral and Negative

4. Tweet_Creation: DateTime feature showing when the tweet was posted

5. Retweet_Count: Number of retweets for the tweet

6. Reply_Count: Number of replies for the tweet

7. Like_Count: Number of likes for the tweet

8. Quote_Count: Number of quoted retweets for the tweet

9. Hashtags: Number of hashtags used in the tweet

10. Media_Type: Categorical variable representing the type of media in the tweet and represented by None, Media or Video

11. Followers_Count: Number of followers of the tweet author

12. Following_Count: Number of accounts followed by the tweet author

13. Tweet_Count: Total number of tweets by the tweet author

14. Listed_Count: Number of lists (curated selection of Twitter accounts)

15. User_Created: Datetime feature of the account's creation date

16. Tweet_Length: Vectorized length of each tweet

17. Day_of_Week: Extracted from the tweet creation DateTime stamp to identify which day in the week the tweet was posted

18. Hour: This derived feature from the tweet creation timestamp creates an hour for each of the tweets that were posted

3.4 Models Employed

All the machine learning models used for prediction are supervised learning techniques because the target and predictor variables are both given. Such techniques seek to model relationships and dependencies between the target and predictor variables. Each of the models described here offers a brief description, but the details of each of the papers are explored in Chapter 4.

3.4.1 Decision Trees

This model forms a hierarchy of if/else questions to lead to a decision. This model aims to predict a target variable's value by learning simple decisions deduced from the data's features. This algorithm can be used for solving either regression or classification problems but will focus on a regression analysis to maintain consistency with this research.

The algorithm for decision trees utilizes a top-down approach, greedy search and no chance for backtracking within the space of possible branches. It consists of a decision node which is characterized as the features and a leaf node which characterizes a decision on the numerical target. The topmost decision node is classified as the best predictor and called the root node. The tree is structured as:

- Node – a feature
- Branch – a decision
- Leaf – an outcome (target variable)

The step involves calculating the standard deviation of the target deviation which is retweets in this research. The standard deviation of each branch is calculated with the standard deviation of each branch calculated and subtracted from the initial standard deviation before the split of the dataset. The result is the standard deviation reduction. For each selected feature, say media type, the dataset is divided based on the values from these features and run recursively on the non-leaf branches. A stopping criterion is determined, when the coefficient of deviation – the relative dispersion of data points around a mean – hits a threshold [36].

$$Standard\ Deviation\ Reduction(T,X) = S(T) - S(T,X)$$

In this equation; T = retweets; X = [feature in dataset]

3.4.2 Random Forests

In choosing which machine learning model to use to tackle the problem proposed in this paper, one of the models settled upon was Random Forests. Random Forests is well suited for either a classification problem or regression problem. However, the target variable of this research is a continuous variable which makes the situation a regression one. The non-linear nature of this model makes it a more excellent option over linear models, though this will still be tested in this paper. This model is an ensemble of decision trees – costumes combine multiple machine learning models to form more powerful models. In the case of a decision tree overfitting on training data, numerous trees constructed randomly would produce overfitting, but it will vary, thus reducing the average result of overfitting.

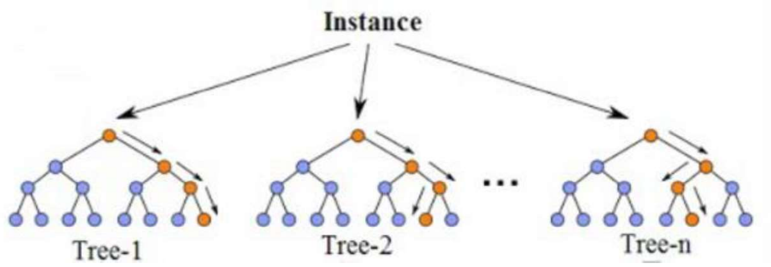


Figure 3.4.2: Random forest structural diagram

3.4.3 Support Vector Regressor

A Support Vector Regressor is built on a Support Vector Machine**.** This algorithm creates a hyperplane that separates data into classes. The model is helpful as it assumes some non-linearity in data and presents a proficient prediction model. Its implementation is easy, and it is robust to outliers. Though this classification may seem suited only for a classification problem, it is also utilised for regression problems. With the Support Vector Regressor, the hyperplane is calculated within the boundary line, which measures the hyperplane's distance and is used to create a margin between the data points. Unlike a simple regression model that aims to minimize the error rate, a Support Vector Regressor fits the error rate within the margin of epsilon (Ɛ). The aim of the SVR is to fit the highest number of data points without violating the margin. The Support Vector defines the hyperplane that holds extreme data points in the dataset. It discovers a hyperplane in an n-dimensional space, characterized by the number of features. Two of the parameters of this algorithm are the kernel and penalty cost C. The parameter utilized for this problem; the Radial Basis Function (RBF) kernel a penalty cost of 1.0 [38].

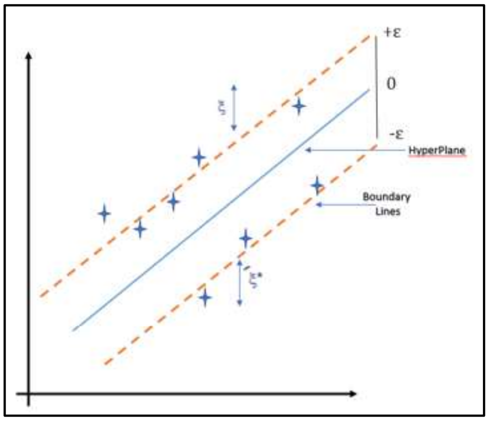


Figure 3.4.3: Support vector regressor sample diagram

3.4.4 Artificial Neural Network

This subfield of machine learning employs the use of deep learning. A deep learning model consists of input, output, and hidden layers. Such models offer an advantage over some of the models employed in this paper as they can do complex analyses and build new features from a limited set of data. This additional method may provide more insight into the modelling stage as many models – ranging from simple to complex – are introduced. Deploying the Artificial Neural Network is done through Google Colaboratory due to its offerings of higher disk capacity.

3.4.5 Naïve Bayes

All the models mentioned already are implemented to predict the retweet, but Naïve Bayes is used to classify sentiment for tweets. It is a classification technique derived from Bayes' Theorem and assumes that the presence of one feature in a class is independent of another feature.

3.5 Approach

The outcome that will be achieved at the end of this paper can be primarily divided into five major sections: Data Collection, Data Processing, Data Exploration, Data Modelling, Interpretation. Figure 3.1 below is visual representation of the research flow this paper adopts.

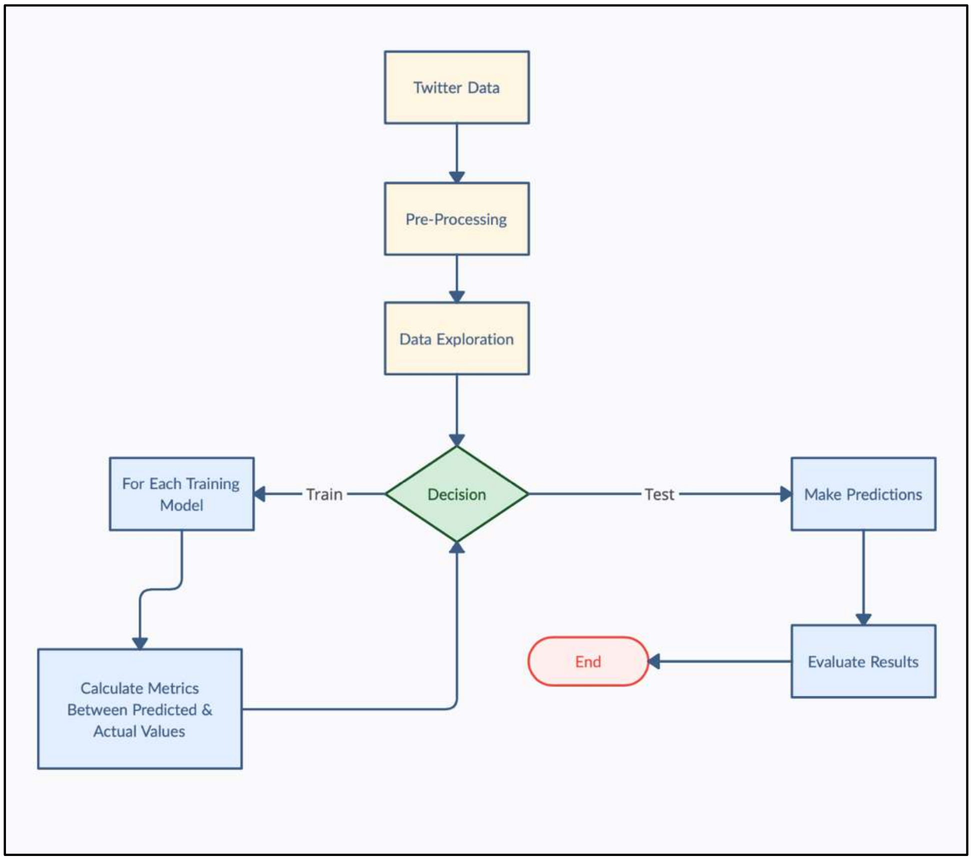


Figure 3.5: Flow chart of research

1. Data Collection

   This constitutes the collection process from Twitter from five foreign consumer goods. All the required data is generated using Python with the assistance of various libraries and packages. The data is divided into a training and testing set where the modelling process, as the name suggests, occurs on the training data, and testing of the solution's viability occurs on the testing set.

2. Data Processing

   This constitutes processing all the columns into the suitable data form, such as encoding categorical variables, eliminating empty fields, removing duplicate values, or chancing any unsuited data formats. This process also includes the sentiment analysis of the individual tweets used in the training and prediction process. At this stage, all sorts of standardisation and normalisation of the input variables occur for the data to be best interpreted by the various machine learning models used.

3. Data Exploration

   This explores the relationship between the predictor variables and the target variables better to understand the extent of causality and perceived correlation and discover patterns, characteristics, and points of interest.

4. Data Modelling

   This section will highlight the different chosen models outlined in the section above to predict the retweet rate by analysing which models produce the best results and gauge if a 'best result' can be achieved.

5. Data Testing

The other part of the data set is used to test the accuracy of the various models adopted. The outcome of the testing will determine which model is best suited to handle this research.

3.6 Technologies Employed

This section addresses the various technical tools and resources employed throughout the duration of this project to conduct experiments, clean data and interpret results.

3.6.1 Jupyter Notebook

Juptyer Notebooks is a web-based interactive computing platform for creating notebook documents developed by Project Jupyter. The platform makes it easy to combine code executions, rich text, equations, plots, and multimedia. Users can write modular bits of code in chunks known as cells and provide documentation to developed software in Markdown, HTML or LaTeX.

This was chosen as the means for the execution of the research because it is ideal for data science tasks like data cleaning and transformation, exploratory data analysis, data visualisation, machine learning and statistical modelling. As a web-based platform, it is accessible on any device and leverages big data tools.

3.6.2 Tweepy

Tweepy is a Python library integrated with the Twitter API to access Twitter's endpoints. It was used to retrieve all the available data points from Twitter. Before it was accessed, an app was created with a developer account on Twitter for specific keys and tokens to generate an initial connection to Twitter's endpoints.

3.6.3 Matplotlib

Matplotlib is another library used in this research for the visualisations – plots and graphs. It is employed for simple, complex, or interactive displays and will be the main library used to visualise model predictions and results and exploratory analysis.

3.6.4 Pandas

Pandas is a Python library used for data analysis and manipulation. It is used to create DataFrames, file importing and conversions, data processing, analysing, and plotting.

3.6.5 Numpy

NumPy is a package for Python scientific computations, arrays and matrices, such as comprehensive mathematical functions and randomisations. This package is employed for calculations between various independent variables necessary for future predictions and experimentations.

3.6.6 Sci-kit Learn

This machine learning library for Python features many models for classification, regression and clustering as needed by supervised and unsupervised machine learning models. It was chosen as the library to employ from the various models used based on its simplicity since it is an abstracted view of the more complicated versions of machine learning models. Since it is an essential library for preprocessing tasks as well, it was used to split the retrieved data into a training and test dataset that is utilised throughout experimentation. It is essentially built upon NumPy, SciPy and matplotlib – two of which are elaborated on above.

3.6.7 Keras

Keras is a high-level neural network library that runs on top of TensorFlow, an open-sourced platform with numerous machine learning tasks. This library is adopted for modelling the problem after an Artificial Neural Network.

Chapter 4: Implementation

To achieve the paper's objective, the implementation process is categorised into the typical process of a machine learning problem. The approach throughout this section would be classified into three experiments: the first one with 4,558 data points (inclusive of outliers), the second one with 32,186 data points (inclusive of outliers) and the third with 32,186 (without outliers). Each of these experiments runs through the same structure. After collecting data, the different features of the dataset are sifted through to remove any missing values, perform sentiment analysis, transform specific features, extract other features, view the importance of all features, and delete any unneeded features. From here, models like decision tree regressor, decision trees with feature selection, random forests, and support vector regressor. The models are trained on 75% of the data mentioned in the previous section, with the remaining 25% of the data used for testing. The accuracy of each model is measured to show whether it can prove the paper's objective.

1st Dataset	Number of Examples	2nd Dataset	Number of Examples
Training	3,419	Training	24,140
Testing	1,139	Testing	8,046
Total	4558	Total	32,186

Table 4: Distribution of Training and Test Data for both Experiments

4.1 Data Collection

This process involved engaging with Twitter’s API and Tweepy, a Python library used for accessing the Twitter API. Data was collected in the months of April and March. The access keys from the API enabled the use of Tweepy, through which the *get_user* function accessed the Twitter user's unique ID, and the *get_users_tweets* role accessed the tweets on a user's timeline. Within the *get_users_tweets* function, tweet fields, media fields, and user fields were specified. Some other accounts were considered in the retrieval process, and though they were publicly accessible, media information related to their past tweets was unavailable. This restricted the type of accounts that were targeted.

The media key was crucial to understanding whether a tweet had media attached to a tweet, in pictures, videos or animated GIFs. In the function, retweets were excluded because it was assumed to fall under a user's response category. The nature of replies is also similar. However, answers were maintained as a feature to examine the effect of retweeting on being a reply or not. Nonetheless, removing both replies and retweets from the data sample would have reduced the number of tweets that could be retrieved. The *get_users_tweets* function also has a limited number of responses that could be retrieved, and not all requests resulted in the maximum results achieved due to the pagination process. All the data from the individual users were then merged to form one dataset.

4.2 Data Processing

Data processing for the data set involved:

1. One-Hot Encoding: This is a technique that changes categorical features into dummy variables, and these were applied to the features:
    - day of the week – notifying the seven days in a week
    - media type - declaring whether an associated media field has no media, a picture, a video or an animated GIF

- Reply - notifying whether the tweet is a reply or not.

After one-hot encoding, one of the derived dummy variables was dropped to prevent the curse of dimensionality. This phenomenon explains that as the number of features increases, generalising a model becomes more challenging, so more training data is needed. Dropping the additional feature gives the model one more minor feature to worry about. However, the process of feature reductions steps is incorporated later in the analysis process.

2. Sentiment Analysis on Tweets: Letting machines understand the sentiment of the text is an ongoing problem that has been solved with many different approaches. The result of the problem often classifies a text into positive, negative, or neutral. Though it has not been perfected, past models can be employed with current issues. To achieve higher accuracy in determining the sentiment of the given tweets in the dataset, this research employed an existing problem using Naïve Bayes. Naïve Bayes is a classification method that predicts the probability of different classes and is mainly used for test classification as it is known to give good results [44]. The approach adopted for this paper employed [45] used an existing dataset of about 1.6 million tweets and classified tweets into negative and positive. It resulted in an 85% accuracy during the testing phase and a 77% accuracy with testing data. The process involved removing English stop words such as, *in, the, an,* and *a, tokenising* words by splitting them into smaller units of single words and cleaning the words to replace shortened terms and abbreviations. The outcome of this model is tested with this research's dataset. Upon further investigation, not all the words were analysed accurately as some seemingly positive words are categorised as negative.

3. Feature Extraction: Not all the hypothesised features that may be helpful in the prediction of engagement are available from the initial dataset. The hour of the tweet, the length of the tweet and the day of the tweet. As these characteristics form part of creating a tweet, it seemed imperative to add them as features.

4. Dropping Features: Though replies and like count also count as a measure of digital engagement, the focus is on retweets. Through the exploratory analysis described below, it was evident that multicollinearity existed between several likes and the number of replies. They also act as resultant features because there is no evidence of this feature until after a post-publication. Due to this, these features were not included in the model. Other features that were also redundant for the modelling process had tweet text, timestamp of the tweet, the usernames, whether the user is verified, when the user was created and the tweet ID.

4.3 Data Exploration

This is the subsection in which different predictor features were explored in correlation to the target feature through univariate and bivariate analysis. The exploration was examined between the target variable and other predictor variables and between the predictor variables. The pictorial representations show much skewness in variables from this portion, such as like and reply count and retweets. This implementation portion also revealed the correlation and significance between the predictor and target variables. For the feature selection process, tools such as the Pearson's Correlation, Chi-square test, ANOVA test, and Extra Tree Classifier.

These feature selection processes produced different results in determining which features are the most significant. Figures 4.3.1, 4.3.2, 4.3.3 and 4.3.4 show the results of the elements of the essential features for modelling.

The correlation matrix is based on the Pearson-type correlation but is influenced by outliers, nonnormality and unequal variances. It ranges from a scale of -1 to 1. The closer the value is to either the negative or positive side, the higher the correlation between the predictor and target variables. The feature importance module showed that the top ten essential features were Tweet Length, Not Reply, Hour, Followers Count, Sentiment, No Media, Hashtags and Tweet Count. The Chi-square test produces a value representing the interdependence of features; as this value increases for a feature, it shows higher importance and significance. In statistics, the Chi-Square test tests the independence of two events. The feature selection process selects features dependent on the target variable. A smaller Chi-square value shows the independence of observed features. So, the highest values are found in Tweet Count, Followers Count, Following Count, Tweet Length, Not Reply, Hour, Hashtags and Video. The ANOVA test is a statistical measure for the analysis of variance, which shows the features independent of the target variable. The importance of the feature is measured using a feature's p-value – a value higher than 0.05 is statistically insignificant and is not an influencer of the target variable. For each of the results, the like and reply count are excluded since they are highly correlated and serve as another measure of engagement. The results of each feature selection process are outlined in the figures below.

The decision tree regressor model discussed in the next section also utilises some feature importance to produce the more significant features.

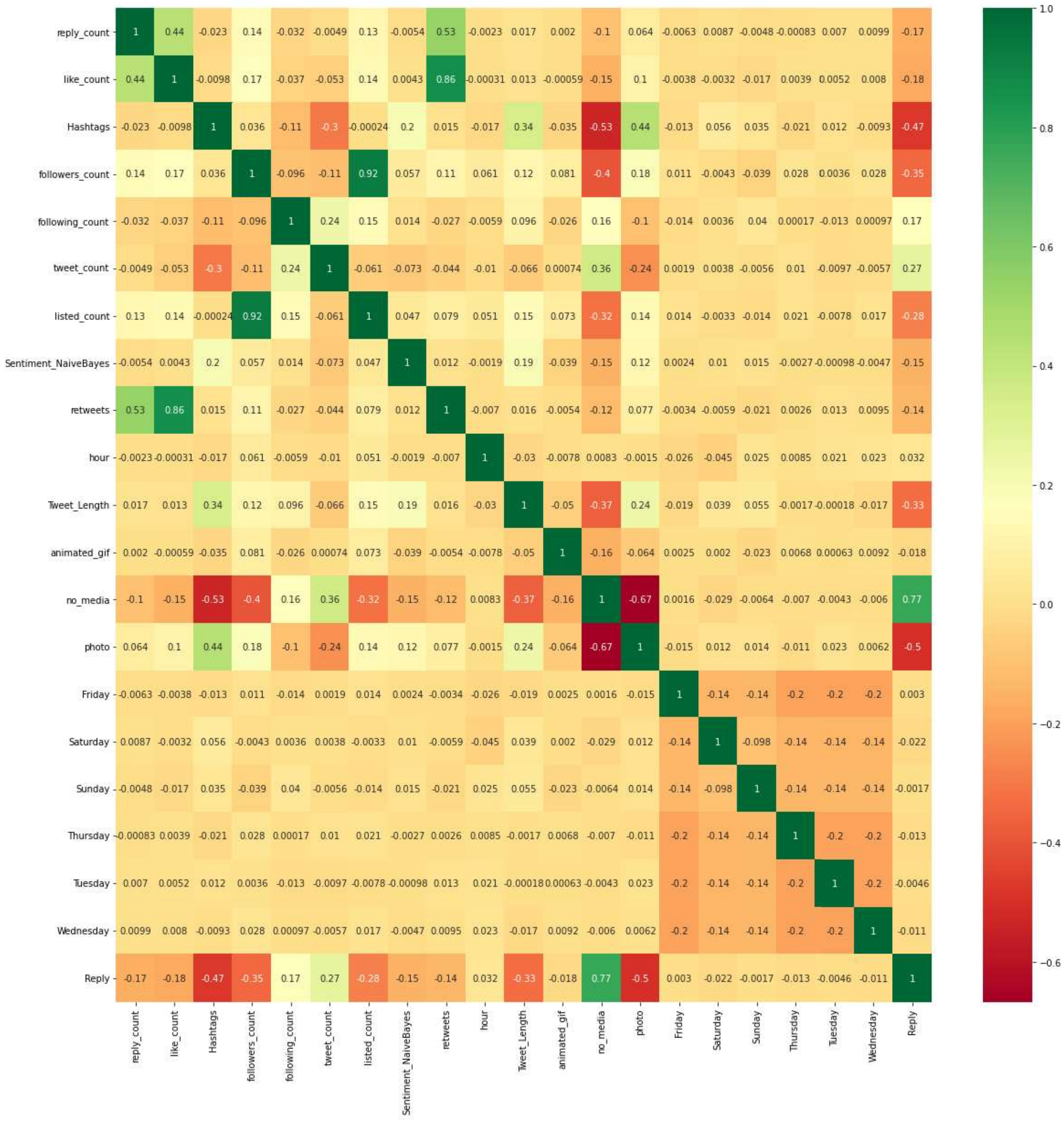


Figure 4.3.1: Results of Correlation Matrix Showing Feature Correlation

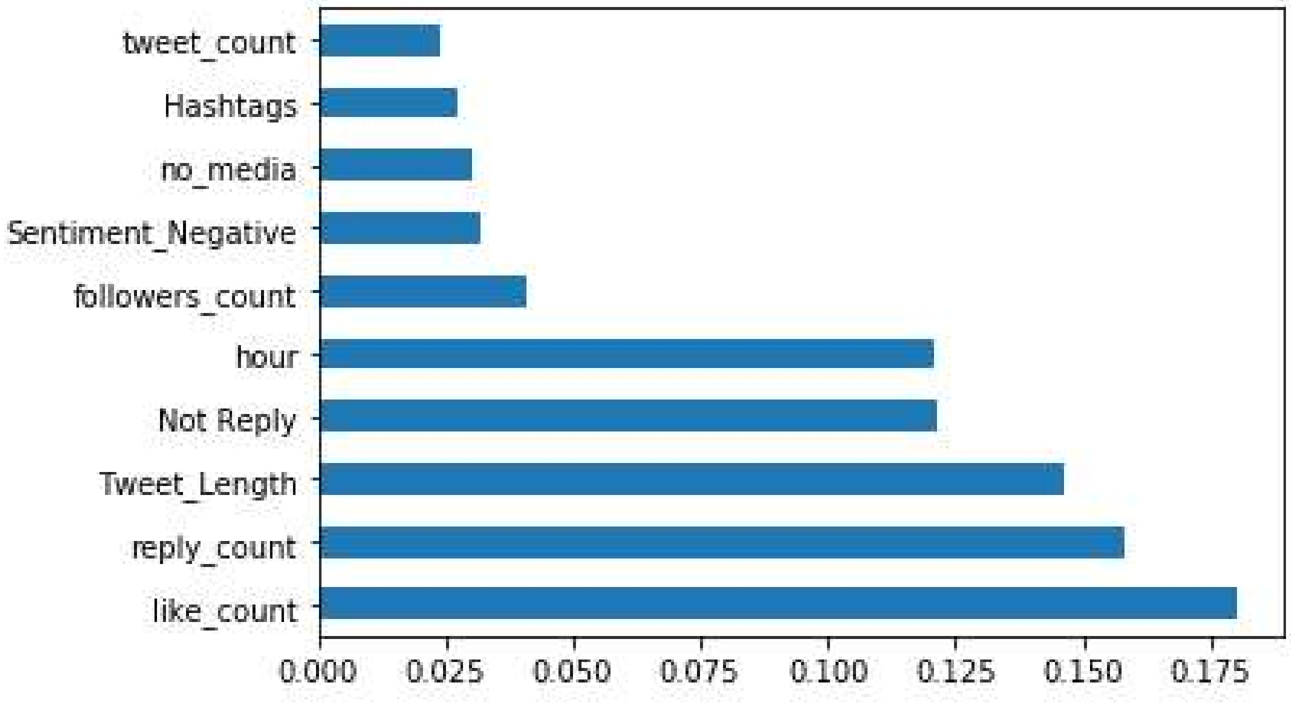
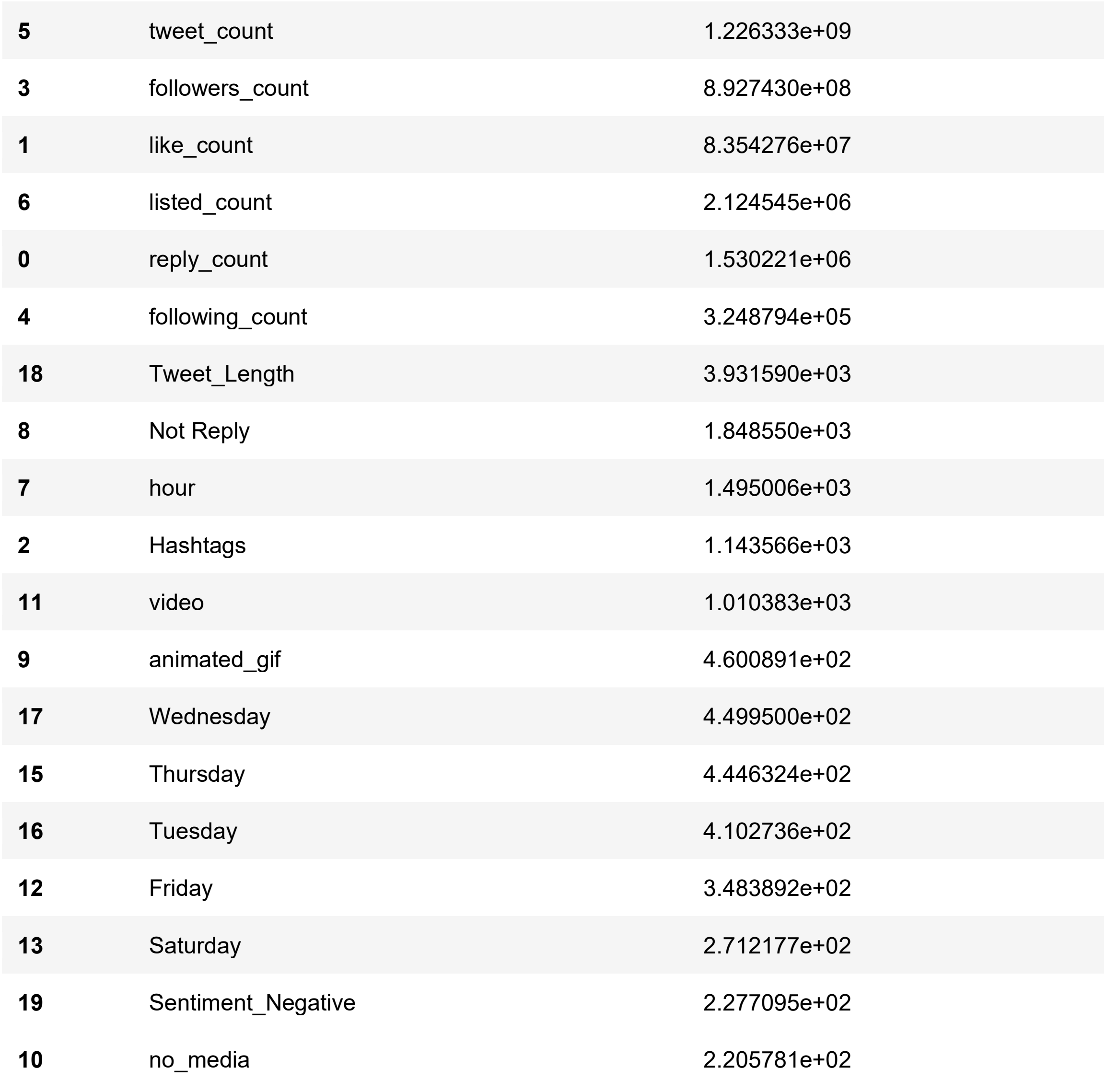
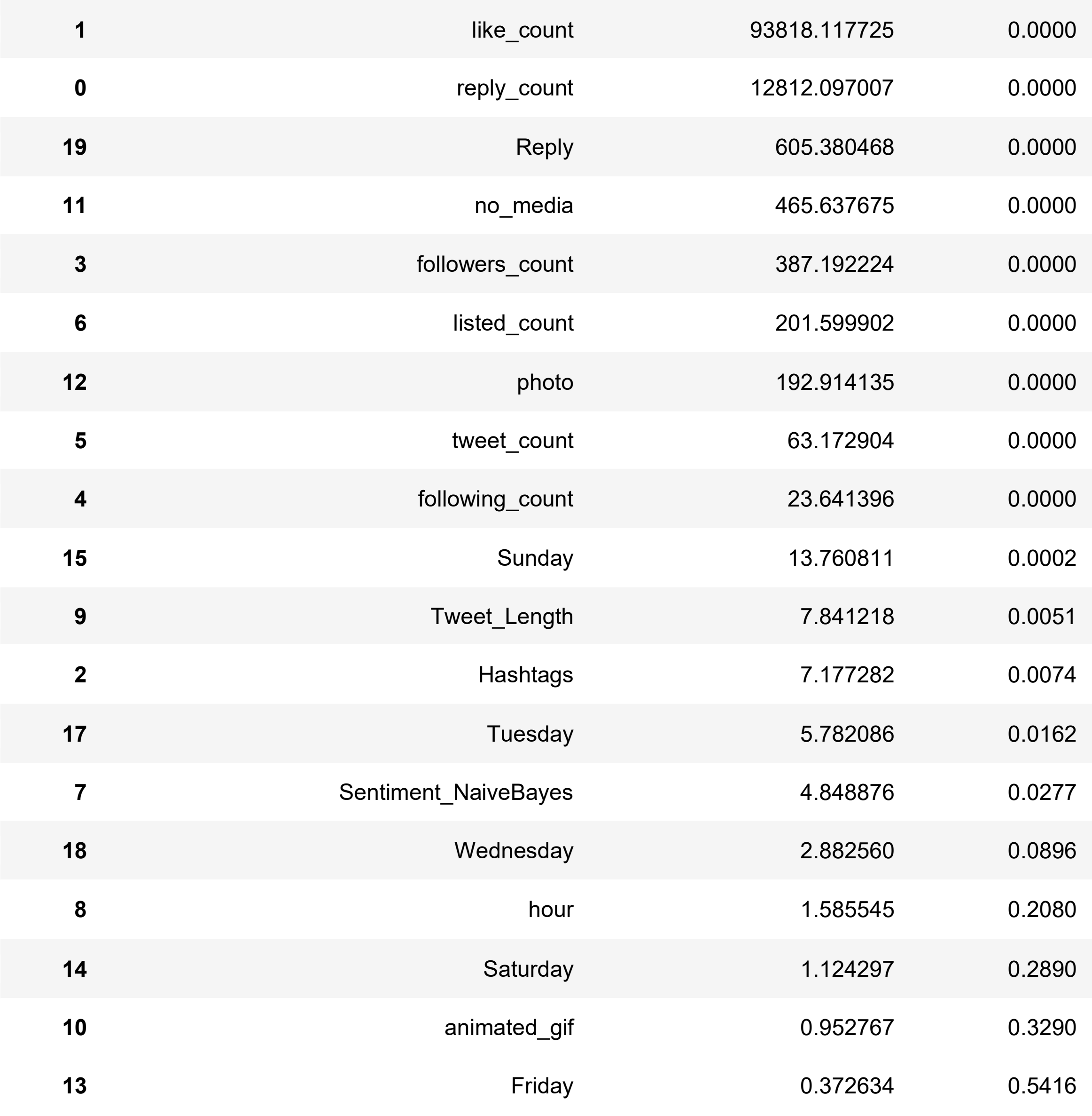


Figure 4.3.2: Results of Tree-Based Classifier Showing Feature Dependence



5	tweet_count	1.226333e+09
3	followers_count	8.927430e+08
1	like_count	8.354276e+07
6	listed_count	2.124545e+06
0	reply_count	1.530221e+06
4	following_count	3.248794e+05
18	Tweet_Length	3.931590e+03
8	Not Reply	1.848550e+03
7	hour	1.495006e+03
2	Hashtags	1.143566e+03
11	video	1.010383e+03
9	animated_gif	4.600891e+02
17	Wednesday	4.499500e+02
15	Thursday	4.446324e+02
16	Tuesday	4.102736e+02
12	Friday	3.483892e+02
13	Saturday	2.712177e+02
19	Sentiment_Negative	2.277095e+02
10	no_media	2.205781e+02

Figure 4.3.3: Results of Chi-Square Test Showing Feature Dependence (Smaller Dataset)



1	like_count	93818.117725	0.0000
0	reply_count	12812.097007	0.0000
19	Reply	605.380468	0.0000
11	no_media	465.637675	0.0000
3	followers_count	387.192224	0.0000
6	listed_count	201.599902	0.0000
12	photo	192.914135	0.0000
5	tweet_count	63.172904	0.0000
4	following_count	23.641396	0.0000
15	Sunday	13.760811	0.0002
9	Tweet_Length	7.841218	0.0051
2	Hashtags	7.177282	0.0074
17	Tuesday	5.782086	0.0162
7	Sentiment_NaiveBayes	4.848876	0.0277
18	Wednesday	2.882560	0.0896
8	hour	1.585545	0.2080
14	Saturday	1.124297	0.2890
10	animated_gif	0.952767	0.3290
13	Friday	0.372634	0.5416

Figure 4.3.4: Results of ANOVA Test Showing Feature Importance (Larger Dataset)

4.4 Data Modelling

Before each test was run, a standardization tool from scikit-learn known as MinMaxScaler, standardized all the input values to put them on a default scale of 0 to 1. This kind of unit variance scaling ensures that features with larger values do not overpower those with smaller values.

4.4.1 Removing Outliers

This research first includes all the outliers from all the features, especially with the first experiment which started off with 4,558 data points. Eliminating the outliers reduced the dataset to only 30% which made it unusable to solve the problem. Especially given the number of features, continuing this process would have reinforced the curse of dimensionality. As such after testing on both the smaller dataset and larger dataset without removing the outliers, the next trial involved removing the outliers on the larger dataset. However, resorting to removing the outliers reduced the dataset by approximately 10,000 data points, leaving only two-thirds of the dataset.

The choice of testing on the dataset with outliers was a result of modeling previous problems that did not mention exclusively removing outliers. Outliers are meant to be removed under the circumstance where there is erroneous data entry, sampling errors and motivated misreporting [33]. According to [33] debates are still ongoing on the cause of action for ‘legitimate cases sampled from the correct population’. A transformation or recoding strategy is advised. Though a unit variance scaling was adopted, it was still not enough to minimize the statistical harm. Therefore, after viewing the performance of the evaluation metrics on the set with outliers, an experiment was initiated to see the effect of the model performance without outliers present. The results from all these test cases are discussed below.

Chapter 5: Results

The section below outlines the research results conducted in various subsections of the outline process stated in chapter three of this paper.

5.1 Measuring Accuracy of Models

For regression models, two of the most common metrics to understand the accuracy of a model are the Mean Absolute Error (MAE) and Mean Squared Error (MSE). A greater MAE and MSE means that the amount of error is large. Error is defined as the difference between the predicted value and the actual value. The MAE is the average of all absolute errors of individual prediction errors overall occurrences of a test set [35]. It is measured as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | x_i - x |$$

Where n is the number of errors or samples, and $|x_i - x|$ is the absolute error of occurrences summed up. The MSE, on the other hand, is the mean of the squared prediction error over all the occurrences of a test set [34]. The prediction is measured as the difference between the predicted and actual values. It is also measured by:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Where n is the sample size or number of errors, y_i is the actual value, and $\tilde{y}_i$ is the predicted value. The summations of all these errors, considering the whole equation, produce the resulting value. Each of the models is pinned against MAE and MSE. The R-squared is another model evaluation metric which Is the percentage of variance explained by a model, effectively stating whether the difference in the target variable can explain the difference in feature variables.

5.2 Results of Experiment 1

	Model	MSE	MAE
Tweet Features	Artificial Neural Network	126762.61	76.39
Chi-Square Test		127000.22	77.22
Tweet Features	Decision Tree Regressor	111706.38	70.41
Chi-Square Test		126426.51	80.36
Tweet Features	Random Tree Regressor	140609.81	86.96
Feature Importance		146889.54	83.58
Tweet Features	Support Vector Regressor	138709.44	79.65
Chi-Square Test		137867.56	78.90

Table 5.2: Table with Evaluation Metrics of Experiment 1

This section outlines the various models used on the first dataset of 4,558 data points. Each model goes through a reiterative process by using insights from the different feature selection processes to run versions of the same model to understand which model would yield the best results. Each of these models uses variations of feature selection processes to train and test the data.

5.3 Results of Experiment 2

	Model	MSE	MAE

Tweet Features	Artificial Neural Network	1795782.25	258.69
Chi-Square Test		1769469.38	252.87
Tweet Features	Decision Tree Regressor	9003709.97	301.60
Chi-Square Test		4316275.25	285.58
Tweet Features	Random Tree Regressor	3799707.95	289.50
Feature Importance		2301169.64	270.88
Tweet Features	Support Vector Regressor	3101542.64	269.80
Chi-Square Test		2981234.05	255.75

Table 5.3: Table with Evaluation Metrics of Experiment 2

This section outlines the various models that were used on the first dataset of 32,186 data points, with outliers. Each of the models goes through a reiterative process by using insights from the different feature selection processes to run versions of the same model, to understand which model would yield the best results.

5.4 Results of Experiment 3

	Model	MSE	MAE
ANOVA subset	Artificial Neural Network	77.27	3.47
Full Set		30.65	2.59
ANOVA subset	Decision Tree Regressor	48.24	2.41
Full Set		132.53	4.25
ANOVA subset	Random Tree Regressor	76.60	3.52
Full Set		76.64	3.50
ANOVA subset	Support Vector Machine	118.80	3.99
Full Set		128.53	4.01

Table 5.4: Table with Evaluation Metrics of Experiment 3

This section outlines the various models that were used on the first dataset of 32,186 data points, without outliers. Each of the models goes through a reiterative process by using insights from the different feature selection processes to run versions of the same model to understand which model would yield the best results.

5.5 Discussion of Experiments

From the experiments above, it is evident that the set that produces acceptable results is Experiment 3 with a dataset without outliers. It was believed that through Experiment 1 had very erroneous results; a larger dataset would prove better. However, from the results, the increase in the dataset size did not influence the models positively until the outliers were removed. Though Experiment 1 and 2 essentially uses features from a Chi-Square test instead of an ANOVA to subset the dataset, the elements from both these tests are similar. They would not positively impact either Experiment 1 or 2.

The best model in Experiment 3 is the Decision Tree Regressor based on a subset produced from an ANOVA test, with the lowest Mean Squared Error and Mean Absolute Error. Even though the MSE is still relatively high, it can outperform models in Experiments 1 and 2. Upon prediction on a set of 10 random feature variables, the predicted values versus the actual values are as follows:

Predicted Values	Real Values
1	7
38	0
14	0
1	23
90	0
1	0
38	71
68	0
1	0
10	0

Table 5.5: Table with Evaluation Metrics of Experiment 3

5.6 Features Influencing Popularity

This section explores the exciting findings that influence the tweet's popularity and what digital marketers could be mindful of to optimise their potential reach. Though it does not offer the number of retweets, these individual characteristics of the explored features still contribute to the holistic approach to understanding social media networks for increasing brand awareness. Features such as the day of the week prove interesting because not all days are significant, and the same applies to the time of the day.

Because many tweets were collected from one user at a time, instead of being wholly dispersed, it isn't easy to glean insights from features such as followers count. However, this does not indicate that user-related attributes are not helpful.

There is evidence of outliers for almost all these figures, which would influence the insights gained, but filtering the dataset without them would have deleted a large portion of the data points.

5.6.1 Retweets, Day of the Week and Hours

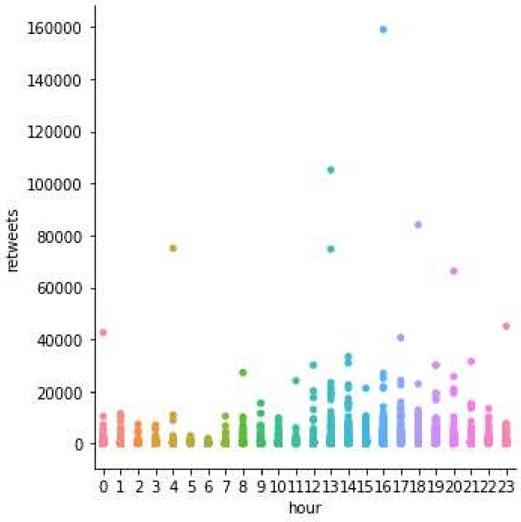


Figure 5.6.1.1

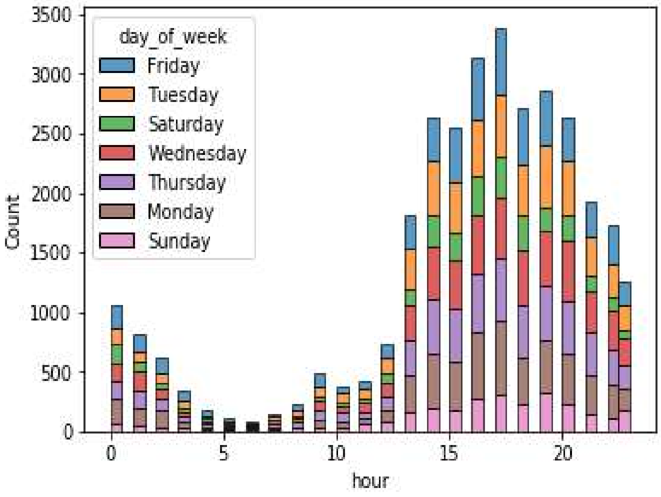


Figure 5.6.1.2

When publishing a post, the time of day and day of the week are always among the factors considered. From Figure 5.1, more activity attracting retweets from about 13:00 to

18:00. The early morning hours are conducive to positive as it is intuitive that there would be less activity around those hours. Based on the level of retweets. From Figure 5.2, the stacked days of the week highlight some of the significant days as the count can be translated to the retweet rate. The days with a higher count include Friday, Thursday, Tuesday, and Wednesday from the prevalent times. Monday also appears to be one of the prevalent days. Interestingly, the weekends do have as man tweets for the popular times.

5.6.2 Retweets and Tweet Length

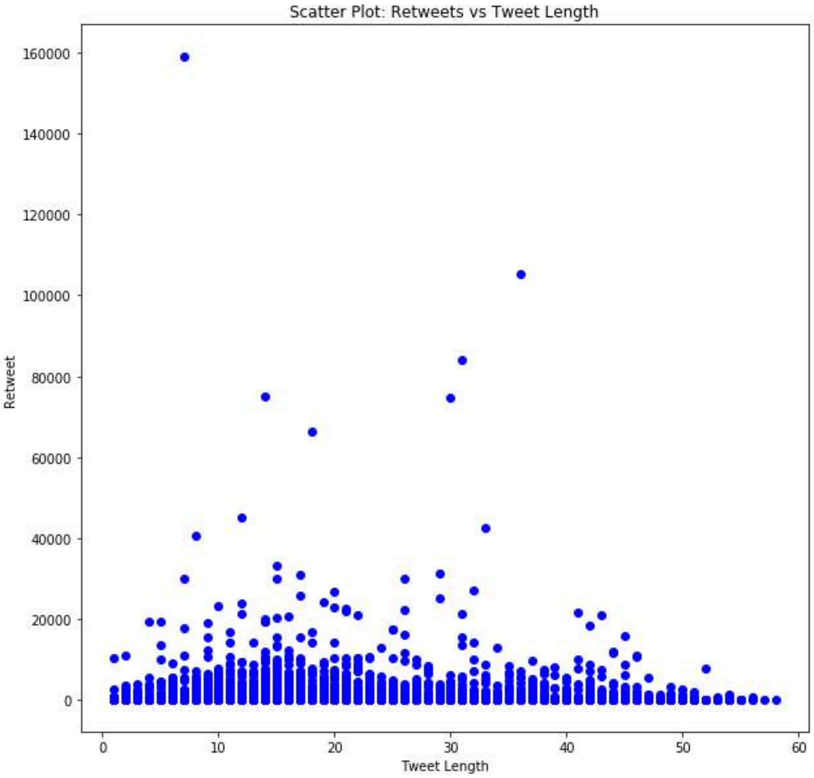


Figure 5.6.2

Twitter has a maximum of 240 characters, and for the collected tweets, the length spanned from 1 to 58 characters. It is quite telling that fewer words are used to convey a message for most of the brands. Most of the retweet rate is based around 10 to 20 characters. Upon examination of some of the tweets, most were short. It is rare to see a tweet with between 45 to 58 characters getting high retweets based on the collected data. Based on this, it is safe to say that most tweets are coupled with some form of media or no media. While longer tweets may be associated with non-commercial users, business brands may not benefit from this characteristic.

5.6.3 Retweets and Media Type

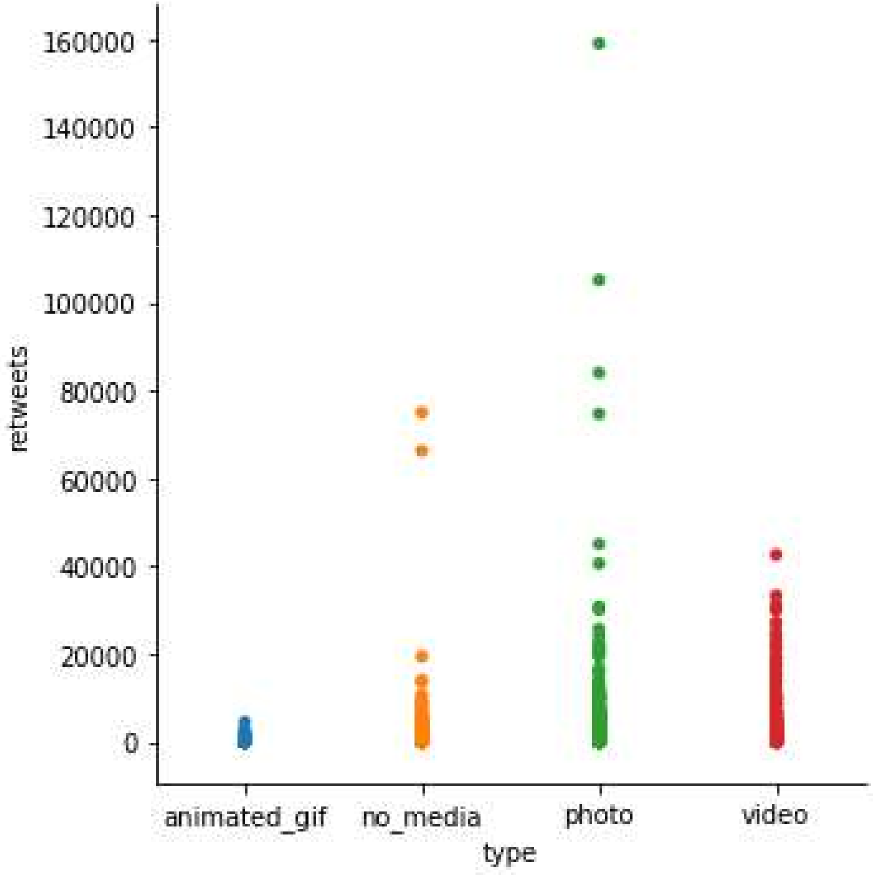


Figure 6.3

In Figure 5.4, the retweet level is influenced more by photos than videos, before no media. Animated GIFs appear not to have as much influence on the number of retweets. Though Twitter is a micro-blogging app, which indicates that more may be used, much communication occurs with photos because a picture says a thousand words, as it is known. Content creators can leverage these tools more and toggle between the three options to provide higher popularity.

5.6.4 Retweets and Sentiment

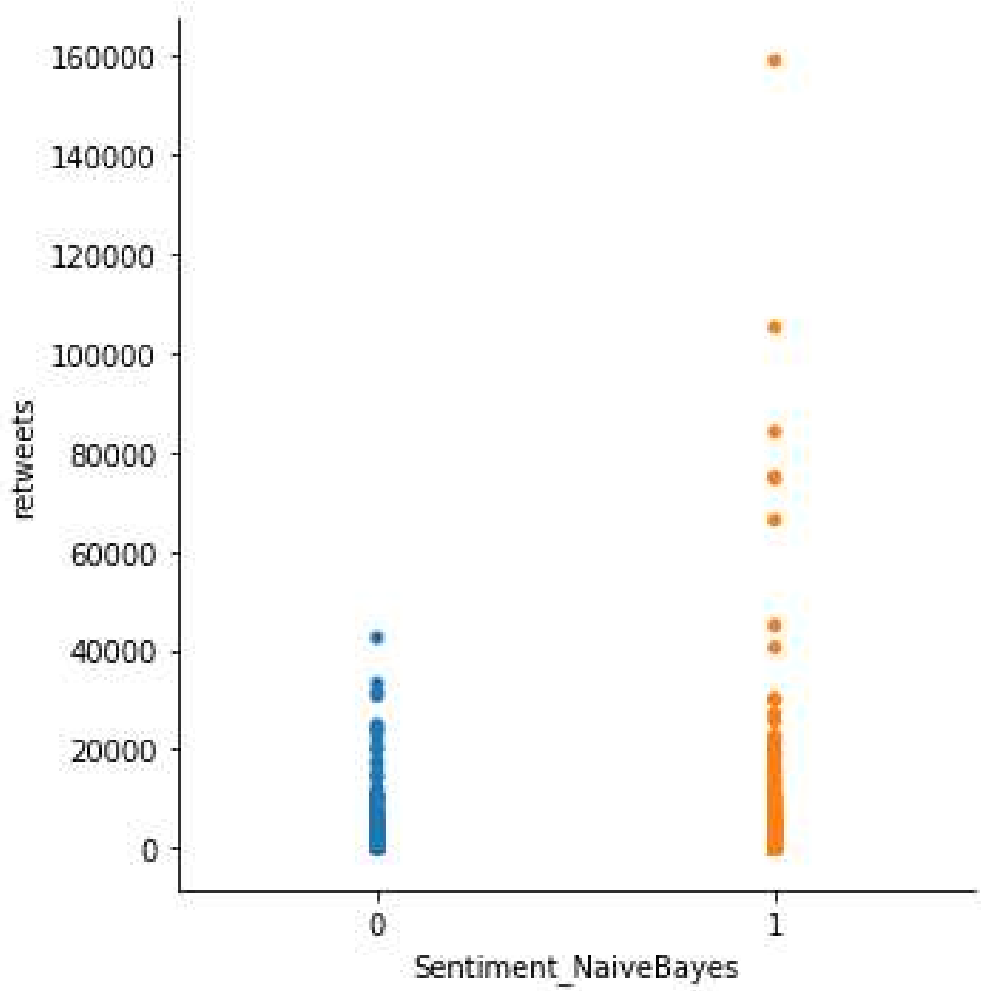


Figure 5.6.4

In Figure 5.5, “0” represents a negative sentiment, while “1” represents a positive sentiment in this figure. The rate of retweets is not significantly affected by a tweet being negative or positive. Still, positive tweets have a higher retweet rate, and it is more likely to find tweets with a higher retweet being positive. However, it is also essential to consider the strength of the sentiment classifier. Generally, due to the nature of tweets published by businesses, there tends to be more emphasis on creating a positive and inviting tone.

5.6.5 Retweets and Hashtags

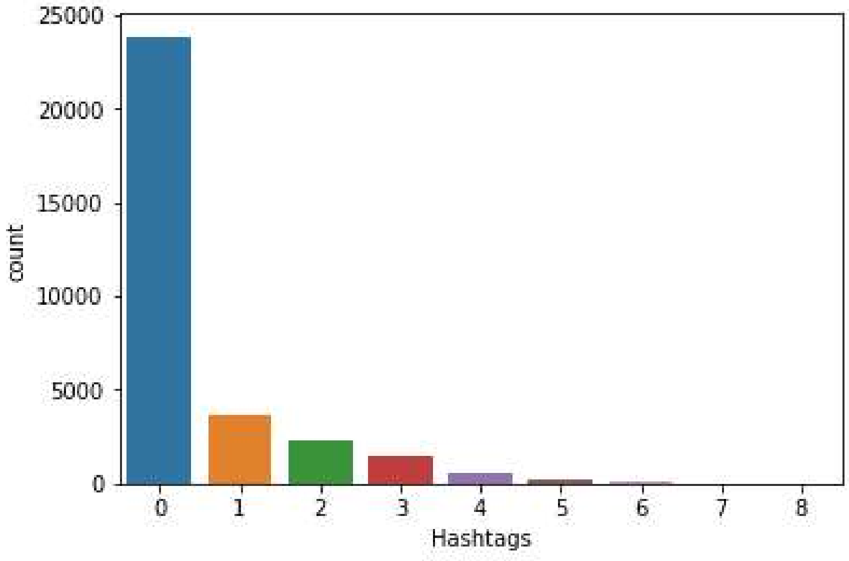


Figure 5.6.5.1

The exploratory data analysis showed a higher level of retweets for tweets with no hashtags, unlike [30], which found that hashtags influenced a retweet rate. In both experiments, this discovery remained true. Since these are foreign brands, it could be that hashtags are not as necessary within that context. At the same time, putting it within the context of businesses at home would mean that digital marketers can explore developing content that doesn’t focus on hashtags but on the content, day, or time of the post. The figures display the count of hashtags and plot them against retweets.

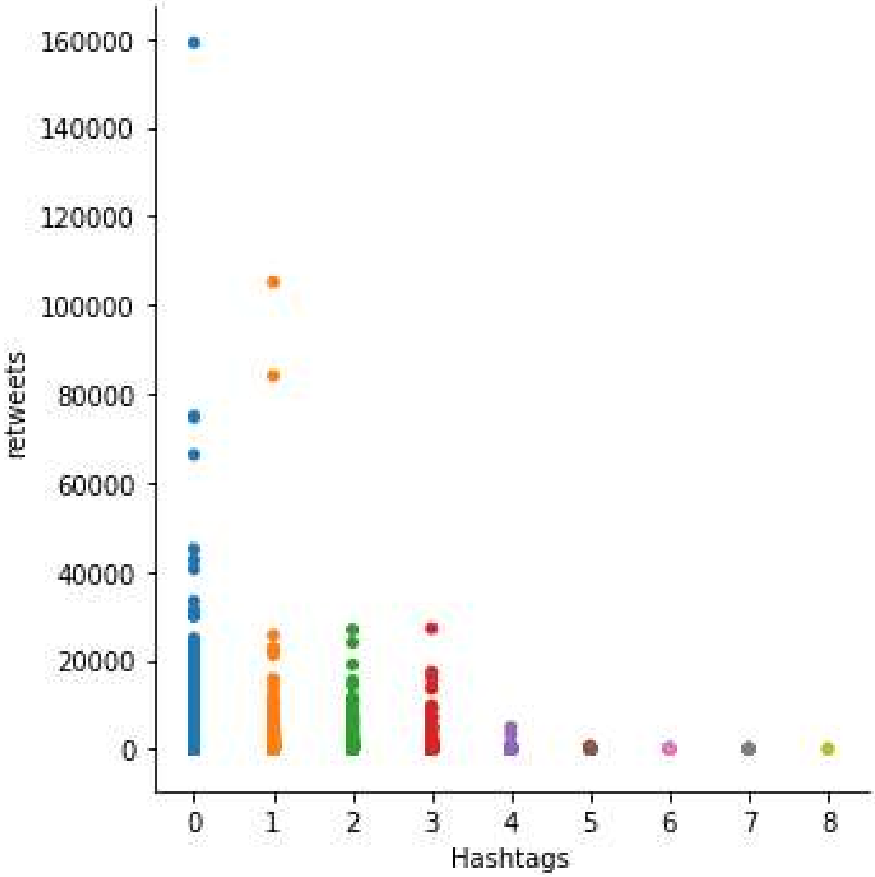

160000
140000
120000
100000
80000
60000
40000
20000
0
retweets
0
1
2
3
4
5
6
7
8
Hashtags

Figure 5.6.5.2

Chapter 6: Conclusion and Further Work

6.1 Summary of Conclusions

This paper aimed to explore the types of regression models that would work best for predicting tweet engagement using user-related and post-related features. The prediction problem was treated solely as a regression problem rather than a classification problem by identifying the count of engagement. While this method sought to determine an accurate model with Instagram and Twitter data, Instagram proved unusable given scope and security constraints. Twitter proved usable with its accessibility to public tweets of business accounts.

Of all the models used for training and testing, none proved to produce an acceptable accuracy score to replicate for real use when the dataset had outliers present. Two experiments took place where the first one involved testing and training on 4,500 data points and the second experiment involved testing and training on 32,000 data points. Each of the experiments went through the same data cleaning, transformation, feature selection, testing, and training. While there was a slight increase in the accuracy level from experiment one to two, the margin was too insignificant to yield acceptable results. Separating the tweet related features each model used did not help achieve better results.

However, removing outliers from the larger dataset preserved two-thirds of the data. This isolated experiment, it showed that removing a significant number of outliers can still yield somewhat acceptable results. Though there was still a unacceptable error between predicted and real values, this dataset outperformed both previous datasets with outliers present. Thus, treating this problem solely as a regression did not yield better results than in [24,31] which adopted the problem as a classification problem.

6.2 Limitations

Though the classification models explored in this research did not yield better results than if this was to be treated as a classification problem, the following are the limiting factors that are believed to have impacted and influenced the results of the study negatively:

1. Though the size of the retrieved dataset is considered large enough according to typical base level experiments, the second experiment with 32,000 data points was still not enough to create enough accuracy for replication.
2. The lack of accessibility to public information, especially public metrics on Instagram, restricted analysis of its past posts. Though this may have been curbed by directly scraping Instagram's website, it is not acceptable according to their terms of use.
3. It cannot be ignored that a large immeasurable social context affects the amount of retweets that a post will get. Since it cannot be factorised into a mathematical equation, it should be primarily considered when formulating a position for publication.

6.3 Suggestions for Further Work

The following measures can be adopted and considered to improve the work done in this paper in line with only Twitter since it was the medium that proved feasible enough:

1. Increasing the amount of data and the variability of the type of data used, instead of focusing on business-to-business accounts, other forms of businesses could result in a higher accuracy result and a lower mean squared error.

2. Another feature that could be considered is the categorisation of the tweets, such as sports, health, fashion, etc.
3. To make the sentiment meaningful, a dictionary of Ghanaian vernacular, informally known as pidgin, can be created with an unsupervised model used only on the sentiment analysis of tweets to understand the post subjectivity.
4. Understanding the context of the tweets and how it also affects the retweet rate, such as which tweets are more generic and focused on current trends, tweets featuring a celebrity, or tweet's more in line with the company's product or service.

References

[1] Youlong Zhuang and Albert L. Lederer. 2014. An Instrument for Measuring the Business Benefits of E-Commerce Retailing. *International Journal of Electronic Commerce* 7, 3 (December, 2014), 65-99. DOI: https://doi.org/10.1080/10864415.2003.11044274

[2] You Lingxian, Kou Jiaging and Wang Shihuai. 2019. Online Retail Sales Prediction with Integrated Framework of K-mean and Neural Network. *ICEME 2019: Proceedings of the 2019 10th International Conference on E-Business, Management and Economics,* July 15, 2019. 115-118. https://doi.org/10.1145/3345035.3345048

[3] Hilde A.M. Voorveld and Guda van Noort. 2014. Social Media in Advertising Campaigns: Examining the Effects on Perceived Persuasive Intent, Campaign and Brand Responses. *Journal of Creative Communications* 9, 3 (November, 2014), 253-258. DOI: https://doi.org/10.1177%2F0973258614545155

[4] Brian E. Weeks and R. Lance Holbert. 2013. Predicting Dissemination of News Content in Social Media: A Focus on Reception, Friending and Partisanship. *Journalism & Mass Communication Quarterly* 90, 2 (April, 2013), 212-232. DOI: https://doi.org/10.1177%2F1077699013482906

[5] Sarah Shugars and Nicholas Beauchamp. 2019. Why Keep Arguing? Predicting Engagement in Political Conversations Online. *Sage Open* 9, 1 (March 2019). DOI: https://doi.org/10.1177%2F2158244019828850

[6] Ibrahim Kircova, Yilmaz Yaman and Sirin Gizem Kose. 2018. Instagram, Facebook or Twitter: Which Engaged Best? A Comparative Study of Consumer Brand Engagement

and Social Commerce Purchase Intention. *European Journal of Economics and Business Studies* 10, 1 (March 2018), 279 – 289. DOI: http://dx.doi.org/10.26417/ejes.v10i1.p279- 289

[7] Angeline Close Scheinbaum. 2016. Digital Engagement: Opportunities and Risks for Sponsors: Consumer View Point and Practical Considerations For Marketing via Mobile and Digital Platforms. *Journal of Advertising Research* 56, 4 (Decemeber, 2016), 341- 345. DOI: http://dx.doi.org/10.2501/JAR-2016-040

[8] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, Roger Zimmerman. 2018. A Multimodal Approach to Predict Social Media Popularity. 2018 IEEE Conference on Multimedia Processing and Retrieval, 180 - 195. DOI: 10.1109/MIPR.2018.00042

[9] James G. Webster. 2014. *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. The MIT Press. Cambridge, MA.

[10] James G. Webster. 2010. User Information Regimes: How Social Media Shape Patterns of Consumption. *Northwestern University Law Review* 104, 2, (March 2010), 593 – 612.

[11] Youlong Zhuang and Albert L. Lederer. 2014. An Instrument for Measuring the Business Benefits of E-Commerce Retailing. *International Journal of Electronic Commerce* 7, 3 (December, 2014), 65-99. DOI: https://doi.org/10.1080/10864415.2003.11044274

[12] You Lingxian, Kou Jiaging and Wang Shihuai. 2019. Online Retail Sales Prediction with Integrated Framework of K-mean and Neural Network. *ICEME 2019: Proceedings of the 2019 10th International Conference on E-Business, Management and Economics,* July 15, 2019. 115-118. https://doi.org/10.1145/3345035.3345048

[13] Hilde A.M. Voorveld and Guda van Noort. 2014. Social Media in Advertising Campaigns: Examining the Effects on Perceived Persuasive Intent, Campaign and Brand Responses. *Journal of Creative Communications* 9, 3 (November, 2014), 253-258. DOI: https://doi.org/10.1177%2F0973258614545155

[14] Brian E. Weeks and R. Lance Holbert. 2013. Predicting Dissemination of News Content in Social Media: A Focus on Reception, Friending and Partisanship. *Journalism & Mass Communication Quarterly* 90, 2 (April, 2013), 212-232. DOI: https://doi.org/10.1177%2F1077699013482906

[15] Sarah Shugars and Nicholas Beauchamp. 2019. Why Keep Arguing? Predicting Engagement in Political Conversations Online. *Sage Open* 9, 1 (March 2019). DOI: https://doi.org/10.1177%2F2158244019828850

[16] Lorenzo Vecchi and Eliane Francisco-Maffezzoli. 2020. Predicting Digital Engagement on Instagram. *Proceedings of the European Marketing Academy*,

May 26-29, 2020, Budapest. http://proceedings.emac-online.org/pdfs/A2020-63299.pdf

[17] Kholoud Khalil Aldous, Jisun An and Bernanrd J. Jansen. 2019. View, Like, Comment, Post: Analyzing User Engagement by Topic at 4 Levels across 5 Social Media Platforms for 53 News Organizations. *Proceedings of the International AAAI Conference on Web and Social Media*, June 11-14, 2019, Munich, Germany. 47-57. https://ojs.aaai.org/index.php/ICWSM/article/view/3208

[18] Dominque Jackson. 2020. 11 Facebook Metrics Every Brand Needs to Track. (August 2020). Retrieved October 18, 2021 from https://sproutsocial.com/insights/facebook-metrics/

[19] Barry Levine. 2017. This new AI-powered social media marketing tool can predict engagement or write the post for you. (March 2017). Retrieved March 24, 2017 from https://martech.org/new-ai-powered-social-marketing-tool-can-predict-engagement-write-post/

[20] Ibrahim Kircova, Yilmaz Yaman and Sirin Gizem Kose. 2018. Instagram, Facebook or Twitter: Which Engaged Best? A Comparative Study of Consumer Brand Engagement and Social Commerce Purchase Intention. *European Journal of Economics and Business Studies* 10, 1 (March 2018), 279 – 289. DOI: http://dx.doi.org/10.26417/ejes.v10i1.p279- 289

[21] Angeline Close Scheinbaum. 2016. Digital Engagement: Opportunities and Risks for Sponsors: Consumer View Point and Practical Considerations For Marketing via Mobile and Digital Platforms. *Journal of Advertising Research* 56, 4 (Decemeber, 2016), 341- 345. DOI: http://dx.doi.org/10.2501/JAR-2016-040

[22] Allan Jay. Number of Social Media Users in 2022/2023: Demographics . *Finances Online*.Retrieved April 26, 2022 from https://financesonline.com/number-of-social-media-users/

[23] Paul Gil. 2021. What Is Twitter & How Does It Work? *LifeWire*. Retrieved April 26, 2022 from https://www.lifewire.com/what-exactly-is-twitter-2483331

[24] Rangsipan Marukatat. 2016. A retweet prediction of Thai tweets. *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)* (2016), 1000-1003. DOI: https://doi.org/10.1109/imcec.2016.7867361

[25] Gokhan Aydin, Nimet Uray, and Gokhan Silahtaroglu. 2021. How to Engage Consumers through Effective Social Media Use—Guidelines for Consumer Goods Companies from an Emerging Market. *Journal of Theoretical and Applied Electronic Commerce Research* 16, 4 (2021), 768-790. DOI: https://doi.org/10.3390/jtaer16040044

[26] Later. 2019. Instagram Marketing: The Definitive Guide. Retrieved December 12, 2021 from https://later.com/instagram-marketing/

[27] Alena Soboleva. 2018. Marketing With Twitter: Investigating Factors That Impact on the Effectiveness of Tweets. Retrieved April 26, 2022 from

https://researchdirect.westernsydney.edu.au/islandora/object/uws%3A47360/datastream/PDF/download/citation.pdf

[28] Michael Stelzner. 2021. 2021 Social Media Marketing Industry Report. Retrieved April 26, 2022 from socialmediaexaminer.com/report/

[29] Yan Yan, Zhaowei Tan, Xiaofeng Gao, Shaojie Tang, and Guihai Chen. 2016. STH-Bass: A Spatial-Temporal Heterogeneous Bass Model to Predict Single-Tweet Popularity. *Database Systems for Advanced Applications* (2016), 18-32. DOI: https://doi.org/10.1007/978-3-319-32049-6_2

[30] Ke Wang, Mohit Bansal, and Jan-Michael Frahm. 2018. Retweet Wars: Tweet Popularity Prediction via Dynamic Multimodal Regression. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2018). DOI: https://doi.org/10.1109/wacv.2018.00204

[31] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. *2010 IEEE Second International Conference on Social Computing* (2010). DOI: https://doi.org/10.1109/socialcom.2010.33

[32] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, Roger Zimmerman. 2018. A Multimodal Approach to Predict Social Media Popularity. 2018 IEEE Conference on Multimedia Processing and Retrieval, 180 - 195. DOI: 10.1109/MIPR.2018.00042.

[33] Jason W. Osborne. 2004. The Power of Outliers (and Why Researchers Should Always Check for Them). Practical Assessment, Research and Evaluation 9, 6 (January 2004),

1-8. Retrieved April 25, 2022 from https://www.researchgate.net/publication/242073851_The_Power_of_Outliers_and_Why_Researchers_Should_Always_Check_for_Them

[34] Johannes Fürnkranz, Philip K. Chan, Susan Craw, Claude Sammut, William Uther, Adwait Ratnaparkhi, Xin Jin, Jiawei Han, Ying Yang, Katharina Morik, Marco Dorigo, Mauro Birattari, Thomas Stützle, Pavel Brazdil, Ricardo Vilalta, Christophe Giraud-Carrier, Carlos Soares, Jorma Rissanen, Rohan A. Baxter, Ivan Bruha, Rohan A. Baxter, Geoffrey I. Webb, Luís Torgo, Arindam Banerjee, Hanhuai Shan, Soumya Ray, Prasad Tadepalli, Yoav Shoham, Rob Powers, Yoav Shoham, Rob Powers, Geoffrey I. Webb, Soumya Ray, Stephen Scott, Hendrik Blockeel, and Luc De Raedt. 2011. Mean Squared Error. Encyclopedia of Machine Learning (2011), 653-653. DOI:https://doi.org/10.1007/978-0-387-30164-8_528

[35] Johannes Fürnkranz, Philip K. Chan, Susan Craw, Claude Sammut, William Uther, Adwait Ratnaparkhi, Xin Jin, Jiawei Han, Ying Yang, Katharina Morik, Marco Dorigo, Mauro Birattari, Thomas Stützle, Pavel Brazdil, Ricardo Vilalta, Christophe Giraud-Carrier, Carlos Soares, Jorma Rissanen, Rohan A. Baxter, Ivan Bruha, Rohan A. Baxter, Geoffrey I. Webb, Luís Torgo, Arindam Banerjee, Hanhuai Shan, Soumya Ray, Prasad Tadepalli, Yoav Shoham, Rob Powers, Yoav Shoham, Rob Powers, Geoffrey I. Webb, Soumya Ray, Stephen Scott, Hendrik Blockeel, and Luc De Raedt. 2011. Mean Absolute Error. Encyclopedia of Machine Learning (2011), 652-652. DOI:https://doi.org/10.1007/978-0-387-30164-8_525

[36] Decision Tree Regression. Retrieved April 27, 2022 from https://www.saedsayad.com/decision_tree_reg.htm

[37] Shaunak De, Abhishek Maity, Vritti Goel, Sanjay Shitole, and Avik Bhattacharya. 2017. Predicting the popularity of instagram posts for a lifestyle magazine using deep learning. 2017 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA) (April 2017), 175-178. DOI:https://doi.org/10.1109/cscita.2017.8066548

[38] Priya Pedamkar. Support Vector Regression - eduCBA. Retrieved April 27, 2022 from https://www.educba.com/support-vector-regression/

[39] Soroosh Tayebi Arasteh, Mehrpad Monajem, Vincent Christlein, Philipp Heinrich, Anguelos Nicolaou, Hamidreza Naderi Boldaji, Mahshad Lotfinia, and Stefan Evert. 2021. How Will Your Tweet Be Received? Predicting the Sentiment Polarity of Tweet Replies. 2021 IEEE 15th International Conference on Semantic Computing (ICSC) (2021). DOI:https://doi.org/10.1109/icsc50631.2021.00068

[40] P M Ashok Kumar, K.Guru Charan, G.B V Sai Kumar, K. Amith, and K.Sai Krishna. 2021. Real-Time Hashtag based Event Detection Model with Sentiment Analysis for Recommending user Tweets. 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (2021), 1437-1444. DOI:https://doi.org/10.1109/icicv50876.2021.9388426

[41] Rattasit Sermsai and Sirisup Laohakiat. 2019. Analysis and Prediction of Temporal Twitter Popularity Using Dynamic Time Warping. 2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE) (2019), 176-184. DOI:https://doi.org/10.1109/jcsse.2019.8864227

[42] Personal Information Disclosure and Privacy in Social Networking Sites. Retrieved April 27, 2022 from https://core.ac.uk/download/pdf/80334091.pdf

[43] Moreno Mancosu and Federico Vegetti. 2020. What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. Social Media + Society 6, 3 (2020). DOI:https://doi.org/10.1177/2056305120940703

[44] Rohit Dwivedi. 2021. What Is Naive Bayes Algorithm In Machine Learning? | Analytics Steps. Retrieved April 27, 2022 from https://www.analyticssteps.com/blogs/what-naive-bayes-algorithm-machine-learning

[45] Twitter Sentiment Analysis with Naive Bayes 85%acc | Kaggle. Kaggle. Retrieved April 27, 2022 from https://www.kaggle.com/lykin22/twitter-sentiment-analysis-with-naive-bayes-85-acc