SUNWAY UNIVERSITY

Capstone Project Report

Sarcasm Detection From Cyber Bullying and Cyber Aggressive Cases Online Through Social Media Analytics

by

Yeo Yee Wen
(18076950)

"Bachelor of Science (Hons) Information Systems (Business Analytics)"
:

Supervisor : (AP Dr Angela Lee Siew Hoong)

Date : 2 July, 2021

Project Title : Sarcasm Detection From Cyber Bullying and Cyber Aggressive Cases Online Through Social Media Analytics
Date : 1st July 2021
Student : Yeo Yee Wen
Supervisor : Assoc Prof Dr Angela Lee Siew Hoong

Abstract

Sarcastic Cyberbullying Detection circulated on social media platforms has lead to unfavourable outcomes though hate speech due to the widespread usage of online social networks. As such, the importance of identifying sarcastic cyberbullying cases are crucial in order to maintain the peace and sovereignty of a nation. There are manual fact checking websites that exist but they are limited and unable to cope with the fast moving and large volumes of online news circulating on social media. The solution to this lies in automated factchecking applications which can be automated and scaled to suit the large volume of data. There are still limitations to this as readily available datasets lack multi-dimension information that could be utilized to improve an accuracy of a predictive model's performance when detecting sarcastic cyberbullying cases online. Since the objectives of sarcastic cyberbullying detection is to detect a negative comment from the messages it receives by a user, sarcastic cyberbullying online in posts, comments, or even in articles display distinct linguistic and psycholinguistic features that affects the human brain and emotions in a certain way. Hence, this paper employs using social media analytics to investigate and discover new attributes that can be derived from news texts such as their linguistic and psycholinguistic features that are influential in deriving sarcastic cyberbullying cases. This paper is also an expansion of previous work that used attributes derived from Twitter data. The best machine learning model derived from this paper achieved an accuracy of 77.35% in detecting sarcastic cyberbullying.

Keywords: Sarcasm Detection, Cyberbullying, Cyber Aggression, Social Media Analytics, Classification Modelling, Machine learning, Social Network Analysis

Table of Contents

List of Figures

List of Tables

1 Chapter 1: Introduction

This following chapter consists of seven subsections namely background information, problem statement, research aim, research questions, research objectives & research goals. The first section describes the background of social media, sarcasm and cyberbullying. The other five sections discusses the issues, aim and desire that is being addressed by this project in detailed of the proposed area in the existing conditions.

1.1 Background Information

Ever since the 21st century, the engagement of social media for knowledge transfer & daily communication as a funnel between people has much become very popular (Subrahmanyam et al., 2008; Lenhart et al., 2009; Duggan et al., 2015). Social networks allow people to convey short and long distance communications in a variety of forms via text, images and videos as expressions to communicate with users. Although social interaction via social networks can be mutually beneficial, either towards the improvement of relationships among peers (Ellison, 2007). The utterance in conveying ideas and opinions via social media may lead towards positive outcomes in respect to users being abled to enrich knowledge transfer, diversity and understanding from different viewpoints. However, negative scenarios due to intolerant behaviour via social media can lead towards social issues such as online bullying via the publishing of anonymous hate speech, discriminatory comments, and online harassment on targeted victims (Cortis and Handschuh, 2015).

According to (Privitera and Campbell, 2009), traits that indicate online bullying are through the frequency of time duration in exchanging these harmful messages. As a result, messages that depict online bullying are through the conduct of offensive words either through irony or mockery. This is known as sarcasm. Sarcasm are parts of human interaction which can be indistinct and illustrated in various methods. For example, sarcasm can be identified through verbal distributions, where humans tend to utilize voice and facial expressions to identify if people are being sarcastic or not. Depicting sarcasm is very difficult to understand and identify as it revolves around the usage of user opinions. Therefore, cyberbullying detection is a challenge especially for researchers and practitioners whom are trying to identify the cause of online bullying while also raising awareness. This is why, researchers via online social media platforms have identified share rules to depict if a sentence is sarcastic or not. For example, twitter utilizes “#sarcastic” , “#mocking” , “#irony” and

more as hashtags to be utilized as benchmarks for detecting sarcasm. Therefore, these benchmarks can be utilized for the detection of cyberbullying cases.

The main motivation behind this research is to improve the accuracy of sentiment analysis through the exploring phenomenon of cyberbullying in a more in-depth manner, specifically, in respect towards social media posts and comments which contains negative words that can be then classified for cyberbullying cases. As a result, the classification can be utilized are explored as decision support data which later on when researchers continue with this study, they would eventually be able to perform further categorisation to discover the effect of these messages on top of the social media users. (Riffet et al., 2014).

1.2 Problem Statement

Cyberbullying is an ever growing controversy which potentially has affected individuals online. It is known as a phenomena that is caused in the internet, mainly communicated through cell phones or other electronic devices to wilfully hurt or harass someone. Some may refer this as hate speech in social media. As a result towards the continuous growth and popularity of social media platforms such as Snapchat, Twitter, Instagram, Reddit, Facebook & more. Therefore, cyberbullying is growing to become a more rampant problem. Therefore, the call for cyberbullying prevention & efforts have been acknowledged through procedures, policies, awareness events and more to help deflate this issue (Espelage, D. L., and Hong, J. S. 2017). However, recent reports have indicated the increase in cyberbullying cases Karmakar, S., & Das, S. (2020, August) especially through social media during the covid pandemic period Babvey et al. (2020).

This indication shows that victims of cyberbullying often are suffering more from mental health disorders such as loneliness, depression, low self-esteem, & anxiety issues, and etc (Parris et al, 2020). In more cataclysmic issues, scenarios occur such as victims of online bullying attempt suicide or suffer from interpersonal problems. Moreover, cyberbullying cases are unrestricted compared to traditional bullying (Squicciarini et al, 2015). In traditional efforts to reduce cyberbullying, the development of policies and standards via campaigns are used to adhere bullying behaviour through indication via profane word lists and facial expressions. However, these mechanisms are not effective in social media as they do not scale well. Therefore, this requires a high demand & usage of a framework to more accurately identify, detect & improvise these ever growing cyberbullying cases.

1.3 Research Aim & Goal

The main aim of this research is to research, develop and implement a cyberbullying data analytics model to detect sarcastic cyberbullying messages via online social media. To achieve this, we perform data collection, feature engineering and machine learning model training to discover sarcastic patterns in cyberbullying messages. The ending goal and purpose of this project is to develop/build a personalised web-based dashboard website that forecasts cyberbullying cases online via social media.

1.4 Research Objectives

The objectives of this project are as follows:

- To formally identify the relationships, causes & problems of cyberbullying in social media.

- To formally verify that sarcastic sentiments have a significant difference between normal bullying posts by comparing their sentiment score distribution through predictive accuracies using machine learning.

- Increase awareness via forecast of cyber bullying & aggressive causes, effects & reinforced repercussions from it.

- Improve public knowledge and understanding of cyberbullying effects as a service via web online.

2 Chapter 2: Literature review

2.1 Sarcasm Detection Definition

Through linguistic concepts, sarcasm detection has been an intensive and popular topic of study. Ghosh et al. (2016) studied sarcasm through semantic modelling of sentences using neural networks, where figurative language is understood through sentiment, belief and speaker intention. The paper separated mockery into four distinct sorts, perlocutionary, propositional, 'as'- prefixed and lexical, by examining how each type varies in the way wherein meaning reversal is showed. Kreuz and Caucci (2007) requested college students to read extracts from distributed works and afterwards characterize the content as being sarcastic or not. The investigation found that lexical factors, for example, contributions (for example "hmm", "gosh"), certain equation based articulations (for example "much obliged", "great job") and redundancies (for example "great, simply awesome") were critical indicators of the members' appraising of irony remarks.

A part from that, research in programmed sarcasm identification details the issue as a stratification task, where given textual data, the objective is to effectively arrange it in a way where the focus is purely on sarcasm. Numerous methodologies base their strategies based on semantics as a benchmark for sarcasm investigation. (Sreelakshmi, K., & Rafeeque, P. C, 2018, July) investigated that the impact of utilizing lexical and pragmatic factors in building AI machine learning models to recognize sarcasm via text. Moreover, social media reviews that utilizes lexicon features such as bigrams, trigrams, n-gram, #hashtags and more include a lot of interjections and sometimes punctuations mentioned by (Bharti, S. K., Pradhan, R., Babu, K. S., & Jena, S. K, 2017). Comparing towards other pragmatic features comprehending positive sentiments, negative situations, negative sentiment and positive situation under Twitter's '@user' mention, which embarks tweet being replied by another tweet. (Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B, 2020) illustrated different literal and figurative definitions towards understanding user opinions by reviewing different machine learning algorithms to model and feature text based on punctuation marks, emojis, unexpectedness (e.g. temporal and contextual imbalance), style (e.g. character-grams, skipgrams, polarity skip-grams) and emotional scenarios (e.g. imagery, pleasantness). (Bharti et al, 2016) acknowledged that tweets streamed in real time can be used as a signal for the explicit individuals to understand such sarcasm effectively as big data frameworks aid in the processing between hyperbolic and non-hyperbolic languages.

However, comparing various wide-ranging sarcasm detection approaches, different methods and techniques present different understanding and interpretation incorporated for illustrating the most effective model for sarcasm detection. Chaudhari, P., & Chandankhede, C. (2017, March) targets specific types of sarcasms that focus according to the approach where their algorithm can understand and comprehend meaning from their the textual depicts. Riloff et al. (2013) proposed a bootstrapping algorithm that is capable of learning positive sentiments and negative emotional phrases from sarcastic tweets and this can be used to detect sarcasm. Table 1 below shows a subset of positive phrases and negative situations learned by their model. This method achieved an F1 score of 0.51 on a dataset of 3,000 tweets.

Word Type (Count)	**Phrase (Examples)**
(Positive Phrases)	"missed, loves, enjoy, cant wait, excited, wanted, can't wait, get, appreciate, decided, loving, ..."
(Positive Predicative Expressions)	"great, so much fun, good, so happy, better, my favourite thing, cool, funny, nice, always fun, ..."
(Negative Situations)	"being ignored, being sick, waiting, feeling, waking up early, being woken, fighting, staying, writing, being home, cleaning, not getting, crying, sitting at home, being stuck, ..."

Table 1: Riloff et al (2013). Positive & Negative bootstrapping algorithm

According to their research, there are several limitations to their approaches. Firstly, There are several limitations to these two approaches. Firstly, Riloff et al. (2013) addresses only subsequent information regarding to the lexicon of the tweets. However, sarcasm cannot be expressed via positive sentiments with a negative situation only. Rather, they need to work vice versa as both ways can capture meaningful insights. For example, sarcasm is not only understood via the word sentiments, but also emojis and hashtags need to be considered especially in social media platforms. For example, a smiley face with the words "You are a piece of work" can indicate a positive sentiment remark but understood as being negative. For this research, these limitations are understood. Therefore, a bootstrapping algorithm will not be used and a custom method shall be applied to obtain the sentiments where we crawl the data online and process each of the sentences to obtain further accuracy regarding the features of the text.

2.2 Importance Of Bullying Detection On The Web

Through literature studies, sarcasm detection cyberbullying based upon from sociological and psychological perspectives are key primary indicators through computational methods, to efficiently detect cyberbullying in different social media platforms. The rapid growth of social networking, online communities are entry requirements for the primary growth of online bullying activities which, in the worst scenario, may result in users or people going through suicidal attempts due to psychological disorders. According to Kim et al. (2020), cyberbullying and victimization is mostly common among the youth. It is still inaccurate what are the true reasons, but adolescent suicidal behaviours result due to various reasons, such as loneliness which ultimately could lead to depression. Therefore, significant understanding of cyberbullying is important so that people can understand the theory of cyberbullying effects thus intervening upon their issues before anything happens to a family's beloved Pabian et al. (2014). Due to a lack of existing datasets as well, very few studies on the identification of cyberbullying online have been conducted. At the moment, no effort has been done to prevent cyberbullying that is accurate and dependable. In this study, we will look at cyberbullying in a holistic way and research what that has been done to detect it.

2.3 Approach

In this section, it formally defines the techniques in how the problem is tackled through a formal approach. The core idea behind our approach is that sarcasm is expressed through a collocation between positive and negative contrasting sentiments. A hypothesis can be illustrated where if opposite sentiment scores are of different variations in the text, the likelihood where a sentence may be sarcastic is higher. Therefore, to depict a model for that, techniques through decomposition of a sentence through a binary tree of phrases on top of a sentiment can show the different sentiment annotations.

Based on the tree diagram in Figure 1, we can feature engineer based on the following:

- sentence sentiment score
- maximum sentiment score
- minimum sentiment score
- sentiment score range
- adjacent sentiment contrast score.

The phrase “I enjoy how exhausting my Mondays are weekly.” to explain what each feature represents. Figure 1 is a visual representation of how the sentence is decomposed using Baly et al. (2017) model.

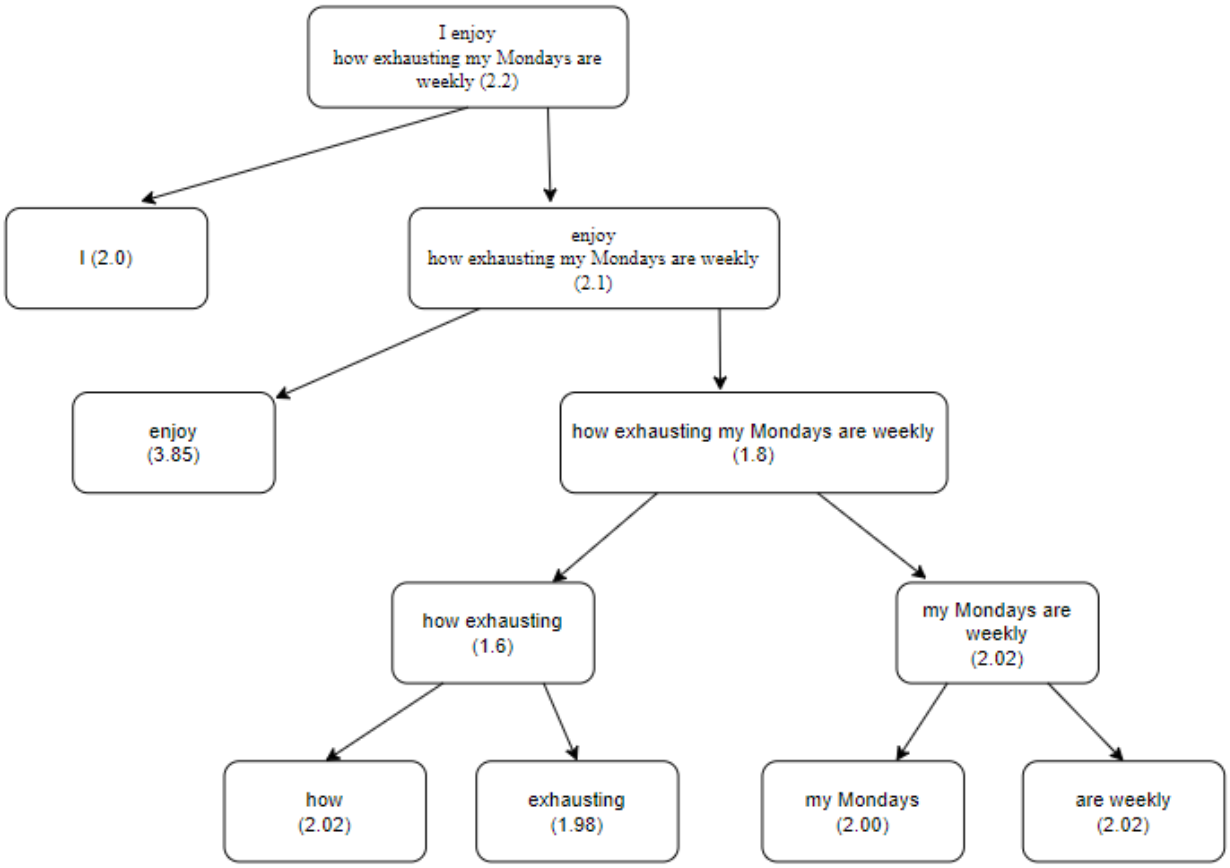


Figure 1: Sentiment annotated tree of phrase “I enjoy how exhausting my Mondays are weekly.”

Based on figure 1, Each node contains a phrase and its sentiment score in parentheses. Sentiment score ranges from 0 to 4, where 0 is very negative, 2 is neutral, and 4 is very positive. Based on the diagram above, the sentiment score represents the node in which is applied of the whole sentence after separations. The illustrated maximum and minimum sentiment scores indicate the highest and lowest sentiment scores of the whole tree. In our example, the phrase “enjoy” has the highest with 3.85 and the phrase “how exhausting” has the lowest with 1.6. The adjacent sentiment contrast score is the maximum difference of sentiment scores of adjacent phrases or, in other words, the difference of sentiment scores of two children nodes with the same parent. This feature is designed to detect if there are phrases of opposite sentiments juxtaposed together. The manner in which sarcasm is communicated is frequently unique across different sites and high performance algorithms. In our examination, we crawl data from social media and we engineer extra slant highlights explicit to

every stage to address the various ways sarcasm might occur on these sites. We model sarcasm detection to detect cyberbullying via a binary classification task to decide whether a piece of text is sarcastic or not sarcastic. Using the features described, we train machine learning classifiers to detect sarcasm and we evaluate the performance by comparing classification accuracies with results of continuous crawling to evaluate our model. Below is a diagram which illustrates the whole process of this research and implementation.

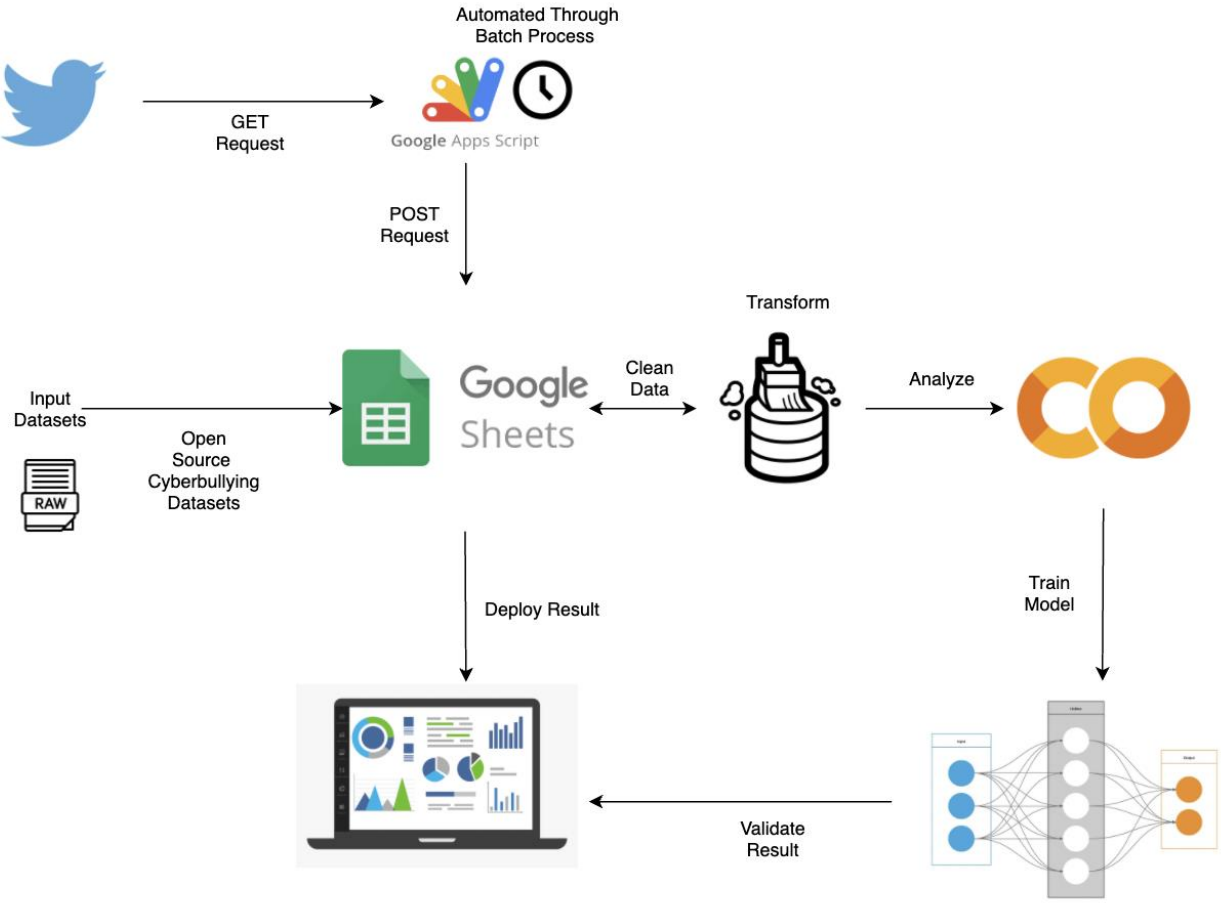


Figure 2: Sarcastic Cyberbullying Research Process & Implementation

The above architecture in Figure 2 resembles the implementation of our work. Data from Twitter is crawled using Google's AppScript through the utilization of the Twitter API along with the usage of open source cyberbullying datasets from Mendeley Data. Data from the following sources are extracted and stored into google sheets which is then cleaned, transformed & analyzed in google colab where models are trained using various algorithms in the application of machine learning to validate the automated result. The cleaned & transformed analysis is also fed back to google sheets on a separate tab where visualizations are generated to be deployed on the web.

2.4 Proposed Framework – (SDCB)

In this chapter, we provide the introduction of a proposed framework, SDCB (Sarcasm Detection Cyber Bullying). The following below illustrates the workflow of this framework from (2.5 – 2.9). The (SDCB) framework is a standardize framework that is created to prioritize its application to be applied in our Methodology.

2.5 Social Media Data Collection

In this section, Twitter API is used along with open sourced cyberbullying datasets to collect tweets from twitter for sarcastic cyberbullying detection.

2.6 Social Media Data Cleaning

In this section, Tweets streamed using the Twitter API will be cleaned in multiple ways.

- Stop word removal, Tokenization, Segmentation & More depending on the unstructured-ness of the data.

2.7 Social Media Data Modelling

In this section, the data will be stored in the database of choice. The choice of data storage that we have opt to utilize is Google Sheets.

2.8 Feature Engineering

In this section, sentiment features extracted from the text is used to filter & capture certain characteristics specific to sarcasm and cyberbullying. We have derived 4 characteristics to derive sarcastic cyberbullying which is through aggression, social attacks, toxicity & Racism.

2.9 Machine Learning

The machine learning technique that will be used to detect cyberbullying associations with online bullying is through binary task classification techniques. This later on will be evaluated through hyperparameter optimization testing to fine tune score distribution based on testing performance.

3 Chapter 3: Comparison Between Different Supervised Classification ML Models

3.1 Supervised Machine Learning

Machine learning in its most normalized form uses algorithms that are programmed for analysing inputted data, making predictions within acceptable limits, and learn and optimize one's work (Souad, Taleb & Adla, Abdelkader, 2020). While providing new data, accuracies of machine learning models are often improved upon the usage of these algorithms. There are several variations on how one would choose group & amplify machine learning algorithms, but they can be divided into sub components, depending on the purpose and how one would teach the basic machine. These three categories are supervised, unsupervised and semi-supervised. Through supervised machine learning, algorithms are labelled upon a training dataset which is used first to train the basic its basic form. Next, this trained algorithm is fed to an unlabelled test dataset, and classifies them into groups. Supervised learning algorithms are suitable for two problems: classification or regression problems. For classification problems, their default variables are often discrete and (Alloghani, Mohamed & Al-Jumeily Obe, Dhiya & Mustafina, Jamila & Hussain, Abir & Aljaaf, Ahmed, 2020). For example, this variable can be divided into various components and dissected into groups, such as 'bullying' or 'non-bullying'. For example, even in the case of "irony" or "non-irony". The meaning is that the corresponding output variable is the actual value of the regression problem, such as the risk that sarcastic cyberbullying online will be affected by characteristics from sarcasm online on social media. The following subsections briefly describe teacher machine learning algorithms commonly used for predicting cyberbullying.

Classification is a supervised machine learning technique used to categorize data into pre-determined classes or outcomes. In this paper, a binary classification model will be used to classify sarcastic bullying tweets into either one of the two outcomes: bullying or non - bullying. The dataset built from online sources are labelled as 1 = bullying or 0 = non-bullying, is first split into a training set and a validation set Upon this, a model is then fitted with the training set to produce a mathematical model that is able to predict the target variable. Subsequently, the validation set is used to measure the accuracy of the built model in predicting a targetted variable. In the field of binary classification, there are a few popular algorithms which are Logistic Regression, Linear Discriminant Analysis KDD, Support Vector Machines and Naïve Bayes. For the purpose of this research, these five algorithms will be utilised to build the predictive models and the model which gives the best performance is selected to determine sarcastic cyberbullying.

3.1.1 Logistic Regression

Logistic regression (LR) is a comprehensive, complete & powerful tool used for supervised classification techniques. It is often considered as an extension upon the normal or ordinary regression model, and can only model dichotomous to variables that usually indicate the occurrence or non-occurrence of events. Logistic regression targets to achieve a probability that a new instance belongs to a certain class. Since it is a probability, that the result is between 0 and 1 (Lee, Wei-Meng, 2019).Therefore, to implement LR as a binary classifier, we assign a threshold to distinguish two classes. For example, if the probability value of an input instance is greater than 0.50, it will be normalized, classified & identified as "Class A"; otherwise it will just be known as "Class B". Upon this, the LR model can also be extended to model categorical variables with more than two values (Berger, Dale, 2017). Therefore, the broad version of LR is called multinomial logistic regression which explains that in order to obtain multiple logit models, it is conceivable to run Y1 independent binary logistic regression models for Y possible outcomes, one of which is selected as the "pivot", and then the other Y1 (El-Habil, Abdalla, 2012). The results are regressions for the pivot results respectively.

$$y = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Figure 3: Logistic Regression Algorithm

In contrast, this would proceed as follows, indicating that the Y outcome, or known as the last outcome is chosen as the pivot. This formulation is known as already transformed and commonly is used in compositional data analysis (Xia, Fan & Chen, Jun & Fung, Wing & Li, Hongzhe, 2013). In our case, since we are dealing with more characteristics to detect sarcastic cyberbullying, we proceed towards 1 step further in LR, which is MLR (Multinomial Logistic Regression). Multinomial Logistic Regression works in a certain way where the null hypothesis, is statistically exclusive / lingo, meaning if there is no relationship with the characteristic of factor A, B, or C. Therefore, Multinomial Logistic Regression helps to evaluate if the hypothesis is true, by evaluating the coefficients for each term in a model. In contrast, the coefficients are used to predict numerical relationship between consumer income and the probability is associated upon interpreting the “p

value". Meaning, the "p value" evaluates the chance of seeing our results, assuming that there is no relationship between factor A between B or C. For example, a "p value" which is less that or equal to 0.05 results a model to be statistically significant (Hamilton, Lawrence & Seyfrit, Carole, 1994). Figure 4 below shows a distributional example between the evaluation of Multinomial Logistic Regression.

Dependent Variable: Any Cyberbullying Data Online (e.g. format A, B, C, etc)

Independent Variable: Sarcasm

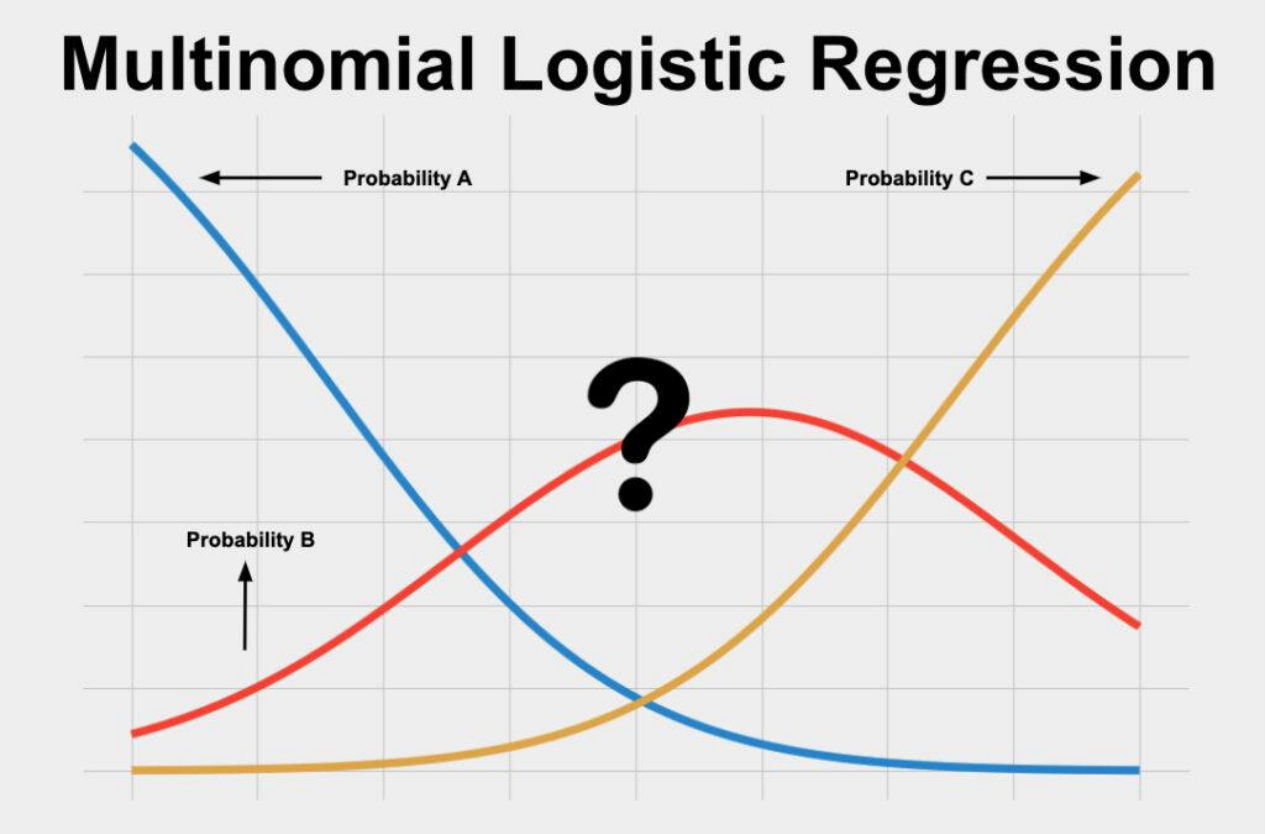


Figure 4: Multinomial Logistic Regression

3.1.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), Normal Discriminant Analysis (NDA) or Discriminant Function Analysis are generalized names of the “Fisher's linear discriminant” method, which is a method used in statistics and other fields to find the characterization or separate two or more categories from a linear combination of multiple types of characteristics within an object or an event. The resulting combination can be used as a linear classifier or, more commonly, dimensionality reduction before further classification (Diaf, A. & Boufama, Boubakeur & Benlamri, Rachid, 2013).

LDA is closely related towards variance analysis (ANOVA) and regression analysis, which also attempts to express the dependent variable as a linear combination of other characteristics or measurements. However, (ANOVA) works by utilizing categorical and independent variables that are continuously dependent attributes, while discriminant analysis has continuously dependent variables and categorical dependent attributes (ie, class labels). Logistic regression and normal regression are often similar to LDA as compared than ANOVA, because they also illustrate a categorical variable by its value of a continuous independent variable. In applications such as data from twitter, where the assumption are based upon independent variables or normally distributed factors, LDA is not unreasonable, as commonly, this method covers most of the technical applications when applied as an algorithm (Singh, Anagh & Prakash, B. & Chandrasekaran, K, 2016).

LDA is also closely correlated to the principal component analysis (PCA) and factor analysis because they are both looking towards understanding linearity within combinations of attributes that can better explain the data. (LDA) explicitly attempts to model the differences between data classes. On the other hand, (PCA) does not consider any category differences, and factor analysis constructs featuring combinations based on differences rather than similarities. The difference between discriminant analysis and factor analysis is that it is not an interdependent technique; as independent variables and dependent variables must be distinguished when measuring the quality of a variable with continuous quantity (Jolliffe, Ian & Cadima, Jorge, 2016).

LDA also works as the equivalent technique, known as “discriminant correspondence analysis”. Discriminant analysis is used for a priori known groups (as opposed to cluster analysis). In this, each use case should have a score on one or more quantitative predictors and a score on a

group of measurements. In simple terms, discriminant function analysis is the act of assigning attributes to groups, classes, or categories of the same type (Abdi, Hervé, 2007).

When comparing LDA with conventional logistic regression, logistic regression is proposed for binary classification problems, and is further extended to multiple class classifications, but is rarely used for this purpose. Therefore, when the class separation is good, logistic regression is restricted and unstable. Logistic regression also seems to be an unstable method when it comes to some examples of estimated parameters. However, linear discriminant analysis can address all of the above and be used as a linear method for multiple class classification problems (Perme, Maja & Blas, Mateja & Turk, Sandra, 2004). The way the algorithm works is that every variable, dimension, or attribute in the dataset has a Gaussian distribution, that is, the characteristic has a bell-shaped curve. Also, each characteristic has the same variance and the same amount of variation around the average. This suggests that each characteristic is assumed to be randomly sampled. Therefore, the independent variables / characteristics lack multicollinearity. If the correlation between independent characteristics increases, the predictive power will decrease.

$$S_b = \sum_{i=1}^{g} N_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T$$

$$P_{lda} = \arg\max_{P} \frac{\left| P^T S_b P \right|}{\left| P^T S_w P \right|}$$

Figure 5: Linear Discriminant Analysis Algorithm

The above LDA algorithm works as follows: it calculates the separability among different classes, which is the distance between the means of different classes, also known as the variance between classes. Find the mean value of each class and the distance between the samples, also known as the within-class variance. Create a low-dimensional space that maximizes variation between classes and minimizes variation within the class. Assume P as a lower dimensional spatial projection, that is, Fisher's criterion. The following is an example of what it looks like when LDA is

applied.

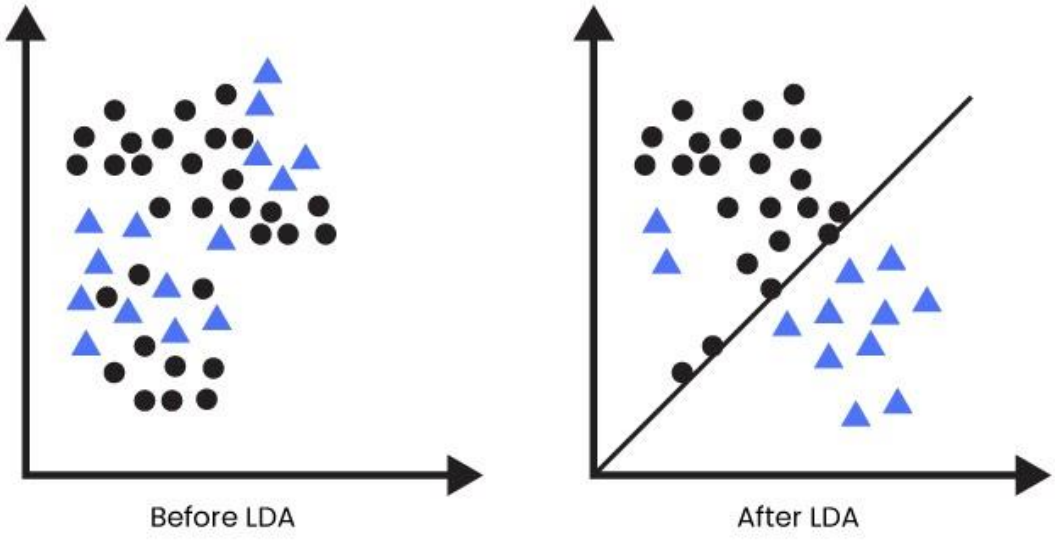


Figure 6: Classification Example of various factors before and after implementing LDA

In conclusion, LDA classes have multiple features, having the ability to effectively and efficiently classify and yield over some kind of overlapping factors. Therefore, when there are increasing features or factors overlapping the result of the classification problem, the result is able to dissect its features are properly classify terms (Balakrishnama, S. & Ganapathiraju, Aravind, 1998). For example, classifying aggressive sarcastic terms between racism sarcastic terms through social media text.

3.1.3 KNN

In statistics, the k-nearest neighbors rule (k-NN) is a non-parametric classification methodology utilized for classification and regression purposes. In regular cases, the inputted data consists of the trained “k nearest” examples from an information set. The result is highly regarded depending if KNN is being implemented for the standard classification or regression model.

In k-NN classification, its nearest neighbours where (k is a positive integer, usually the dataset is small). If “k = 1”, then the item is solely allotted to the category of that single nearest neighbour. In KNN regression models, the resulting output is known as the property worth for an object or attribute. This is valued as an average based of upon the k nearest number of neighbours. KNN may be a form of classification wherever the operation is simply approximated regionally and every one computation is delayed till a function is evaluation (Lodwich, Aleksander & Shafait, Faisal & Breuel, Thomas, 2011). Since this rule depends on distance for classification, if the options represent totally different physical units or are available immensely different scales then normalizing the trained information will improve its accuracy dramatically.

Similar to nearly other algorithms, KNN works based upon the result of deeply stacked mathematical theories when it is used. Once implementing KNN, the primary step is to remodel informational attributes by pointing them into featured Euclidean vectors, or any other mathematical value. The rule then works by realizing the gap between the mathematical values of those points. The foremost common algorithm to find distance utilizing the Euclidian distance, is as shown below.

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Figure 7: KNN Algorithm

KNN utilizes the Euclidean distance mathematical formula to compute the distance between each attribute point and tested attribute. It then works by finding the probability of these endpoints, by understanding which of the tested data are similar to one or the other. KNN is then classified and attributed upon its highest probabilities. To visualize this formula, it would look something like this below at Figure 8;

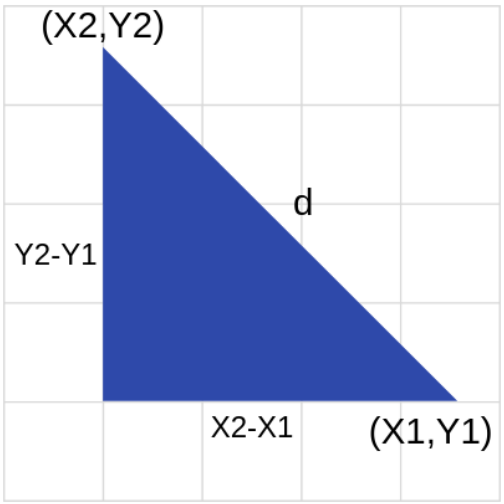


Figure 8: Euclidean Distance Diagram

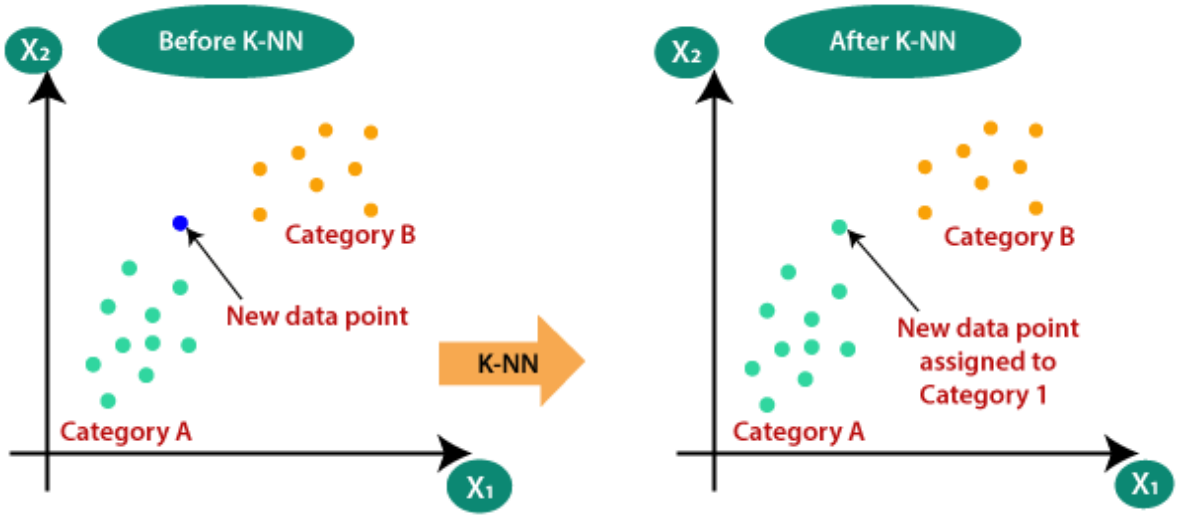


Figure 9: Implementation Prior & After KNN

Based on Figure 9, KNN works by selecting the number of K neighbours. Once this is done, the Euclidean formula is applied to calculate the distance between Y number of neighbours. Once this is done, new data points are assignment on top of the 2 corresponding categories. If there are more than 1 categories, usually, KNN works to count and create the most optimum number of maximum neighbours to apply its use case. A lot of times, KNN is utilized for segmentation purposes, such as segmenting a few data factors on 1 category (Surlakar, Prachi & Araujo, Sufola & Sundaram, K, 2016).

3.1.4 Naïve Bayes

The Naïve Bayes algorithm works as classifiers which are considered to be a family of easy "probability classification" supported by applying the applications from Bayes' theorem to study their (naïve) independence assumptions between the options (see Bayes classifier). (Naïve Bayes) classifiers are highly scalable. However, they require a variety of parameters, lineated within a number of variables (features/predictors) during a learning problem. In machine learning, the maximum-likelihood learning or training of a model is done by evaluating a closed-form expression, that takes a linear time, instead of by expensive unvaried approximation as used for several alternative styles of classifiers (Friedman, Nir & Geiger, Dan & Goldszmidt, Moises, 1997). In contrast, this is a probabilistic machine learning model which is built fundamentally from the Bayes theorem which assumes that the predictors are independent of each other. Bayes theorem is given by:

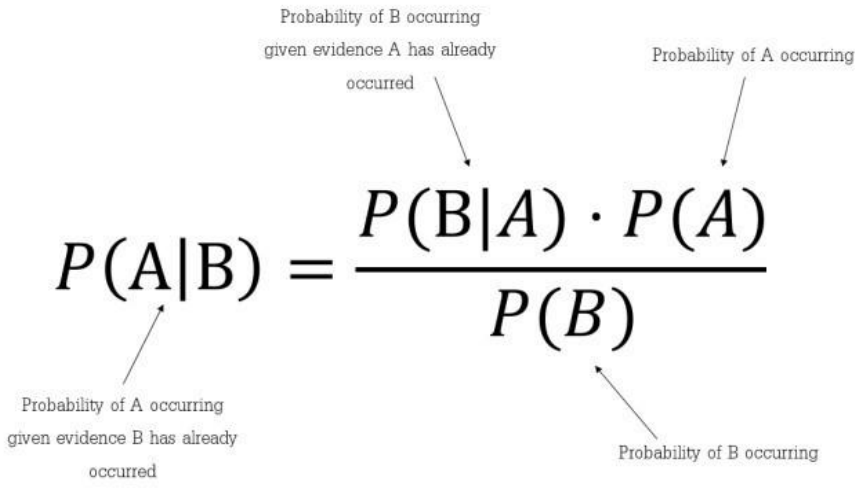


Figure 10: Naïve Bayes Probability Algorithm

Figure 10 is evaluated based on the following, where "$P(B|A)$ is the probability of predictor given by class, $P(A)$" and the probability of "class where $P(B)$ is the probability of the predictor". However, the drawbacks of using the above formulation is that when the number of N features are too large , it causes the probability of a model to become infeasible. Therefore, if this happens, the model must be reformulated. Using the Bayes' theorem, the conditional probability can also be decomposed as the above's written algo (Rish, Irina, 2001). In practice, one is only interested in the numerator of a fraction, because the denominator does not depend on C and the values of the features are given, so that the denominator is effectively constant. Meaning, the numerator is equivalent to the joint probability of a model.

3.1.5 SVM

In machine learning, the application of support-vector machines is that it is a supervised learning model associated with learning algorithms that can analyse categorical data for regular classification and normal regression analysis. SVM works in a way when it is set or even labelled with a subset of a training attribute examples, it has the ability to mark each belonging attribute to one of two or various categories. This algorithm is based upon searching for a hyperplane in a N-dimensional space which can be used to classify attributes into distinct categories. The hyperplane has the maximum distance which separates the data points of different classes (Evgeniou, Theodoros & Pontil, Massimiliano, 2001).

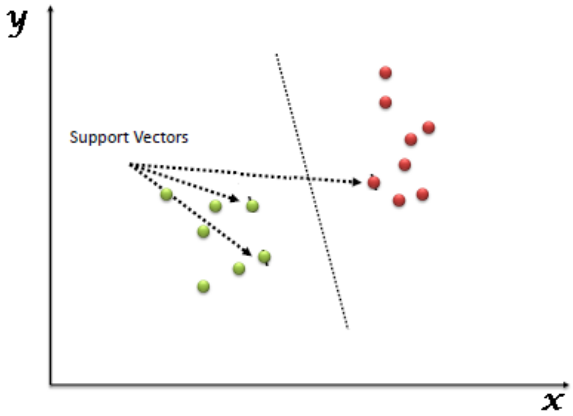


Figure 11: SVM Algorithm implementation

The SVM training rule builds models that will assign new vectors to at least one class or the other, creating a (non-probabilistic) “binary linear classifier”. SVM maps training examples with the purpose in areas to maximise the gap between the 2 classes; ex supported Figure 11. Once trained, new vectors are then created upon the mapped attributes into the same area vectorized and are foreseen to belong to a category based on that facet of the gap they fall (Borg, Ingwer & Mair, Patrick & Cetron, Joshua, 2020). Facet theory and multidimensional scaling. Additionally, in utilizing linear classification, SVMs are economical in performing non-linear classification exploitation, which is what's referred to as a kenel trick, by utilizing an implicit mapping to know high dimensional feature spaces (Kurita, T, 2004). However, a lot of times, data is unlabelled. Therefore, at that point, supervised learning isn’t possible yet. Therefore, a unsupervised learning approach can be utilized, to discovered by building a natural cluster that has the ability to map out information into groups. This is sometimes known as segmentation in data analysis and it is one amongst the foremost wide used machine learning algorithms in industrial applications.

4 Chapter 4: Methodology: O.S.E.M.N

4.1 O.S.E.M.N Framework

This research paper will focus on expanding our work to discover additional features that are significant in identifying sarcastic cyberbullying. Although, the Cross Industry Standard Process for Data Mining (CRISP-DM) Model is a widely adopted paper that provides a clear set of guidelines that are used to structure and plan out data analysis projects (Bosnjak, Zita & Grljevic, Olivera & Bošnjak, Saša, 2009). We have decided to utilize a different framework called OSEMN as this framework is more for a process for independent research. The OSEMN process for data science stands for Obtain, Scrub, Explore, Model & Interpret, which describes the analytics workflow used in this approach. The framework for the following methodology was originated from Hilary Mason and Chris Wiggins in 2010. According to Dineva, Kristina & Atanasova, Tatiana. (2018), they described it as the follow-up of standardized processes in data science. According to the OSEMN framework, the life cycle of a data science project contains 5 phases; which is to obtain, scrub, explore, model data & interpret results.

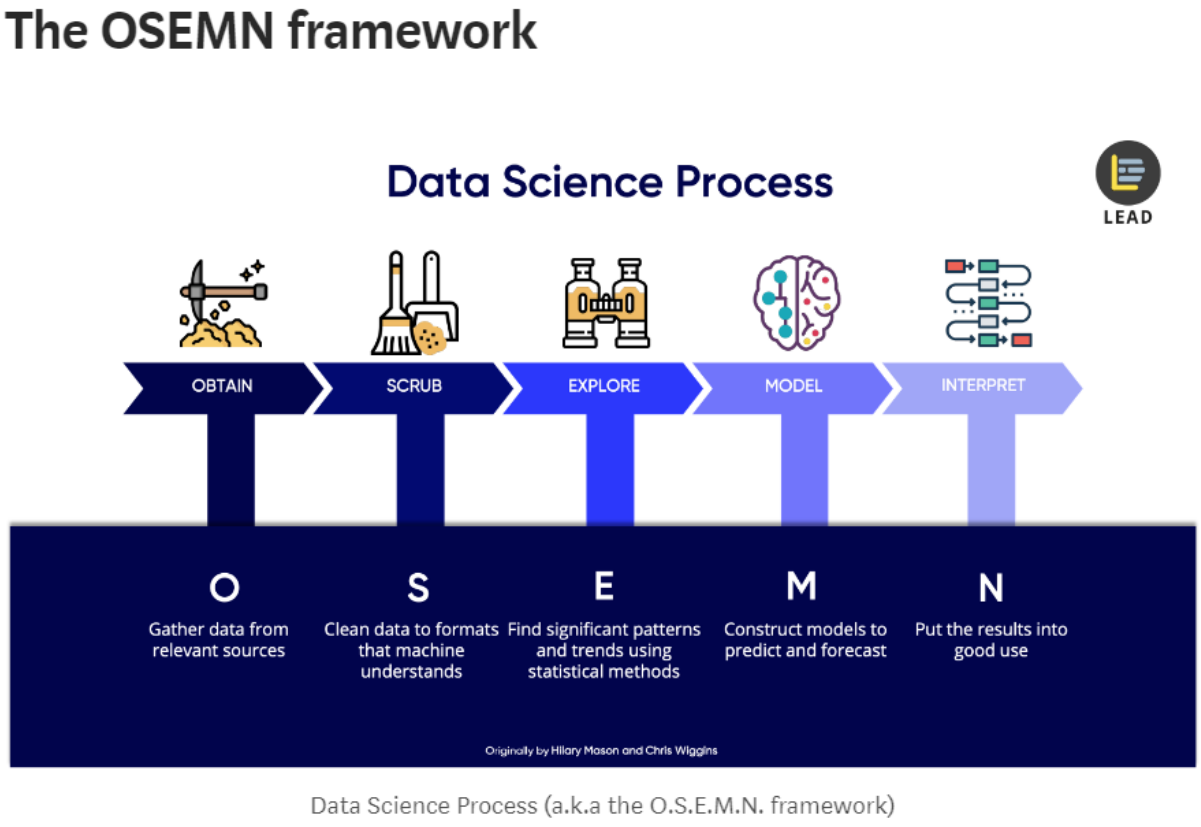
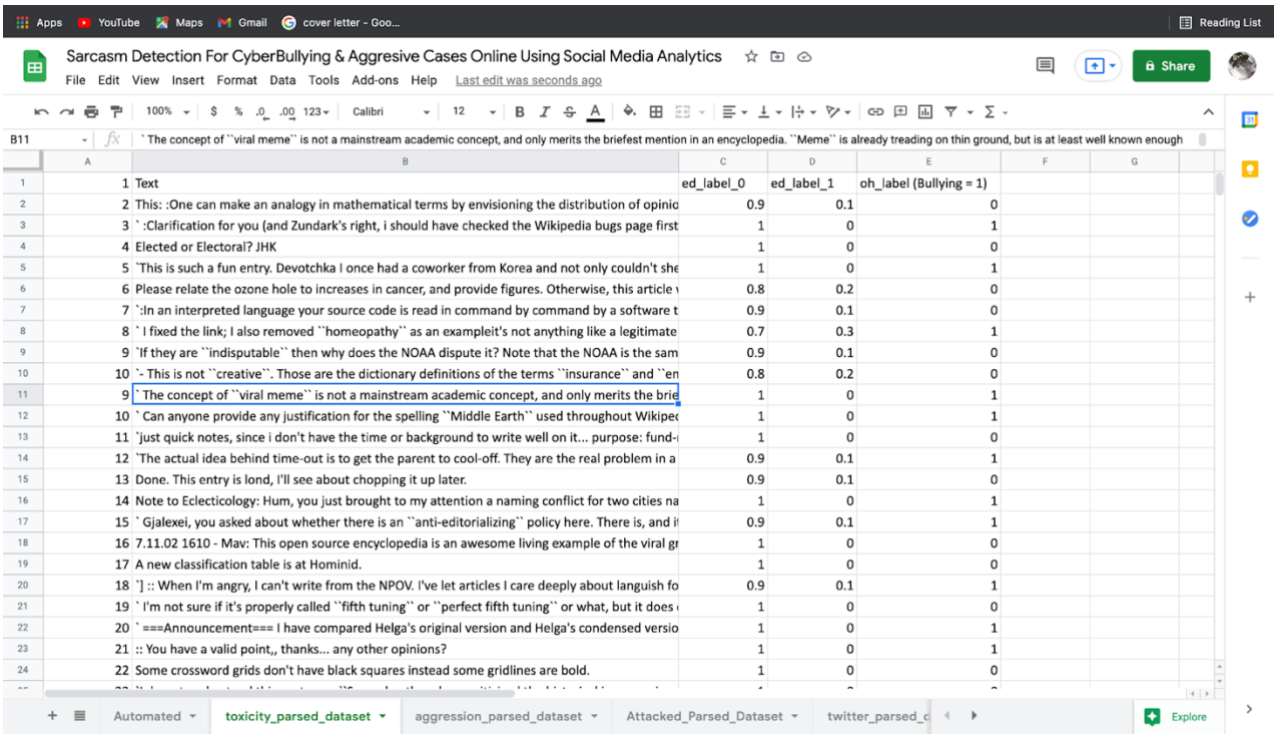


Figure 12: The OSEMN Framework

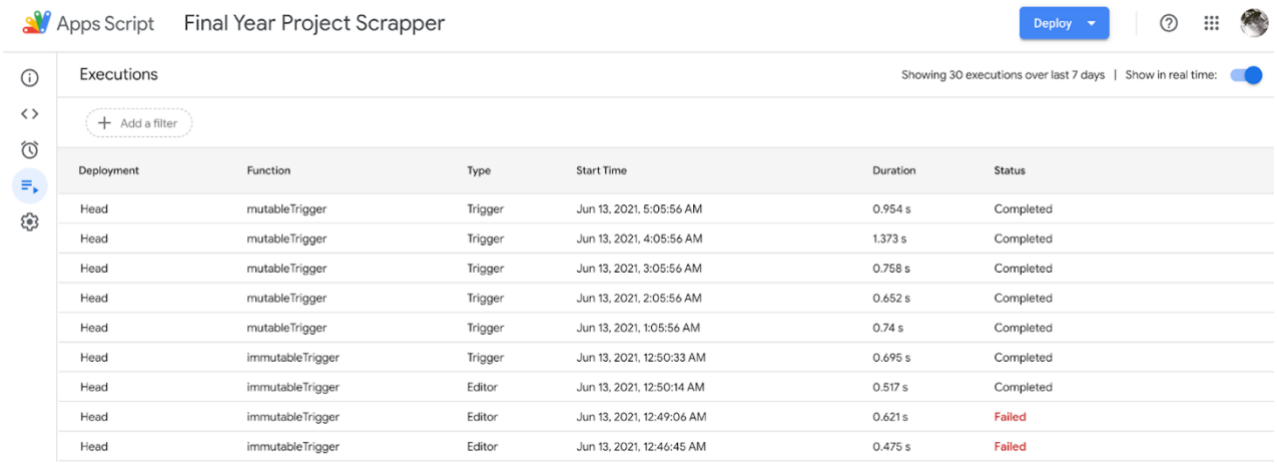
4.1.1 Obtain Data

The data that has been identified for this research and study is from twitter, obtained through data scraping using the Twitter API, scheduled on Google Apps Script to automate data collection of sarcastic cyberbullying online from twitter, and all the open source datasets, except youtube_parsed_dataset.csv & kaggle_parsed_dataset.csv from Mendeley Data. The open sourced datasets consist of cyberbullying data which is used to support our research in sarcastic cyberbullying detection.



	A	B	C	D	E
1	1	Text	ed_label_0	ed_label_1	oh_label (Bullying = 1)
2	2	This: :One can make an analogy in mathematical terms by envisioning the distribution of opinio	0.9	0.1	0
3	3	` :Clarification for you (and Zundark's right, i should have checked the Wikipedia bugs page first	1	0	1
4	4	Elected or Electoral? JHK	1	0	0
5	5	`This is such a fun entry. Devotchka I once had a coworker from Korea and not only couldn't she	1	0	1
6	6	Please relate the ozone hole to increases in cancer, and provide figures. Otherwise, this article	0.8	0.2	0
7	7	`:In an interpreted language your source code is read in command by command by a software t	0.9	0.1	0
8	8	` I fixed the link; I also removed ``homeopathy`` as an exampleit's not anything like a legitimate	0.7	0.3	1
9	9	`If they are ``indisputable`` then why does the NOAA dispute it? Note that the NOAA is the sam	0.9	0.1	0
10	10	`- This is not ``creative``. Those are the dictionary definitions of the terms ``insurance`` and ``en	0.8	0.2	0
11	9	` The concept of ``viral meme`` is not a mainstream academic concept, and only merits the brie	1	0	1
12	10	` Can anyone provide any justification for the spelling ``Middle Earth`` used throughout Wikipe	1	0	1
13	11	`just quick notes, since i don't have the time or background to write well on it... purpose: fund-	1	0	0
14	12	`The actual idea behind time-out is to get the parent to cool-off. They are the real problem in a	0.9	0.1	1
15	13	Done. This entry is lond, I'll see about chopping it up later.	0.9	0.1	0
16	14	Note to Eclecticology: Hum, you just brought to my attention a naming conflict for two cities na	1	0	1
17	15	` Gjalexei, you asked about whether there is an ``anti-editorializing`` policy here. There is, and i	0.9	0.1	1
18	16	7.11.02 1610 - Mav: This open source encyclopedia is an awesome living example of the viral gr	1	0	0
19	17	A new classification table is at Hominid.	1	0	0
20	18	`] :: When I'm angry, I can't write from the NPOV. I've let articles I care deeply about languish fo	0.9	0.1	1
21	19	` I'm not sure if it's properly called ``fifth tuning`` or ``perfect fifth tuning`` or what, but it does	1	0	0
22	20	` ===Announcement=== I have compared Helga's original version and Helga's condensed versio	1	0	1
23	21	:: You have a valid point,, thanks... any other opinions?	1	0	1
24	22	Some crossword grids don't have black squares instead some gridlines are bold.	1	0	0

Figure 13: Data Description Of Open Sourced Datasets.



Deployment	Function	Type	Start Time	Duration	Status
Head	mutableTrigger	Trigger	Jun 13, 2021, 5:05:56 AM	0.954 s	Completed
Head	mutableTrigger	Trigger	Jun 13, 2021, 4:05:56 AM	1.373 s	Completed
Head	mutableTrigger	Trigger	Jun 13, 2021, 3:05:56 AM	0.758 s	Completed
Head	mutableTrigger	Trigger	Jun 13, 2021, 2:05:56 AM	0.652 s	Completed
Head	mutableTrigger	Trigger	Jun 13, 2021, 1:05:56 AM	0.74 s	Completed
Head	immutableTrigger	Trigger	Jun 13, 2021, 12:50:33 AM	0.695 s	Completed
Head	immutableTrigger	Editor	Jun 13, 2021, 12:50:14 AM	0.517 s	Completed
Head	immutableTrigger	Editor	Jun 13, 2021, 12:49:06 AM	0.621 s	Failed
Head	immutableTrigger	Editor	Jun 13, 2021, 12:46:45 AM	0.475 s	Failed

Figure 14: Automated Twitter API crawling triggers using Google Apps Script

The description of each attribute from their respective datasets provide the understanding of how each variable was derived as below:

Attribute	Description
No	Unique identifier of a tweet
Text	Bullying Tweet Text
ed_label_0	0 = Non - Bullying
ed_label_1	1 = Bullying
oh_label (Bullying = 1)	Labelled Classification For Bullying

Table 2: Description of toxicity_parsed_dataset

Attribute	Description
No	Unique identifier of a tweet
Text	Bullying Tweet Text
ed_label_0	0 = Non - Bullying
ed_label_1	1 = Bullying
oh_label (Bullying = 1)	Labelled Classification For Bullying

Table 3: Description of aggression_parsed_dataset

Attribute	Description
No	Unique identifier of a tweet
Text	Bullying Tweet Text
ed_label_0	0 = Non - Bullying
ed_label_1	1 = Bullying
oh_label (Bullying = 1)	Labelled Classification For Bullying

Table 4: Description of attacked_parsed_dataset

Attribute	Description
No	Unique identifier of a tweet
Text	Bullying Tweet Text
ed_label_0	0 = Non - Bullying
ed_label_1	1 = Bullying
oh_label (Bullying = 1)	Labelled Classification For Bullying

Table 5: Description of twitter_parsed_dataset

Attribute	Description
No	Unique identifier of a tweet
Text	Bullying Tweet Text
ed_label_0	0 = Non - Bullying
ed_label_1	1 = Bullying
oh_label (Bullying = 1)	Labelled Classification For Bullying

Table 6: Description of twitter_sexist_dataset

Attribute	Description
No	Unique identifier of a tweet
Text	Bullying Tweet Text
ed_label_0	0 = Non - Bullying
ed_label_1	1 = Bullying
oh_label (Bullying = 1)	Labelled Classification For Bullying

Table 7: Description of twitter_racism_parsed_dataset

4.1.2 Scrub Data

After obtaining data, we refer towards scrubbing or referenced as cleaning the data. Firstly, we look at the term “garbage in, garbage out” philosophy in data quality. At this phase, we look towards dropping all irrelevant data such as stop word removal, emoji filtering & data deletion. We clean data for obtaining sarcastic cyberbullying detection using a number of text pre-processing techniques. Before extracting the linguistic and psycholinguistic features, there were a number of text pre-processing methods that were used to prepare the data. The data cleaning is done using Python and libraries within it such as nltk and contractions.

The first technique utilized is through sentence segmentation which is known as a process to convert paragraphs into sentences. This is done through the nltk python library in google colab. Once the data is available in sentences, we proceed towards the process of tokenization, which is to break down sentences into regular words. Secondly, the number of sentences in the twitter text and open sourced cyberbullying datasets were counted based on the total number of periods, exclamation marks and question marks which indicate the end of sentences. Next, contractions were expanded to get each individual word. For example, “they’ve” will be expanded to “they have”. Upon this, all punctuations were then removed from the text and the texts are split into strings each containing individual words. Lastly, a lemmatization function is applied to convert each individual word back into their root word. As an example, the words “makes no sense” and “no sense” both comes from the root word “stupid” The individual words are then mapped to their respective linguistic and psycholinguistic features explained earlier and aggregated accordingly to transform the data. It can be seen that the dataset contains 63.40% of sarcastic_bullying_cases and 36.40% are non_sarcastic_bullying cases. The dataset is later concatenated & merged to build a final dataset.

Attribute	Description
Index	Unique Category Of Each Tweet Factor
Text	Bullying Tweet Text
Id	Unique Identifier For All Tweets
Annotation	The Bullying Characteristic Types
oh_label (Bullying = 1)	Labelled Classification For Bullying

Table 8: Description of final_dataset_for_training

4.1.3 Explore Data

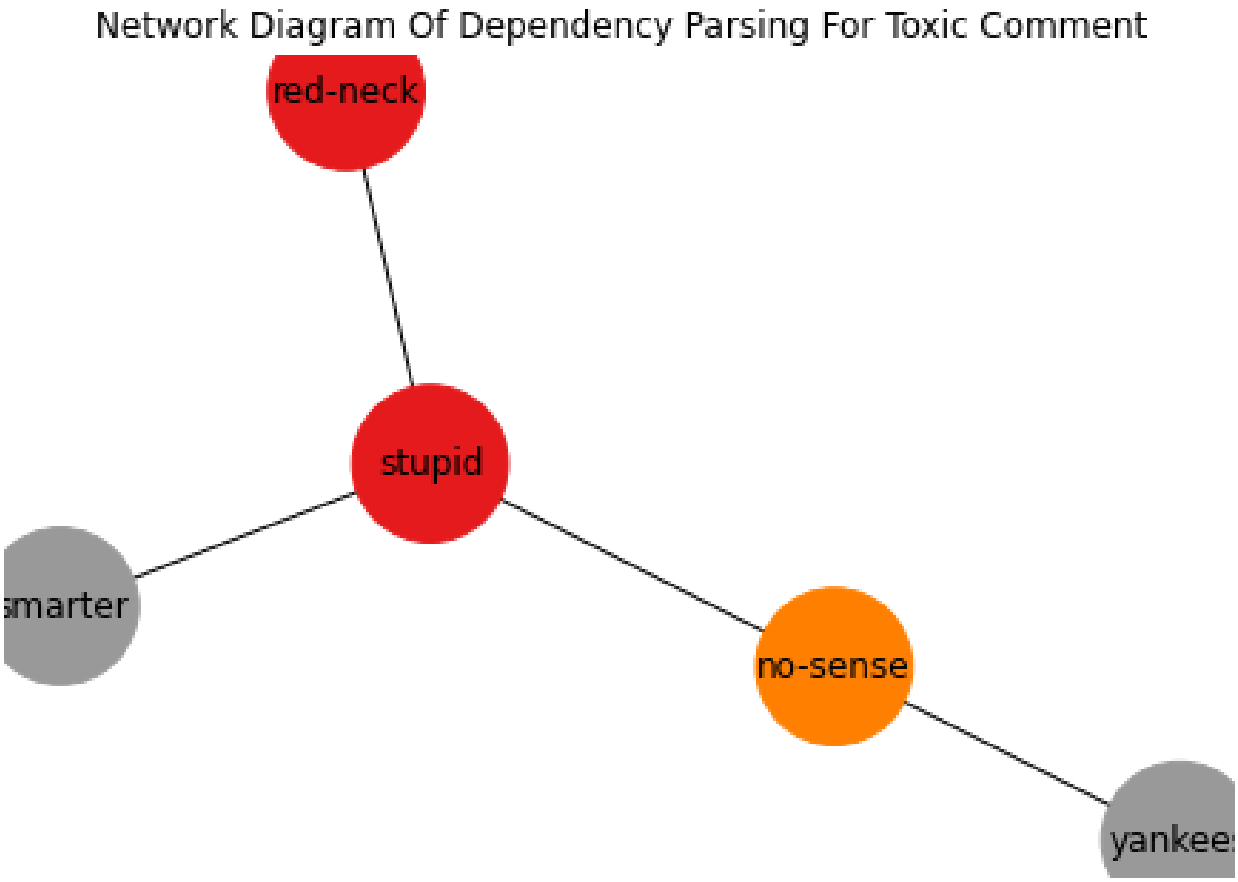


Figure 15: Network Diagram Of Dependency Parsing For Toxic Comment

After Lemmatization, we proceed towards performing dependency parsing which is to find relationships between the words. Dependency Parsing (DP) works by analysing the grammatical structure of words from sentences. Based on their comments, the words are matched to their respective components to their following relationships in a network diagram. The mechanism of this approach is based on the concept that there is a direct link between every linguistic unit of a sentence. This allows us to detect cyberbullying derived from sarcasm. Based on Figure 15, it helps us indicate the different word correlation sentiment as when the terms are connected, they have a relationship with one and the other. This means, that sarcastic terms are different as they are ambiguous. This also indicates that each comment may be a positive or negative context when deriving sarcasm. For example, Figure 15 says that the toxic comment made by the tweeter shows that he believes that the opposing comment is a “red neck” and what the opposer says, makes no sense, believing he is smarter. This indicates a negative sentiment on a sarcastic comment or tweet.

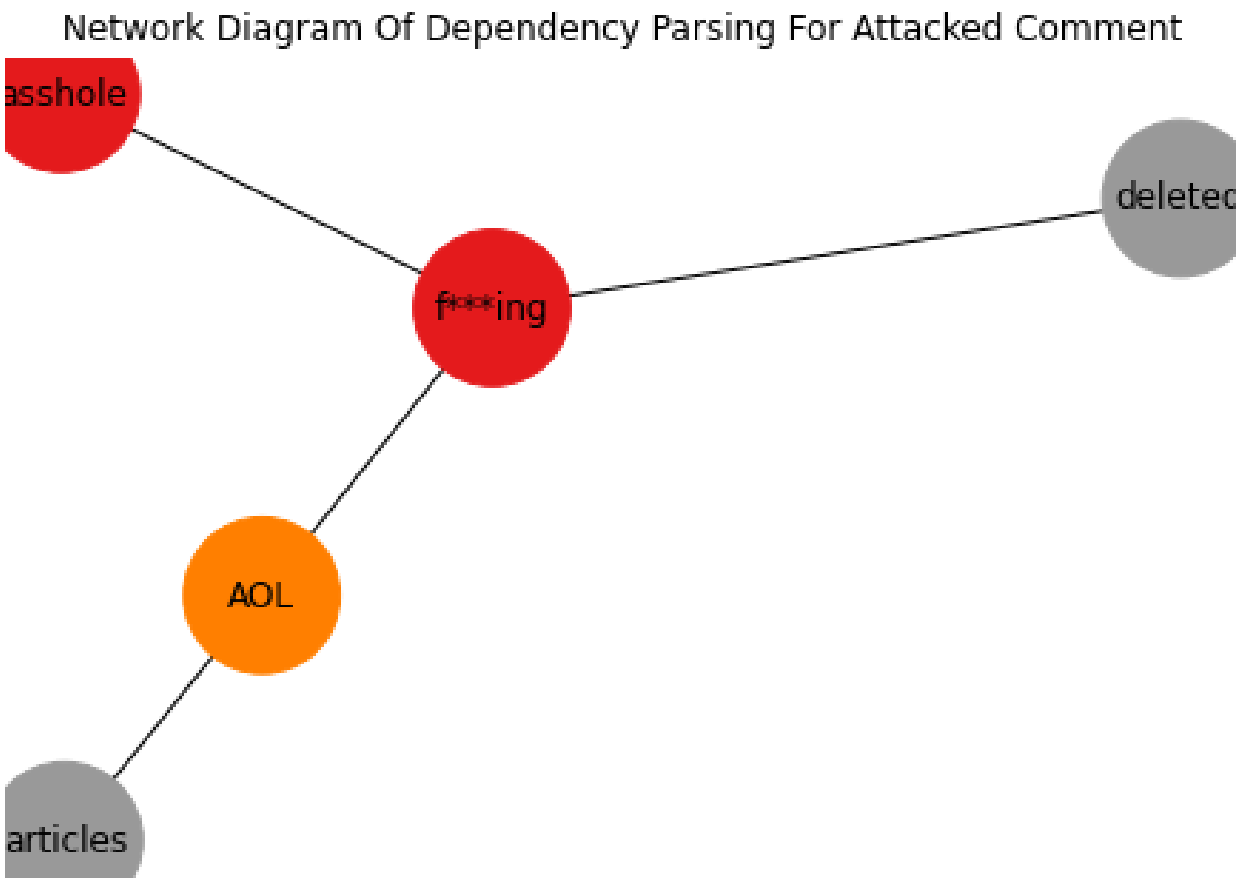


Figure 16: Network Diagram Of Dependency Parsing For Attacked Comment

After Lemmatization, we proceed towards performing dependency parsing which is to find relationships between the words. Dependency Parsing (DP) works by analysing the grammatical structure of words from sentences. Based on their comments, the words are matched to their respective components to their following relationships in a network diagram. The mechanism of this approach is based on the concept that there is a direct link between every linguistic unit of a sentence. This allows us to detect cyberbullying derived from sarcasm. Based on Figure 16, it helps us indicate the different word correlation sentiment as when the terms are connected, they have a relationship with one and the other. This means, that sarcastic terms are different as they are ambiguous. This also indicates that each comment may be a positive or negative context when deriving sarcasm. For example, Figure 16 says that the attacked comment made by the tweeter shows that he believes that the opposing comment is an “asshole” and what the opposer says, is that AOL category of articles shared by the opposer should be deleted. This indicates a negative sentiment on a sarcastic comment or tweet.

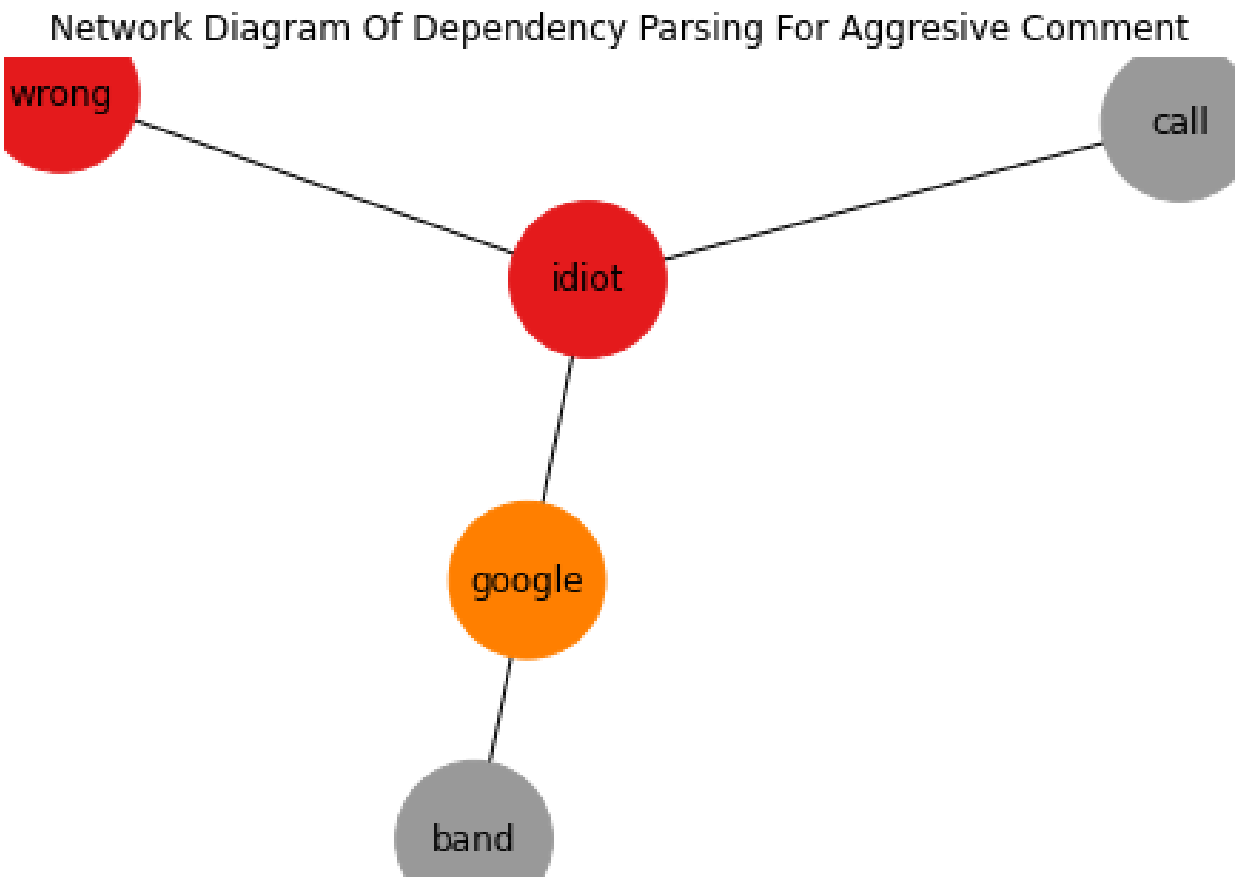


Figure 17: Network Diagram Of Dependency Parsing For Aggressive Comment

After Lemmatization, we proceed towards performing dependency parsing which is to find relationships between the words. Dependency Parsing (DP) works by analysing the grammatical structure of words from sentences. Based on their comments, the words are matched to their respective components to their following relationships in a network diagram. The mechanism of this approach is based on the concept that there is a direct link between every linguistic unit of a sentence. This allows us to detect cyberbullying derived from sarcasm. Based on Figure 17, it helps us indicate the different word correlation sentiment as when the terms are connected, they have a relationship with one and the other. This means, that sarcastic terms are different as they are ambiguous. This also indicates that each comment may be a positive or negative context when deriving sarcasm. For example, Figure 17 says that the aggressive comment made by the tweeter shows that he believes that the opposing comment made by the opposer is wrong and he/she is an idiot, indicating that what the opposer should do is to google what he is trying to find out from “band” and call them. This indicates a negative sentiment on a sarcastic comment or tweet.

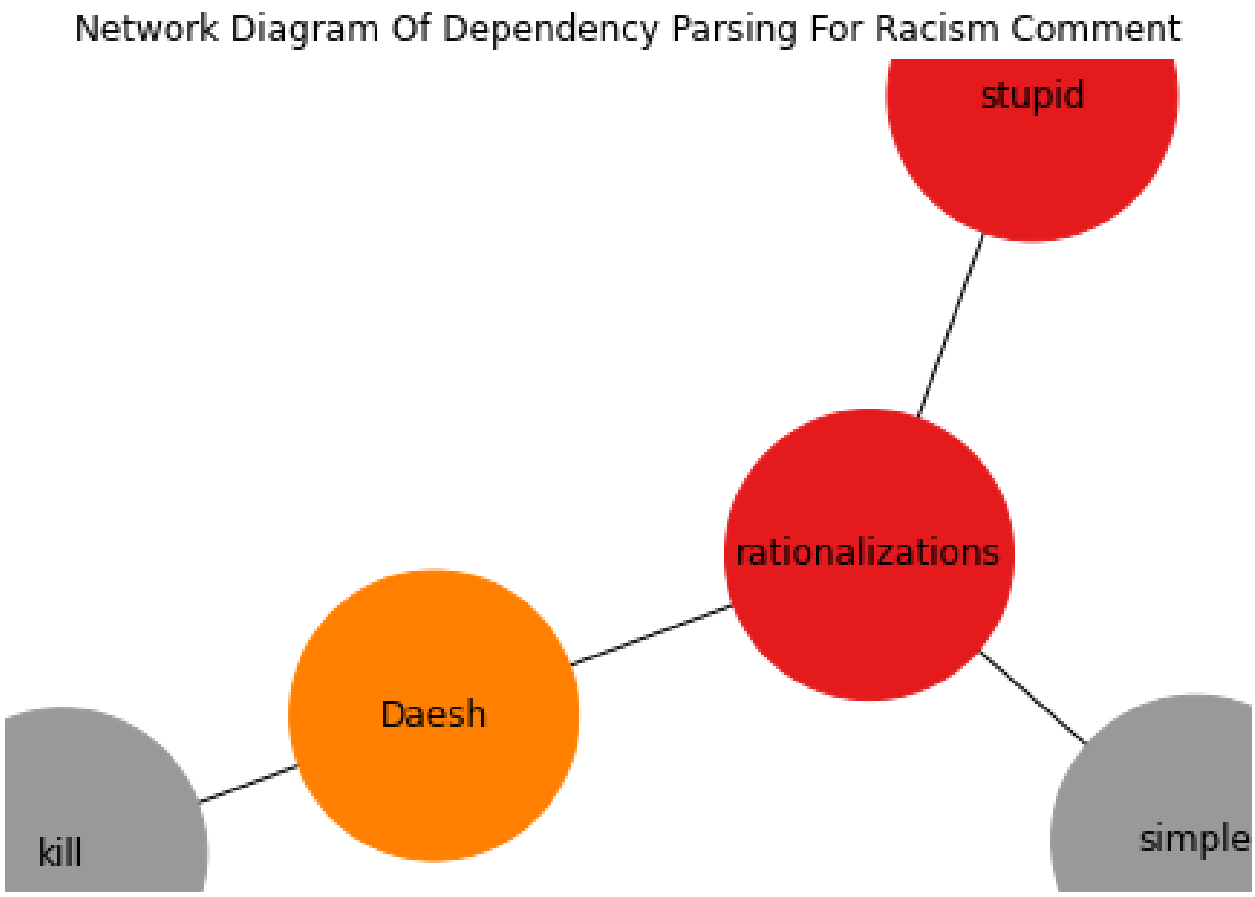


Figure 18: Network Diagram Of Dependency Parsing For Racism Comment

After Lemmatization, we proceed towards performing dependency parsing which is to find relationships between the words. Dependency Parsing (DP) works by analysing the grammatical structure of words from sentences. Based on their comments, the words are matched to their respective components to their following relationships in a network diagram. The mechanism of this approach is based on the concept that there is a direct link between every linguistic unit of a sentence. This allows us to detect cyberbullying derived from sarcasm. Based on Figure 18, it helps us indicate the different word correlation sentiment as when the terms are connected, they have a relationship with one and the other. This means, that sarcastic terms are different as they are ambiguous. This also indicates that each comment may be a positive or negative context when deriving sarcasm. For example, Figure 18 says that the racist comment made by the tweeter shows that he believes that the opposing comment is "stupid" and the tweeter has no simple rationalization to what he/she is tweeting. This indicates a negative sentiment on a sarcastic comment or tweet.

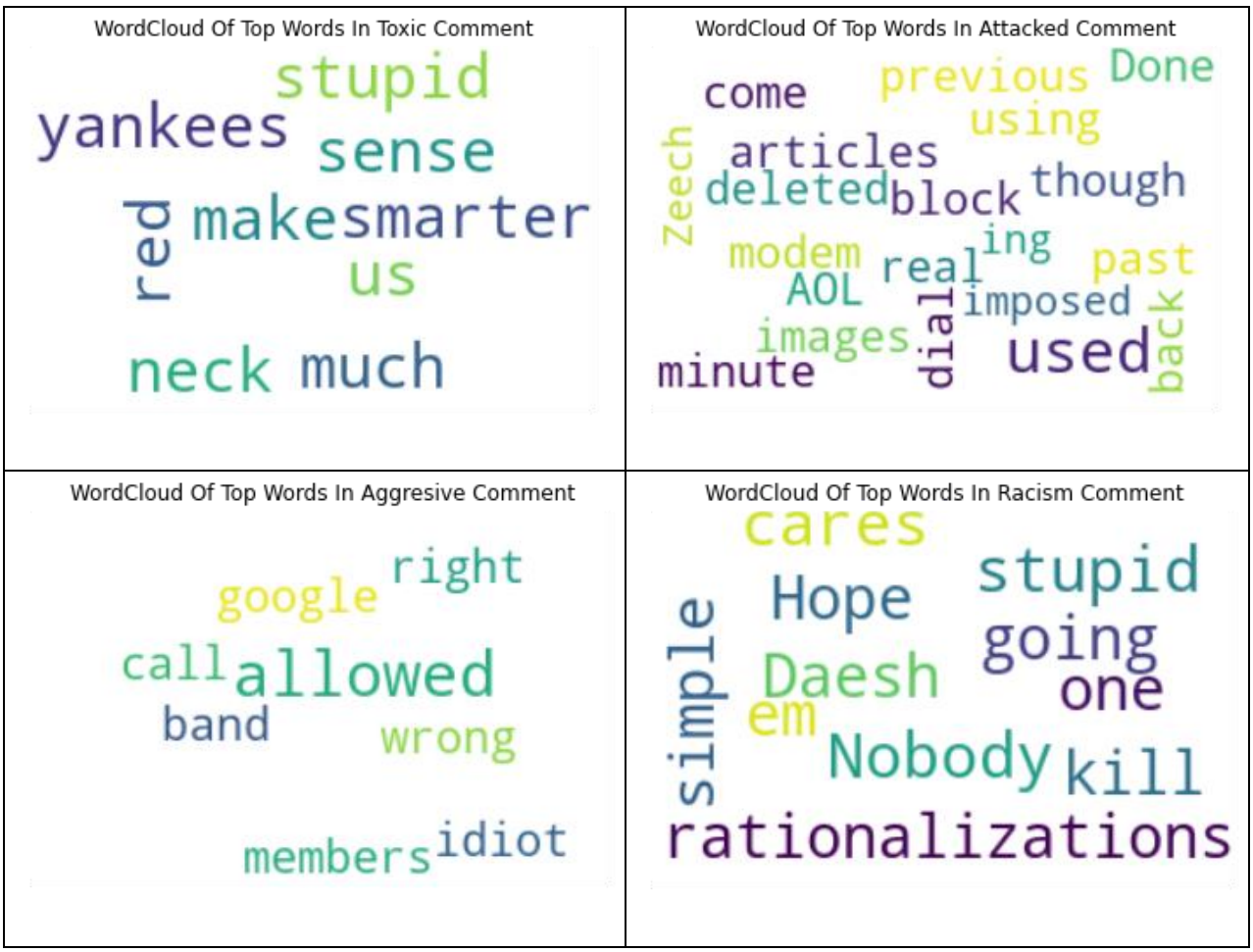


Figure 19: Word Clouds Of Top Words In Toxic, Attacked, Aggressive & Racist Comments

Through our studies from social network analysis, we are able to identify those relationships among various factor characteristics, which are used to derive sarcastic cyberbullying in social media. Based on Figure 19, the word clouds indicate the ideas which are in focused. The topics which are of focus show that a large polarity of the sarcastic comments are negative, indicating that cyberbullying is not context dependent, but sarcasm is being used by tweeters in social media to conduct hate speech and bully others in social media.

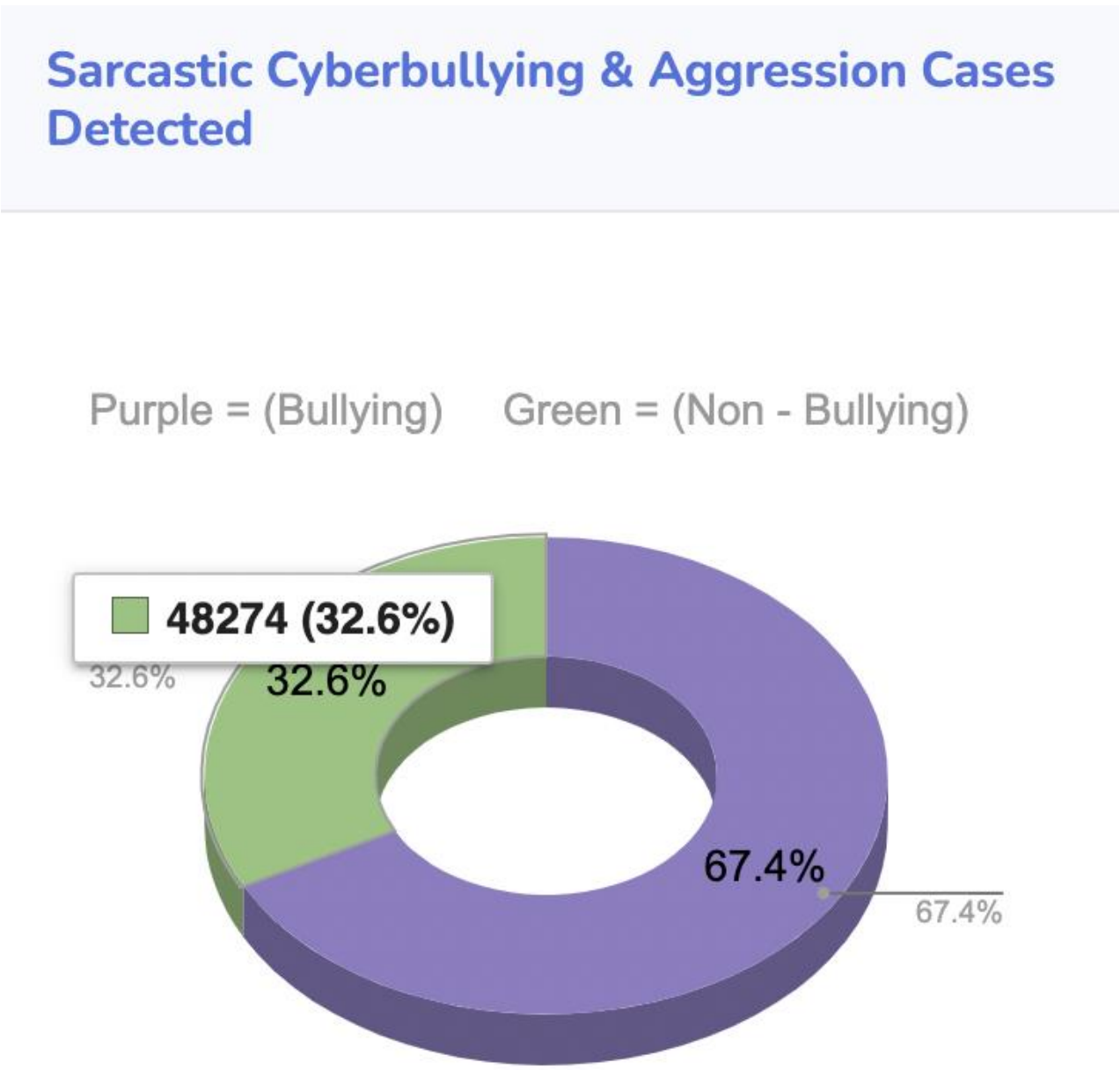


Figure 20: Sarcastic Cyberbullying & Aggression Cases Detected

Based on Figure 20, we have identified a large portion of Bullying cases available online in social media. The statistics show that 67.4% (99582) of the bullying cases detected are derived based on the factors analysed from the social network analysis diagrams and a quarter portion detected in social media is able to detect non – bullying cases which is 32.6 % (48274). This gives us an indication that online bullying should not be taken too lightly as various factors are derived upon sarcasm to interject hate speech with an opposing tweeter.

Figure 21: Detected Bullying Cases Over Period (1st June 2021)

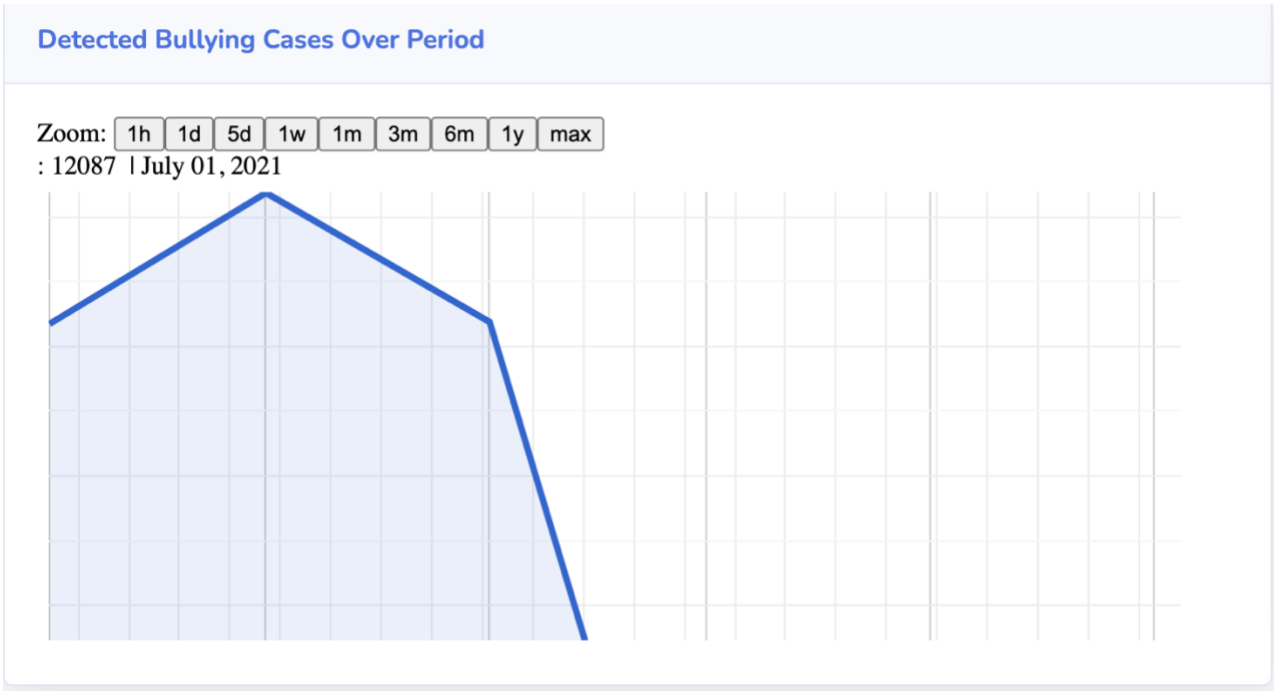


Figure 22: Detected Bullying Cases Over Period (1st July 2021)

On top of the detected bullying cases, time series analysis is applied on top of the data to identify the bullying trend increase from early April 2021 (74321) cases to present (12087) cases. Through the detection of our automated tool, we are able to detect a distinct increase in bullying cases over a timestamp period.

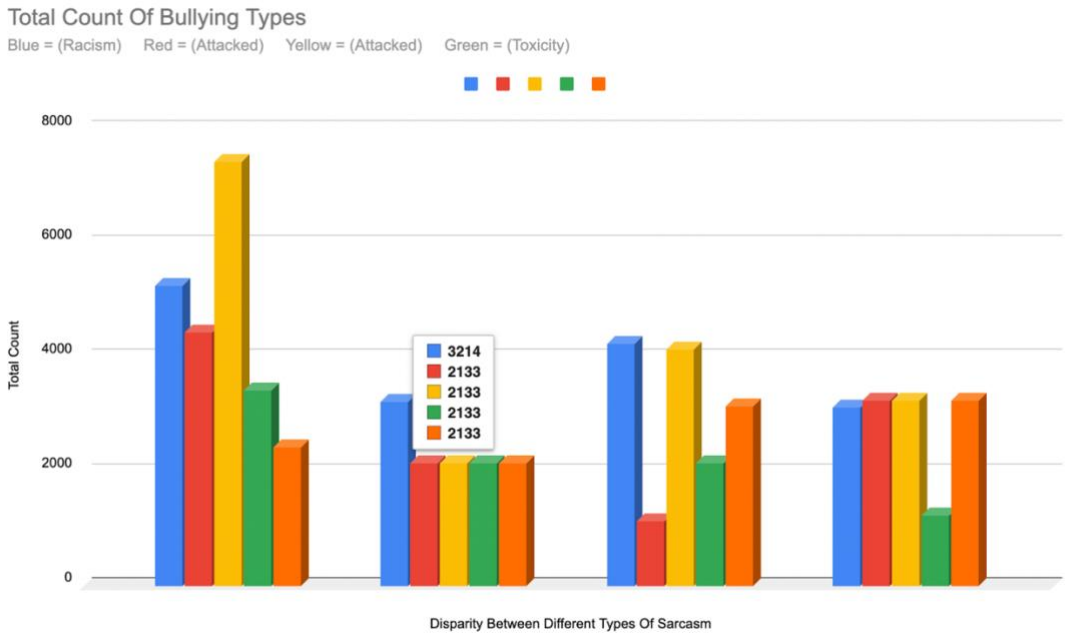


Figure 23: Total Count Of Bullying Types

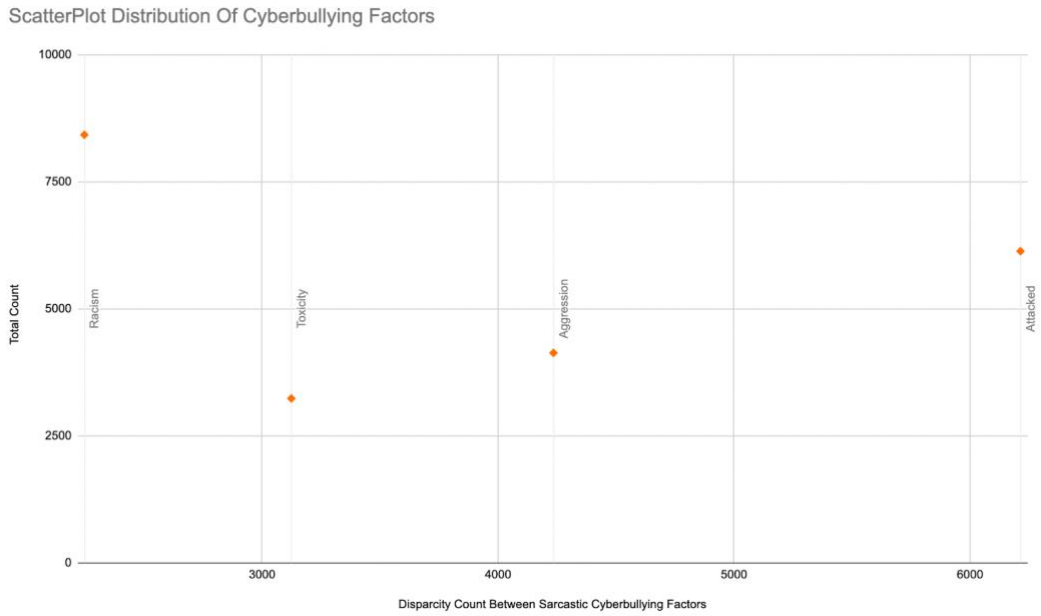


Figure 24: Scatterplot Distribution Of Cyberbullying Factors

Based on Figure 23 & 24, the Total Count Of Bullying Types allows us to identify & understand the distribution of what type of factors are mostly contributed to hate speech in conjunction with sarcastic cyberbullying while the Scatterplot Distribution allows us to understand the linearity between each cyberbullying factor. From the diagrams above, we can see that attacked comments are among the highest and toxicity along with aggression are highly related.

4.1.4 Model Data

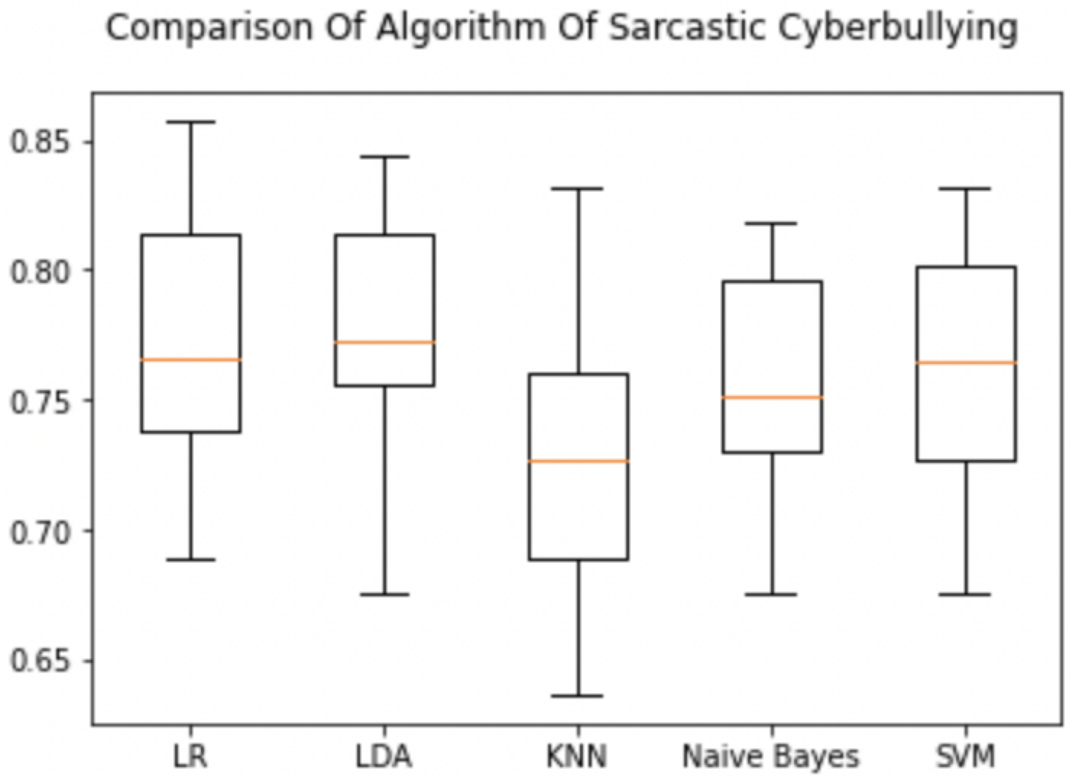


Figure 25: Comparison Of Algorithm Of Sarcastic Cyberbullying

Algorithm	Accuracy
Logistic Regression	LR: 0.769532 (0.052966)
Linear Discriminant Analysis	LDA: 0.773462 (0.051592)
KNN	KNN: 0.726555 (0.061821)
Naïve Bayes	Naive Bayes: 0.755178 (0.042766)
SVM	SVM: 0.760424 (0.052931)

Table 9: Accuracies Of Machine Learning Algorithms

In this section, the training dataset is trained using python's scikit_learn machine learning library to analyse the accuracies in detecting sarcastic cyberbullying in social media. The training data was trained on top of 5 different machine learning algorithms which are Logistic Regression, Linear Discriminant Analysis, KNN, Naïve Bayes & SVM. Through our findings, LDA provides the highest accuracy at 77%. Therefore, this algorithm was utilized for our categorical analysis.

4.1.5 Interpret Data

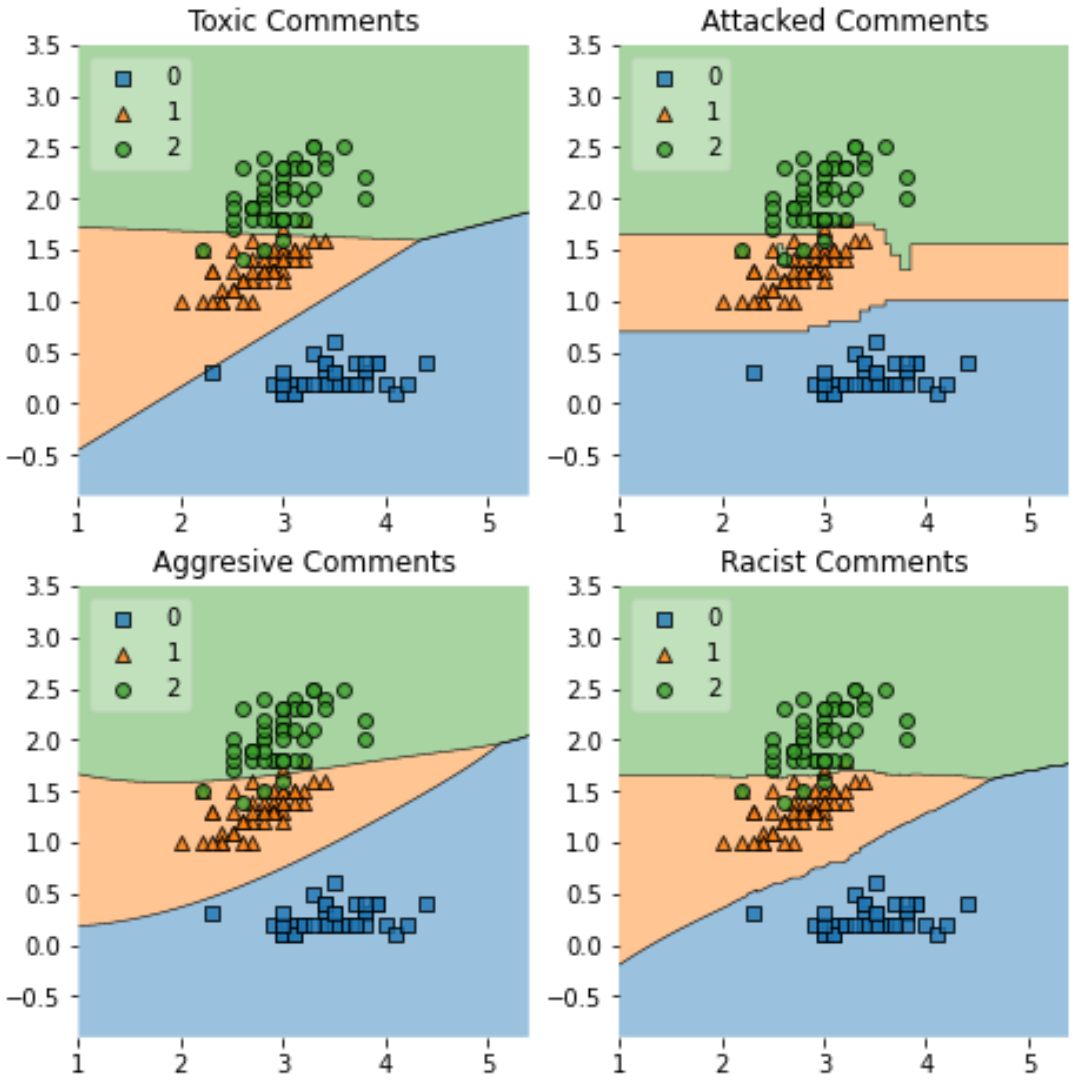


Figure 26: Data Distribution Sentiment Analysis Interpretation

From our training dataset, as our model, at 77% accuracy is able to proportionally detect sufficient relevancy in sarcastic cyberbullying from toxic, aggressive, racist & attacked characteristics, we now shall put it into good use. We have used another machine learning library called mlxtend to perform sentiment analysis on top of these sarcastic cyberbullying cases. Through our analysis, we have identified, based on the blue distribution in the diagram that a very large majority / portion is naturally negative from the comments while green is naturally positive while orange is neutral. From the mlxtend model, it is able to efficiently predict large cases from racist, aggressive & toxic comments. Generally, sarcasm is a negative sentiment. Therefore, our model being able to dissect and perform opinion mining on a sentiment analysis level gives us the understanding that sarcastic cyberbullying can be related based upon various hate speech characteristics.

5 Chapter 5: Results and Discussion

5.1 Interpretation Of The Best Model

Through the use of the attributes specified in Table 8 with 90,217 observations to train and validate, the sarcastic cyberbullying detector's best model is achieved was achieved at an accuracy of 77.35%. Thus, as our model was tested in Figure 25, the evaluation of results provided shows that Linear Discriminant Analysis is best applied for sarcastic cyber bullying in social media using hate speech factors.

5.2 Evaluating The Dashboard

This section describes the specification of the analytics web app user interface. The user interface (UI) specification can be illustrated below which contains information about available functions and provide an overview of the cyberbullying dashboard detector that users are expected to see when visited. It is available at the following link here.

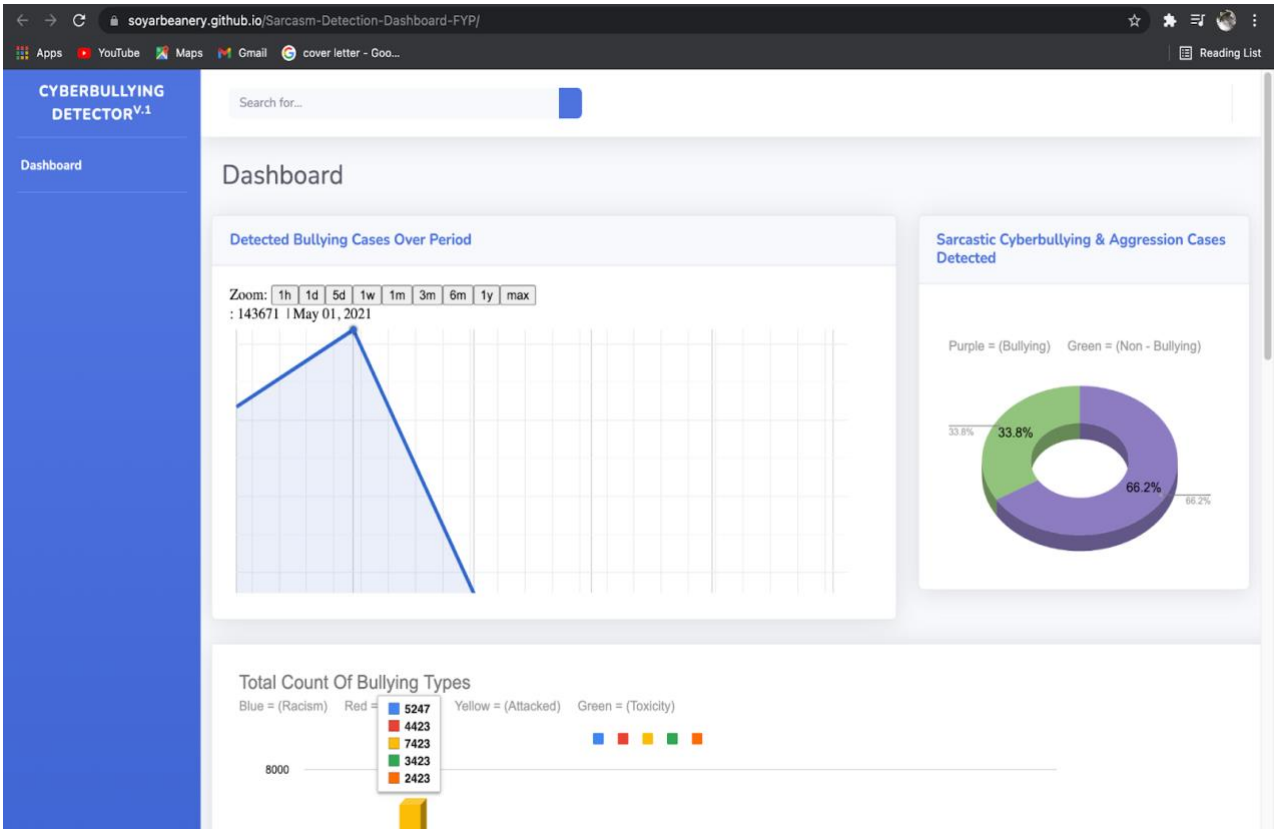


Figure 27: Screenshot Of Detected Bullying Cases In Web Dashboard User Interface

6 Chapter 6: Conclusion

Through our study, we have learnt that cyberbullying is a negative sentiment. This paper has studied on the negative impacts of sarcastic cyberbullying circulating on social media platforms as well as the lacking of attributes that can help the classification models to perform better with higher accuracies for automated fact-checking techniques. The final dataset presented in this paper encompasses tweet-specific attributes alongside with open – sourced dataset attributes derived from linguistics and psycholinguistics features from the training and validation of sarcastic cyberbullying detection models. The study has also shown that linguistics and psycholinguistics features are important factors in detecting sarcastic cyberbullying derived from hate speech characteristics. The achievement of this paper is that it is able to model with 77.35% accuracy which is higher than the other machine learning studies. Thus, we are confident that this research has further contributed to the field of study in detecting sarcastic cyberbullying cases in social media platforms and believes that it can be a strong basis for future research and studies to be based upon.

6.1 Limitations

Although we are achieving a high model performance in this paper, it does not come without its limitations. One of the limitations is that data privacy is a growing issue in most countries. As such, only information from tweets that are only set to 'public' can be gathered. As data privacy laws are becoming more and more restrictive, there may come a day where scraping Twitter data is no longer allowed and this may have an impact towards the current research. The second limitation is that linguistic and psycholinguistic databases are centred around English words. Therefore, sarcastic cyberbullying detection can only be done in the English language for this study. However, in a multilingual country like Malaysia, this is an issue as cyberbullying can still be spread in other languages such as the Malay language and Mandarin. The third limitation is that psycholinguistic features deal with the emotional expressions of words upon humans. This may be a problem as some writers of real news who are naturally expressive in their writing and use more flowery vocabularies. These vocabularies may come across as a distortion of emotional impact to the sarcastic cyberbullying classification models. In reaction to this, the sarcastic cyberbullying detection model may be inclined to classifying these comments as negative based sentiment even when they are positively being sarcastic.

6.2 Implications

This paper proposes an approach to classify sarcastic cyberbullying as a novel contribution to enhance automated cyberbullying detection. Cyberbullying is a growing social problem that inflicts detrimental impacts on online users. The identification of roles is a valuable contribution to future research as it can prompt closer monitoring of bullies and implicitly help victims through potential prevention. Currently, our approaches to identifying cyberbullying related roles focus only on individual comments and are quite heavily relied upon open sourced datasets to gather sufficient data for our training. As a recommendation, we aim to expand this further by considering an entire discussion and the discourse relationships between the posts within the considered discussion. This will enable us to get a better understanding of the roles played by different users in a discussion. Moreover, we intend to integrate cyberbullying and role classification as a single model and optimise performance further to provide an effective solution to cyberbullying problems in social media.

6.3 Future Work

It may come across with great satisfaction that this research has met its goals and objectives. Nevertheless, in the field of research of sarcastic cyberbullying, it is a continuous and ongoing process that can always be improved although it may be in small increments. For further down the line streaming and future work, researchers may possibly and should look into deriving other attributes that may help with increasing the accuracy of sarcastic cyberbullying detection

References

- Alloghani, Mohamed & Al-Jumeily Obe, Dhiya & Mustafina, Jamila & Hussain, Abir & Aljaaf, Ahmed. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. 10.1007/978-3-030-22475-2_1.
- Babvey, P., Capela, F., Cappa, C., Lipizzi, C., Petrowski, N., & Ramirez-Marquez, J. (2020). Using social media data for assessing children's exposure to violence during the COVID-19 pandemic. Child Abuse & Neglect, 104747.
- Borg, Ingwer & Mair, Patrick & Cetron, Joshua. (2020). Facet theory and multidimensional scaling.
- Elsafoury, Fatma (2020), "Cyberbullying datasets", Mendeley Data, V1, doi: 10.17632/jf4pzyvnpj.1
- Kim, J., Walsh, E., Pike, K., & Thompson, E. A. (2020). Cyberbullying and victimization and youth suicide risk: the buffering effects of school connectedness. The journal of school nursing, 36(4), 251-257.
- Karmakar, S., & Das, S. (2020, August). Evaluating the impact of covid-19 on cyberbullying through bayesian trend analysis. In Proceedings of The European Interdisciplinary Cybersecurity Conference (EICC) co-located with European Cyber Week.
- Parris, L., Lannin, D. G., Hynes, K., & Yazedjian, A. (2020). Exploring social media rumination: associations with bullying, cyberbullying, and distress. Journal of interpersonal violence, 0886260520946826.
- Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., & Wright, B. (2020). Sarcasm detection using machine learning algorithms in Twitter: A systematic review. International Journal of Market Research, 1470785320921779.
- Souad, Taleb & Adla, Abdelkader. (2020). Optimization Techniques for Machine Learning. 10.1007/978-981-15-0994-0_3.
- Lee, Wei-Meng. (2019). Supervised Learning—Classification Using Logistic Regression. 10.1002/9781119557500.ch7.
- Dineva, Kristina & Atanasova, Tatiana. (2018). OSEMN PROCESS FOR WORKING OVER DATA ACQUIRED BY IOT DEVICES MOUNTED IN BEEHIVES.

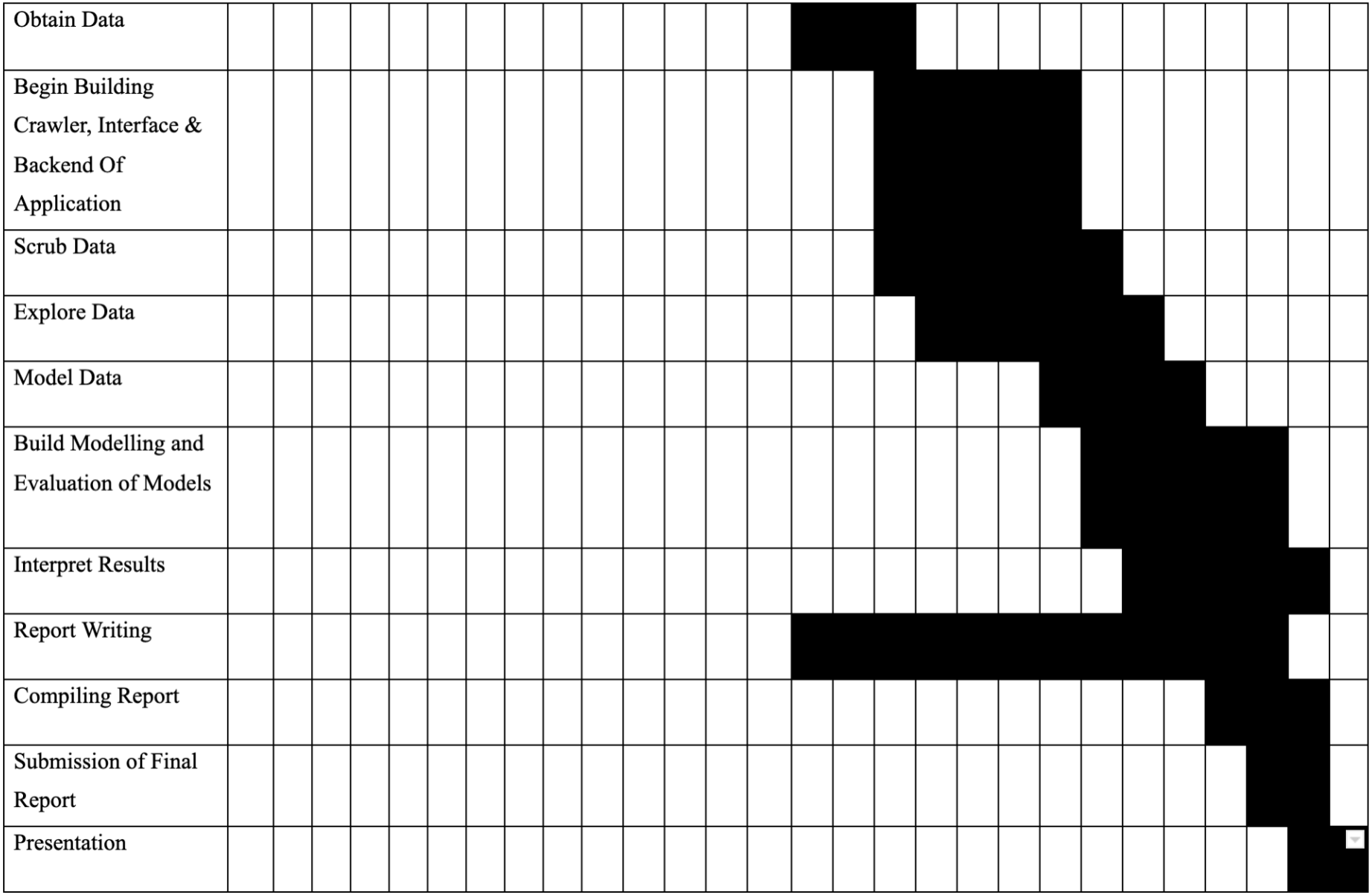
- Sreelakshmi, K., & Rafeeque, P. C. (2018, July). An effective approach for detection of sarcasm in tweets. In 2018 International CET Conference on Control, Communication, and Computing (IC4) (pp. 377-382). IEEE.
- Berger, Dale. (2017). Introduction to Binary Logistic Regression and Propensity Score Analysis.
- Baly, R., Hajj, H., Habash, N., Shaban, K. B., & El-Hajj, W. (2017). A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in Arabic. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 16(4), 1-21.
- Bharti, S. K., Pradhan, R., Babu, K. S., & Jena, S. K. (2017). Sarcasm analysis on twitter data using machine learning approaches. In Trends in Social Network Analysis (pp. 51-76). Springer, Cham.
- Chaudhari, P., & Chandankhede, C. (2017, March). Literature survey of sarcasm detection. In 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET) (pp. 2041-2046). IEEE.
- Espelage, D. L., & Hong, J. S. (2017). Cyberbullying prevention and intervention efforts: current knowledge and future directions. The Canadian Journal of Psychiatry, 62(6), 374-380.
- Bharti, S. K., Vachha, B., Pradhan, R. K., Babu, K. S., & Jena, S. K. (2016). Sarcastic sentiment detection in tweets streamed in real time: a big data approach. Digital Communications and Networks, 2(3), 108-121.
- Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. In Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis (pp. 161-169).
- Jolliffe, Ian & Cadima, Jorge. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 374. 20150202. 10.1098/rsta.2015.0202.
- Singh, Anagh & Prakash, B. & Chandrasekaran, K.. (2016). A comparison of linear discriminant analysis and ridge classifier on Twitter data. 133-138. 10.1109/CCAA.2016.7813704.
- Surlakar, Prachi & Araujo, Sufola & Sundaram, K.. (2016). Comparative Analysis of K-Means and K-Nearest Neighbor Image Segmentation Techniques. 96-100. 10.1109/IACC.2016.27.

- Cortis, K. & Handschuh, S. (2015). Analysis of cyberbullying tweets in trending world events. Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business. ACM, 7.
- Squicciarini, A., Rajtmajer, S., Liu, Y. & Griffin, C. (2015). Identification and characterization of cyberbullying dynamics in an online social network. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ACM, 280-285.
- Pabian, S., & Vandebosch, H. (2014). Using the theory of planned behaviour to understand cyberbullying: The importance of beliefs for developing interventions. European Journal of developmental psychology, 11(4), 463-477.
- Riff, D., Lacy, S. & Fico, F. (2014). Analyzing media messages: Using quantitative content analysis in research, Routledge.
- Diaf, A. & Boufama, Boubakeur & Benlamri, Rachid. (2013). Non-parametric Fisher's discriminant analysis with kernels for data classification. Pattern Recognition Letters. 34. 552–558. 10.1016/j.patrec.2012.10.030.
- E. Riloff et al., "Sarcasm as contrast between a positive sentiment and negative situation," in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, (2013), pp. 704–714.
- Xia, Fan & Chen, Jun & Fung, Wing & Li, Hongzhe. (2013). A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. Biometrics. 69. 10.1111/biom.12079.
- El-Habil, Abdalla. (2012). An Application on Multinomial Logistic Regression Model. Pak.j.stat.oper.res.. 8. 10.18187/pjsor.v8i2.234.
- Lodwich, Aleksander & Shafait, Faisal & Breuel, Thomas. (2011). Efficient Estimation of k for the Nearest Neighbors Class of Methods.
- Bosnjak, Zita & Grljevic, Olivera & Bošnjak, Saša. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. 509 - 514. 10.1109/SACI.2009.5136302.
- Privitera, C. & Campbell, M. A. (2009). Cyberbullying: the new face of workplace bullying? CyberPsychology & Behavior, 12, 395-400.
- Subrahmanyam, K., Reich, S. M., Waechter, N. & Espinoza, G. (2008). Online and offline social networks: Use of social networking sites by emerging adults. Journal of applied developmental psychology, 29, 420-4.

- Abdi, Hervé. (2007). Discriminant Correspondence Analysis. Encyclopedia of Measurement and Statistics.
- Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13, 210-230.
- R. J. Kreuz and G. M. Caucci, "Lexical influences on the perception of sarcasm," in Proceedings of the Workshop on computational approaches to Figurative Language. Association for Computational Linguistics, 2007, pp. 1–4.
- Kurita, T. (2004). Support Vector Machine and Generalization Takio Kurita.
- Perme, Maja & Blas, Mateja & Turk, Sandra. (2004). Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodološki Zvezki. 1. 143-161.
- Rish, Irina. (2001). An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell. 3.
- Evgeniou, Theodoros & Pontil, Massimiliano. (2001). Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.
- Balakrishnama, S. & Ganapathiraju, Aravind. (1998). Linear Discriminant Analysis—A Brief Tutorial. 11.
- Friedman, Nir & Geiger, Dan & Goldszmidt, Moises. (1997). Bayesian Network Classifiers. Machine Learning. 29. 131-163. 10.1023/A:1007465528199.
- Hamilton, Lawrence & Seyfrit, Carole. (1994). Interpreting multinomial logistic regression. Stata Technical Bulletin. 3

Appendix: Gantt Chart

Activities / Tasks	Week																										
	Capstone Project 1 (Diploma Final Year Project: 14 weeks)														Capstone Project 2												
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Planning Document																												
Literature Review																												
Methodology Literature Review																												
Finalize Literature Review																												
Initial Web Design & Model																												
Compilation of Planning Documents																												


Obtain Data
Begin Building Crawler, Interface & Backend Of Application
Scrub Data
Explore Data
Model Data
Build Modelling and Evaluation of Models
Interpret Results
Report Writing
Compiling Report
Submission of Final Report
Presentation