SCHOOL OF ENGINEERING AND TECHNOLOGY

ASSIGNMENT FOR THE
BSC (HONS) IS (BUSINESS ANALYTICS); YEAR 2
BSC (HONS) IS (DATA ANALYTICS); YEAR 2

ACADEMIC SESSION: MARCH 2022; SEMESTER 4 and 5

IST2134: SOCIAL MEDIA ANALYTICS

DEADLINE: (12th July 2022, Friday 11:59pm) via eLearn

STUDENT NAME: Tan Yi Shen	**STUDENT ID: 20026977**
STUDENT NAME: Lai Wei Qi	**STUDENT ID: 20035390**
STUDENT NAME: Grace Fung Wai Yan	**STUDENT ID: 20035713**
STUDENT NAME: Tai Jun-Li	**STUDENT ID: 19035781**
STUDENT NAME: Gwee Jing Ling	**STUDENT ID: 18043596**
STUDENT NAME: Teng Dale	**STUDENT ID: 18079517**

INSTRUCTIONS TO CANDIDATES

- This assignment will contribute 50% to your final grade.

IMPORTANT

The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 50%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Lecturer's Remark (Use additional sheet if required)

I (Names and IDs stated above) received the assignment and read the comments

……..... (Signature/date)

Academic Honesty Acknowledgement

"We, Tan Yi Shen, Lai Wei Qi, Grace Fung Wai Yan, Tai Jun-Li, Gwee Jing Ling and Teng Dale, (student names). verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

Tan, Lai, Grace, Tai, Gwee, Teng 19/6 (Student's signature / Date)

Table of Contents

Abstract 3
1. Introduction 3
1.1. Problem Statement 1
1.2. Objective 2
1.3. Structure of the paper 2
2. Literature Review 2
3. Methodology 5
3.1. Data Collection 5
3.2. Data Preparation 5
3.2.1. Data Cleaning 5
3.2.2. Data Pre-processing 6
3.2.3. Data Splitting 6
3.3. Feature Extraction 6
3.4. Model Selection and Analysis 7
3.4.1. Decision Tree 7
3.4.2. Logistic Regression 7
3.4.3. Naïve Bayes 8
3.4.4. K-Nearest Neighbor (KNN) 8
3.5. Performance Evaluation 9
4. Results Analysis 10
4.1. Model Performance 10
5. Discussion 12
6. Implication 12
7. Conclusion 13
8. Reflection 14
References 18

Abstract

According to various research, it is analyzed that one- third of teens have suffered from cyberbullying activities during the year 2020, which is suspected to be due to the 20% **longer** online social media **sessions like Twitter, Facebook, etc., that causes an inclination of cyberbullying cases** during **COVID-19** lockdowns, comparing to the years before the happening of COVID-19 pandemic. However, due to the occurrence of cyberbullying activities, it is also found that it may result in **negative impacts** towards the victims by affecting one's **mental health** to suffer from depression and anxiety as a result of lowered self-confidence and sense of value, which would result in having suicidal thoughts and self-harm as a way to cope with the negative feeling felt from the victims of cyberbullying. Hence, this paper is intended to propose with 4 types of **unsupervised** and **supervised** machine learning **classification models** namely **Decision Tree, Logistic Regression, Naïve Bayes,** and **K-Nearest Neighbor,** with proper **data pre-processing techniques** used including data cleaning, data reduction, etc., to **identify** and **detect cyberbullying texts** by scraping data from **Twitter**, as Twitter is also among the social medias to have the occurrence of cyber bullying activities. Moreover, **model performance measurement** such as **Confusion Matrix** and performance metrics of **accuracy**, precision, recall, f1score are also implemented, showing that the Decision Tree model performing the best for the scraped data obtained from Twitter with an accuracy result of 78%, while also achieving the highest true positive predicted value of 28.74% at detecting cyber bullying text and lowest false negative actual value of 12.4% at detecting non-cyberbullying texts, compared to the other 3 models in this research.

Keywords: *Cyberbullying, social media, Twitter, unsupervised learning, supervised learning, machine learning, classification models, Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, pre-processing techniques, identify and detect, scrape data, cyberbullying text, model performance measurement, Confusion Matrix, accuracy*

1. Introduction

Cyberbullying or cyber harassment is a way of bullying that occurs through electronical devices such as on computers, tablets, and smartphones, where it would happen online via **social medias**, communities, or entertainment where users can read, interact with, or exchange content, with the means of digital devices. **Preparator** of cyberbullying would often involves sending, publishing, or disseminating unfavorable, hurtful, or malicious material about

someone as well as disclosing sensitive or private information about an individual in a way that causes shame or degradation towards the **victims** of cyberbullying, which has the possibilities to drift into illegal or criminal action [32].

According to multiple research such as Pew Research Center and Security.org, it is analyzed that one-third of **teens** have suffered from cyberbullying activities [29], while 21% of parents mentioned that a kid in their family had already experienced cyberbullying during the year **2020**, where the cause of rising cases of cyberbullying is assumed to be due to the longer online sessions during COVID-19 lockdowns at the year 2020, as it can be seen that people across the world especially children, are using social media 20% more than they were prior to the COVID-19 pandemic lockdowns such as Snapchat, Twitter, YouTube, etc. [30]. Additionally, due to the occurrence of inclination of cyberbullying cases, one should be aware that the **negative impacts** of victims suffering from cyberbullying activities are:

i) **Emotional and Mental Health Effects**: According to a study, 32% of children who are the victims of cyberbullying say they have suffered at least one stress symptom, where the victims of cyberbully could also feel sad, wounded, humiliated, and even worry for their wellbeing. Due to suffering these emotions in the long run, it would lead to the victim succumbing depression and anxiety, where 93% of individuals who encountered cyberbullying expressed despair, helplessness, and hopelessness.

ii) **Physical Effects:** Bullying-related stress can also cause stomach problem related issues like stomach discomfort, and stomach ulcers. Children who are subjected to cyberbullying may also be prone to experience disordered eating patterns, missing meals or engaging in compulsive eating, as well as sleep disturbances due to victims of cyberbullies may experience nightmare and insomnia [31].

Therefore, this paper seeks to implement suitable **machine learning models** to identify and detect whether if a text is harmful or not harmful to be assumed as cyberbullying, to assist in reducing the negative impacts of cyberbullying cases as a paper, by using **4 supervised** and **unsupervised machine learning classification models**, to categorize data into distinct type concerning the probability of binary (Yes/No) event occurring, using models such as Decision Tree, Logistic Regression, Naïve Bayes, and K-Nearest Neighbor, with proper data pre-processing techniques of data cleaning, data transformation, and data reduction as shown below.

1.1. Problem Statement

Cyberbullying is a global issue in countries around the world. A survey covers from 24 countries found that nearly 80% of residents from worldwide considered cyberbullying as a critical problem and insufficient of online bullying measures are taken to address the problem [16]. As the term implies, cyberbullying involves using cyberspace as a way of bullying others, whether or not the bully is aware of it. The widespread usage of Social Networking Sites (SNSs) has significantly increased the likelihood and possibility of cyberbullying and cybervictimization [17]. Bullying place are commonly occurs through social media platforms such as Twitter, YouTube, Facebook, Snapchat and TikTok and sending text messages (i.e. iMessage) via mobile phones as well [18]. According to the cyberbullying statistics in 2022 [19], with a 11% of younger teens and 18% of teens have experienced being bullied through forwarding their private messages to others or publicize them to the public channels.

With the ever rising of cyberbullying cases on social media networks, it is considered one of the serious cybercrimes as it could cause great harm in both physical and mental health with far-reaching consequences. There are several of research papers proved that cyberbullying results in negative effects on the individual human-being, ranging from tension and fear to critical complications such as depression, suicide and self-destruction [8], [9], [10]. The reason behind these complications may be due to criticizing an individuals' race, sexuality, religion or appearance. A study showed that a number of 61% of teens being cyberbullied by appearance, followed by 25% of academic achievement, 17% of racism, 15% of sexual discrimination and financial difficulties as well as 11% of bullies attempts to mock someone religion, reported by [19].

In the recent years, it is found that there is an increasing numbers of unforeseen cases attributed to cyberbullying occurred in different types of social media platforms and applications [20]. A short 10 second Instagram/Facebook story can be shared all over the internet by just a simple click on SNSs including Twitter, Whatsapp, TikTok, Wechat, Discord, etc. within a few minutes. The online world is a freedom of speech after all, which means users are allowed to comment from both positive or negative side based on its perspective to the public and accessible to view any information on online. Therefore, the victim of cyberbullying runs the risk of becoming a target for millions of people in a short period of time. The following are the examples of tragedy cyberbullying cases happened in the past through social media sites:

i. Tyler Clementi, age 18 who committed suicide in 2010 by jumped down from George Washington bridges, United States. He leaped over the bridge because of a leaked video shared by his roommate in Twitter, which allegedly captured him kissing another man without his consent [27].
ii. Megan Meier, age 13 who committed suicide by hang herself to death in 2006 due to humiliation cyberbullying through social networking website, Myspace [28]. Throughout the investigation of the case, a friend of hers' created a fake profile which intended to gather information about her and subsequently humiliate her [27].

With the catastrophe mentioned above, there is an urge to detect cyberbully contents from social media platforms in order to enhance practices of non-cyberbullying and eliminate the online bullying cases in the future.

1.2. Objective

The objective of this study is to detect cyberbullying texts in social media in order to prevent cyberbullying activities from occurring in the future. By using various machine learning approaches, we observe which model performs better at cyberbullying detection to tackle cyberbully issues on social media.

1.3. Structure of the paper

The rest of the study is organized as follows. Section 2 provides a comprehensive review of the related works. Section 3 explains the research methodological approach and describes the machine learning model selections and analysis throughout the study. Section 4 analyses the results and analysis from the derived output. Section 5 discusses the findings related to the Section 4. Section 6 provides implications of this research and conclusion of this study is presented in Section 7. Lastly, the group reflection is presented in Section 8.

2. Literature Review

A study conducted by Reynolds et al. (2011) developed a decision tree model to detect cyberbullying on a random subset of data that was collected from the Formspring.me website [11]. The JRIP algorithm in Weka was used which generates a wide set of rules that is repeatedly reduced until the smallest rule set with the equal success rate is obtained. Furthermore, an instance-based (IBK) algorithm was employed with a k-nearest neighbour approach. The IBK method used in the study was k = 1, k = 3. Finally, a sequential minimal

optimization (SMO) algorithm was used. This is a support vector machine algorithm that is based around functionality. The decision tree generated that was large and complicated would imply that the data has been overfitted. Conversely, a modest and uncomplicated tree would suggest that the model cannot be plainly identified and there is unbalanced training data present in the model. The NORM chart showed that the duplication of positive cases by 7 to 8 times resulted in the final tree having 13 leaves, while the 9 to 10 weightings generated trees with 6 leaves. Ultimately, the positive cases were duplicated 8 times in order to produce a model with the maximum true positive success rate. This creates a decision tree that has contains the greatest equilibrium between tree size and accuracy of the model. The final model was able to detect a true positive rate with 78.5% accuracy.

A study conducted by Muneer and Fati (2020) attempted to creates many models to detect cyberbullying in a worldwide dataset of 37,373 unique tweets from Twitter [12]. Many machine learning classification methods were utilised in this study, including Logistic Regression, Stochastic Gradient Descent, Random Forest, Support Vector Machine, Light Gradient Boosting Machine AdaBoost and Naive Bayes. To evaluate all the algorithms, the precision, accuracy and the F1 score performance metric were documented to verify the cyberbullying recognition rate on the data. The data was split with a 70:30 ratio between the data for training and prediction. The logistic regression machine learning model is able to continuously update the parameters that have been set in order to minimise possible error. The findings of this study resulted in logistic regression being the best model by far, achieving a median accuracy of 90.57%. It also displayed the best F1 score among all the classifiers with a score of 0.93%.

A study done by Mouheb et al. (2019) conducted a Naïve Bayes algorithm for machine learning to automatically detect the cyberbullying present in Arabic social media streams [13]. The scheme that was proposed utilised a dataset that was gathered through Twitter and YouTube. The data here was randomly split 50:50 for the training and testing sets. The best conclusion resulted in a precision rate of 70.5%, a recall rate of 70.6% as well as a F-measure rate of 70.4%. Another paper from Alsubait and Alfageh (2021) compared the performance of various machine learning algorithms in cyberbullying recognition from a labelled dataset of Arabic YouTube comments [14]. The models used include Logistic Regression, Multinomial Naïve Bayes and Complement Naïve Bayes while the feature extraction methods that were used include TFIDF Vectorizer and Count Vectorizer. The F1 score is used to gauge the performance

of these classifiers. The final outcome shows that the Logistic Regression model is able to surpass the other models when the Count Vectorizer feature extraction is used. However, the Complement Naive Bayes model is the one to surpass the other models when the TFIDF Vectorizer feature extraction is used. The TFIDF Vectorizer is able to produce a better F1 score with an average score of 77.9% while the Count Vectorizer's F1 score has an average of 77.5%. Overall, the Complement Naive Bayes model is the classifier that produced the best performance based on the feature extraction.

Ozel et al. (2017) enhanced the cyberbullying detection system using machine learning methods such as Support Vector Machine (SVM), Decision Tree, Multinomial Naïve Bayes (MNB) and K-Nearest Neighbor (KNN) on Turkish texts. The authors constructed a dataset from 900 messages from Twitter and Instagram [5]. A number of 450 messages were classified as cyberbullying and the rest considered as irrelevant cyberbullying contents. To identify whether the accuracy will be maintained or improved, this study uses features collection which are Information Gain and Chi-square methods. As a result, MNB model performed the best without applying features, while KNN model gained the best accuracy when apply with features. In comparison, SVM models was nether than MNB and KNN models, and Decision Tree model gave the lowest accuracy among all classifiers in this research. This study demonstrated that KNN models outperformed the SVM and Decision Tree model in detecting cyberbully.

Nurrahmi and Nurjanah (2018) employed K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) models to learn and detect cyberbullying from the Indonesian text. The data collection was done by utilizing a web scraper tool, Selenium and collected 700 tweets from Twitter as a dataset [1]. The authors then utilized feature extractions which were identified by another paper [2] to better determine text that contained explicit harassment such as the number of bad words from the tweets, the number of words that shows negative and positive emotions, etc. After determining feature extractions, the authors created a labelling system for participants to label the cyberbullying and non-cyberbullying tweets, ranging from 1 to 4, which indicates the weights of cyberbully and non-cyberbullying for participants to choose from the comments of the four controversial posts provided by the authors and select which were the most and least cyberbullying tweets according to participants' opinions. Following that, the data cleaning step is applied to the gathered data which involves removing special characters and URLs from posts on Twitter. The results were obtained from the authors with

301 cyberbullying tweets, 399 non-cyberbullying tweets, 2053 negative words and 129 swear words. KNN and SVM models were then applied and fed by the pre-processed gathered data and the aforementioned feature extractions to perform cyberbullying classification. The labelled tweets were classified by these two models and evaluated using precision, recall and F1-score. KNN algorithm obtained with a F1-score of 0.66% while 0.61% of F1-score from SVM Linear algorithm and 0.67% of F1-score were evaluated in the SVM RBF algorithm. The results showed that KNN algorithm is considered as a preferable classifier at detecting cyberbullying tweets compared to SVM Linear algorithm in this study.

3. Methodology

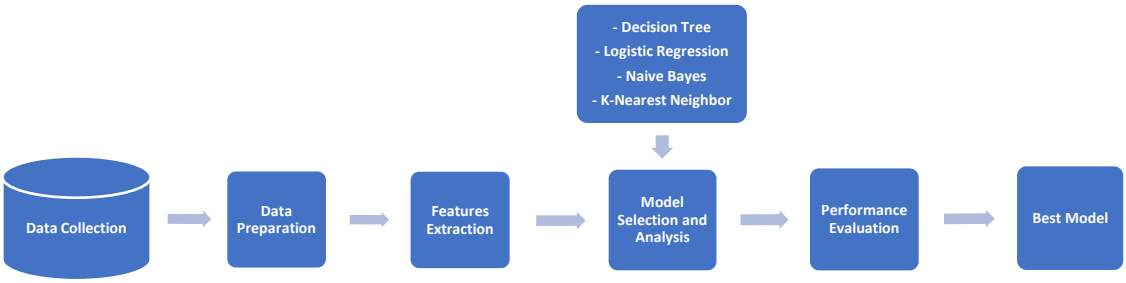

Figure 1: Classification Methodology

3.1. Data Collection

Firstly, the sentiment pipeline was defined. To collect the tweets, the roberta base-sentiment model which is a Natural Language Processing (NLP) model was used. This base model was trained on ~58M tweets and further finetuned for sentiment analysis with the TweetEval standard. Next, TwitterSearchScraper is used to scrape data and append the tweets to a list. A total of 10,000 tweets that contained "amber heard" was collected. The tweets were all posted within the period between 18th of May 2022 and 20th of May 2022. The data collected consists of the tweet date, tweet ID and tweet content. Once compiled, the data is then stored into a data frame and converted to a csv file.

3.2. Data Preparation

3.2.1. Data Cleaning

The cleaning process involves removing the common words that are do not carry much useful information for the analysis. This means implementing stop words for their removal. The automatic stop words are set to English as it is the primary language found in the data.

Furthermore, manual stop words have also been stated such as renditions of certain names that are deemed unnecessary. Additionally, the noise in the data such as hyperlinks, usernames and numbers are also removed.

3.2.2. Data Pre-processing

The 'nltk' library is used for the data pre-processing. All the tweets are also tokenized so that the unnecessary characters such as punctuation is removed. A stemming process using the Porter Stemmer as the stemming algorithm is applied. This is to make the tweet content more transparent so that interpretation can have an improved accuracy. Furthermore, the data is tokenized with a word tokenizer. Lemmatization is conducted so that the context of the tweets become more apparent. The 'Wordnet' lexical analyser is used to complete this process. A sentiment polarity of positive, negative and neutral is also assigned to the tweets based on the content. For this study, TextBlob from the nltk library which supports procedures on textual data. Once the polarity has been assigned, the neutral tweets are dropped. A polarity label is then assigned based on the polarity for easier analysing process. "1" is assigned to negative tweets while "0" is assigned to positive tweets. A word cloud of the most used words based on the polarity label is generated for both the positive words as well as negative words.

3.2.3. Data Splitting

The pre-processed data were then split into training dataset and test dataset with a 70:30 ratio, which 70% of the training data is utilized to train the classification model, which then uses the training data to discover classification rules, and the remaining 30% of the test data is utilized to evaluate the performance of the classification models.

3.3. Feature Extraction

CountVectorization is implemented as feature extraction in this study to count the frequency count that occurs on each words. It allows to tokenize text documents in collection, build a vocabulary of known words and utilizes that vocabulary to encode new documents. The following examples show of using CountVectorization to get vectors:

- Create an instance of the CountVectorizer class.
- To learn a vocabulary from one or more documents, call `fit()` function is needed.
- To encode each documents, call `transform()` function is needed.

It is necessary to convert the source documents into vector representations as to perform machine learning on the text [6]. Python provides a package called spicy.sparce, it is to handle the sparse vectors that contains a lot of zero [7].

3.4. Model Selection and Analysis

A brief description of the machine learning algorithms, with Decision Tree, Logistic Regression, Naïve Bayes and K-Nearest Neighbor (KNN) are provided as follows:

3.4.1. Decision Tree

Decision tree is one of the popular approaches for solving classification and regression issues in machine learning. The decision tree model operates merely by guiding a transaction in a certain way based on the features extracted from the data. It proceeds from a fundamental root question and branches, which the specifics are used to create individual elements that ultimately result in the leaves of the decision tree. When continuous data partitioning is based on a particular parameter, decision trees are non-parametric supervised learning techniques that can be utilized for classification and regression applications. Decision tree is used to build a training model that can predict the class of the response variable and make predictions that classify whether a transaction is a fraud or not [25]. The advantage of this algorithm is that they can handle categorical attributes and do not require feature scaling [4]. These characteristics make decision tree model a preferable algorithm for cyberbully detection since they enable them to handle classification and regression problems better than rules-based models.

3.4.2. Logistic Regression

Logistic regression is a type of machine learning model that is conducted when the dependent variable is a binary or dichotomous. It utilises the logistic function also known as the signoid function at the foundation of the method [12]. As the probability will always be in between “0” and “1”, the S-shaped curve of the logistic function is used to obtain a value between that. For training data to be appropriate for a logistic regression, the sample sizes should be fairly large for the result to be accurate. Binary logistic regression is used to predict events that only have two possible outcomes, also known as binary classification. For example, to detect whether a customer is likely to be infected by a certain disease (yes/no). However, other types of logistic regression also exist where a different number of predicted outcomes is used. A multinomial logistic regression deals with dependent variables that possess multiple outcomes.

For example, to predict if an individual had the flu, an allergic reaction or Covid-19 based on their portrayed symptoms. Lastly, ordinal logistic regression are used for multiple outcomes that are ordered. For example, predicting the severity of Covid cases by classifying them into mild, moderate and serious.

3.4.3. Naïve Bayes

The Naive Bayes classifier is a type of machine learning model that is used for classification tasks through probability. This is done based on the Bayes theorem that can be written as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

By utilising Bayes theorem, the probability of A is found given that B has also occurred. B is the parameters that are used as evidence while A is the hypothesis. The assumptions that are present when using this theorem is that the predictors are all independent and all have an equal influence on the result. These algorithms are mainly used for filtering spam and sentiment analysis. There are multiple types of Naive Bayes classifiers which include Multinomial Naive Bayes, Bernoulli Naive Bayes and Gaussian Naive Bayes. Firstly, the Multinomial Naive Bayes type is used to classify outcomes into multiple categories based on the predictors present. Secondly, Bernoulli Naive Bayes is also used to classify outcomes into categories with the difference being that predictors used are made up of Boolean values. These parameters can only exist as yes or no. For example, to categorise based on if a certain word exist in a document. Lastly, the Gaussian Naive Bayes type is used when the predictors are made up of continuous values. An assumption used in this model is that the values that are sampled will make up a gaussian distribution [12].

3.4.4. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is one of the simplest algorithms for machine learning based on supervised learning techniques. The KNN algorithms assume the similarity between the new case/data and the existing cases and assign the new case to the class that is most similar to the existing class. Moreover, the KNN algorithm classifies new data based on similarity and stores all available data. This implies that as new data is generated, it may be quickly categorized using the KNN algorithm into relevant categories. The KNN algorithm may be used for both Regression and Classification problems, while it is mainly used for Classification issues. In addition, the KNN algorithm is non-parametric, which means it does not make any assumptions

about the underlying data. It is also known as a lazy learning algorithm because it stores the dataset instead of understanding it immediately from the training set. It only performs operations on the dataset when classification is required. The KNN method simply keeps the information during the training phase, and when it receives new data, it categorizes it into a category that is quite similar to the new data [3].

3.5. Performance Evaluation

The models are evaluated by predicting the class of the transactions in the test dataset and the best model will be selected after the performance assessment. This will demonstrate how each algorithm performed and allow us to determine whether it predicts cyberbully with satisfactory or unsatisfactory results. Confusion Matrix (Table 1) provides us a thorough assessment of the model performance in terms of matrix output with "0" denoting Negative or False and "1" denoting Positive or True.

	Predicted Negative (0)	**Predicted Positive (1)**
Actual Negative (0)	True Negative (TNs)	False Positive (FPs)
Actual Positive (1)	False Negative (FNs)	True Positive (TPs)

Table 1: Confusion Matrix

The model performance is assessed by the number of true negatives (TNs), false positives (FPs), false negatives (FNs), and true positives (TPs):

- **True Negatives** – The algorithm predicted NO, but the actual result was NO.
- **False Positives** – The algorithm predicted YES, but the actual result was NO.
- **False Negatives** – The algorithm predicted NO, but the actual result was YES.
- **True Positives** – The algorithm predicted YES, but the actual result was YES.

The primary evaluation metric employed in this study is Accuracy as it is the most frequently used metric among all metrics to weigh the model performance in classification algorithms, although it is not the only method to assess the model. In this study, machine learning models were assessed using evaluation metrics including Precision, Recall, and F-Measure/F1-Score. Several frequently employed performance metrics for text classification are briefly reviewed here:

- **Accuracy** is defined as the ratio of accurate predictions to the total number of observations.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions\ Made}$$

- **Precision** is the ratio of true positive predictions to total predictions.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- **Recall** is the ratio of accurate predictions to all accurate observations in the sample space.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- **F-Measure** is nothing more than the weighted harmonic mean of precision and recall.

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

4. Results Analysis

4.1. Model Performance

Each machine learning model will be evaluated with the confusion matrix and performance metrics, and we will present a comparative analysis to determine which of the models is the best model for predicting cyberbully tweets in this study. The binary numbers are "0" and "1" where "0" denotes as non-cyberbullying while "1" denotes as cyberbullying.

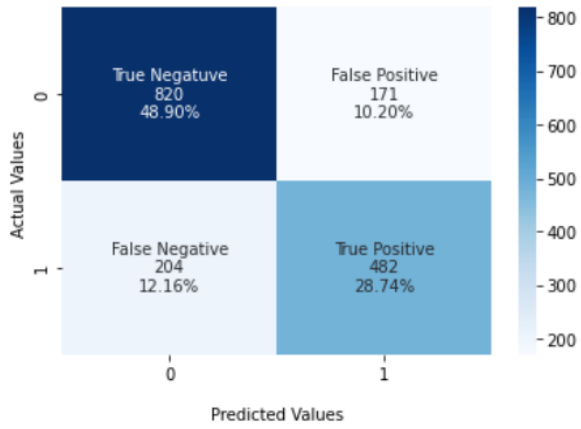

Figure 2: Decision Tree

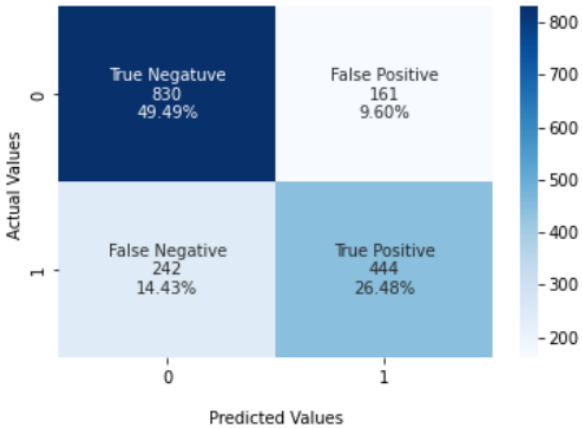

Figure 3: Logistic Regression

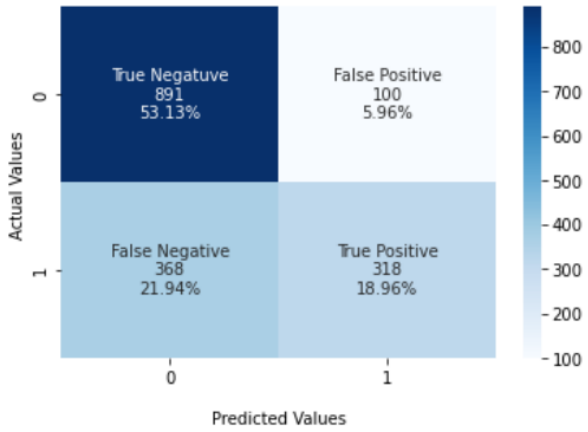

Figure 4: Naïve Bayes

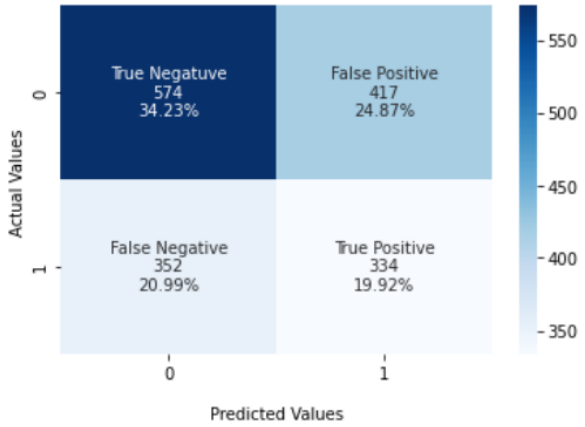

Figure 5: KNN

In the figures above shows the output of confusion matrix for Decision Tree model, Logistic Regression model, Naïve Bayes and KNN model. The true positive results of Decision Tree and Logistic Regression models which obtained values of 482 and 444, respectively, indicating the two highest true positive value compared to other two models. This means that the Decision Tree model were able to correctly predict 28% of text as cyberbully, while 26% of cyberbully text is predicted in Logistic Regression model. With the false negative value of 171 and 161 from Decision Tree and Logistic Regression, both models performed the lowest value, which means they were 1.53% and 1.89% incorrectly predicted as cyberbully. In conclusion, the best model to predict cyberbully are Decision Tree model as it produced the lowest value of false negative value and highest true positive value among all other models.

Machine Learning Models	**Performance Metrics (Percentage)**			
	Accuracy	**Precision**	**Recall**	**F-Measure/ F1-Score**
Decision Tree	78%	74%	70%	72%
Logistic Regression	76%	73%	65%	69%
Naïve Bayes	72%	76%	46%	58%
KNN	54%	44%	49%	46%

Table 2: Comparative Analysis

Table 2 above presents the results for the performance of the supervised machine learning algorithms, with Decision Tree, Logistic Regression, Naïve Bayes, and K-Nearest Neighbor (KNN) being the classifiers to be evaluated in this study. According to the findings, the

Decision Tree, Logistic Regression and Naïve Bayes have a high accuracy of 78%, 76% and 72% correspondingly, whereas KNN model had the lowest accuracy of 56%. The Naïve Bayes, Decision Tree and Logistic Regression models, which had a precision of 76%, 74% and 73% respectively, were accurately predicted positive outcomes, whereas KNN models had the lowest precision of 54%. In comparison to the Decision Tree and Logistic Regression models have higher recall of 70% and 65% than the KNN and Naïve Bayes models, which have a lower recall of 49% and 46% of all observations that the model were correctly predicted. The Decision Tree and Logistic Regression models have higher F-Measures of 72% and 69% because they have a greater harmonic mean between accuracy and recall values than the Naïve Bayes and KNN models, which have lower F-Measures of 58% and 46% accordingly. In conclusion, these findings indicates that Decision Tree model showed the highest accuracy and other performance indicators compared to other classification models while attempting to predict whether the text are cyberbullied or non-cyberbullied.

5. Discussion

The results obtained is limited as the feature extractions that have been applied are only limited to the four methods of Decision Tree, Logistic Regression, Naïve Bayes as well as KNN. This may have been the cause of the low performance metrics that were documented as the highest accuracy, precision, recall and F1 score only managed to reach above 70%. Furthermore, the sample size of the dataset used was set to fetch only 10,000 tweets as the running time of the program would be further decreased if more tweets were added. Many of the studies that were referenced and discussed in the literature review all used datasets that were much larger which may have been a big contribution to more successful results. Therefore, this lower number of tweets used may have negatively affected all the performance metrics in this study. Additionally, the data cleaning and processing that was done were also limited and may have resulted in a less reliable dataset for the analysis.

6. Implication

The present study developed a comprehensive framework for social media analytics and a cyberbullying detection system based on several machine learning algorithms for detecting the cyberbullying content underlying massive amounts of tweets on Twitter. The main contribution of this study is that it presents a detection system to recognize signals of cyberbullying on social

media by analysing historical data to learn the language patterns of the bullies and using the findings discovered in this study to prevent current and future cyberbullying practices. It will be beneficial to society in understanding the prevalence and severity of cyberbullying as well as the features that should be evaluated for when determining whether a social media text contains cyberbullying content. For instance, this study will assist data science researchers in identifying which offensive comments or tweets that would constitute cyberbullying, and the system can be developed to flag online conversations that contain specifically insulting terms. Hence, the social media platforms and online communities can implement text filters accordingly to identify offensive content, activate a warning alert before potential cyberbullying comments are posted to give users an opportunity to reconsider their words prior to posting, and even forbid it from being posted within their platform or community. In addition, these machine learning approaches are adding additional features to enhance the ability and accuracy of detecting cyberbullying on social media platforms, which it improves the performance of the cyberbullying detection system. The findings demonstrate that various machine learning techniques can indeed be applied in detecting cyberbullying texts on social media, with the Decision Tree model having the highest accuracy among all machine learning classifiers.

7. Conclusion

Cyberbullying is a bully committed online that would occur via chat services, gaming platforms, and social media. It is a set of actions meant to frighten, infuriate, or embarrass people from those who are the subject of cyberbullies such as posting incorrect facts or publishing humiliating photographs or videos of others online, etc. [21]. Therefore, to assists in **identifying** and **stopping potential cyberbullying practices** with the aim to **reduce** the **negative impacts** of cyberbullying on victim, as mentioned on section 1.0 of introduction in this document, 4 various machine learning classification model of **Decision Tree, Logistic Regression, Naïve Bayes**, and **K-Nearest Neighbor** have been conducted to determine which texts are deemed harmful and which are not, by using the scraped cyberbully tweets found from Twitter during 18th May to 20th May 2022. Some of the **significant results** from the 4 machine learning models built to detect harmful texts as seen from section 4.2. are:

i) **Highest performing accuracy performance metrics:** The **Decision Tree** had produced the **highest accuracy** result of **78%**, compared to other 3 models, with Logistic Regression having 76% accuracy score, while the other 2 models obtained

accuracy scores of 72% and lower, with reference to the **performance metrics** made to analyze the accuracy of 4 models built by defining the ratio of true positives and true negatives to all positive and negative observations for accuracy metric.

ii) **Highest performing confusion matrix metrics:** The **Decision Tree** had produced **the highest overall predicted labels vs actual labels result,** outputting the **highest true positive predicted value** of 28.74% detecting cyber bullying text and **lowest false negative actual value** of 12.4% detecting non-cyberbullying text compared to other 3 models, with reference to the confusion matrix made to analyze the prediction performance of 4 models built from the scikit learn machine learning library obtained.

With the results mentioned above, the **highest** performing machine learning model for the research analyzed is the **Decision Tree** classification model, showing that Decision Tree model is a prominent model to be used for text classification to identify which text is considered as harmful and which is not by using the scraped data obtained from Twitter. Through researches of related cyberbullying detection using machine learning models, for future work recommendations, it can be suggested that Deep Learning algorithms machine learning model may be used to produce promising results of text classification to detect cyberbullying online as seen from research paper by authors [22],[23], showing a top performing text classifications by using deep learning models with accuracy score above 84%, which is higher than the top 2 machine learning models implemented in this paper.

8. Reflection

Throughout this assignment, we have learned the importance of effective communication to complete the assignment in the given time. Effective communication will prevent misunderstanding and also prevent conflicts between the group members and smoothen the process of completing the assignment. One of the challenges that we faced is that most of us have other assignments deadline in the corner and affecting the progress of the Social Media Analytics assignment. However, all of the group members helped each other out to complete the assignment and gave out the best ideas and opinions to overcome the challenge. Throughout the assignment, our group members collected and gathered up the information and any relevant documents that are needed to complete this assignment. When one of us are facing any problems or difficulties, we tend to seek help from our group members and ask for their opinion. This assignment benefits us in understanding the importance of teamwork and allows us to make a more detailed study research on cyberbullying by applying machine learning

algorithms. Cyberbullying should not be foreseen and actions must be taken to reduce and prevent cyberbullying for the good mental health of everyone and also leading to a friendly and more comfortable environment for everyone to live in.

References

[1] H. Nurrahmi and D. Nurjanah, "Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility," *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 543-548, doi: 10.1109/ICOIACT.2018.8350758.

[2] G. Sarna and M. P. S. Bhatia, "Content based approach to find the credibility of user in social networks: an application of cyberbullying," International Journal of Machine Learning and Cybernetics, vol. 8, no. 2, pp. 677–689, Nov. 2015, doi: 10.1007/s13042-015-0463-1.

[3] "K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint," *www.javatpoint.com*. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[4] B. A. Talpur and D. O'Sullivan, "Cyberbullying severity detection: A machine learning approach," *PLOS ONE*, vol. 15, no. 10, p. e0240924, Oct. 2020, doi: 10.1371/journal.pone.0240924.

[5] S. A. Özel, E. Saraç, S. Akdemir, and H. Aksu, "Detection of cyberbullying on social media messages in Turkish," *IEEE Xplore, Oct. 01, 2017*. https://ieeexplore.ieee.org/abstract/document/8093411/ (accessed Sep. 18, 2020).

[6] V. Andreevich Kozhevnikov and E. Sergeevna Pankratova, "B'Theoretical & Applied ScienceB,'" www.t-science.org, May 30, 2020. http://www.t-science.org/arxivDOI/2020/05-85/05-85-106.html

[7] "Sparse matrices (scipy.sparse) — SciPy v1.8.0 Manual," docs.scipy.org. https://docs.scipy.org/doc/scipy/reference/sparse.html

[8] A. John et al., "Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review," *Journal of Medical Internet Research*, vol. 20, no. 4, p. e129, Apr. 2018, doi: 10.2196/jmir.9044.

[9] J. Escobar Echavarría, L. Elisa Montoya González, D. Restrepo Bernal, and D. Mejía Rodríguez, "Cyberbullying and suicidal behaviour: What is the connection? About a case," *www.elsevier.es*, Oct. 18, 2017. https://www.elsevier.es/en-revista-revista-colombiana-psiquiatria-english-edition--479-pdf-S2530312017300577

[10] C.-F. Yen, T.-L. Liu, P. Yang, and H.-F. Hu, "Risk and Protective Factors of Suicidal Ideation and Attempt among Adolescents with Different Types of School Bullying Involvement," *Archives of Suicide Research: Official Journal of the International*

Academy for Suicide Research, vol. 19, no. 4, pp. 435–452, 2015, doi: 10.1080/13811118.2015.1004490.

[11] Reynolds, K., Kontostathis, A., & Edwards, L. (2011). Using Machine Learning to Detect Cyberbullying. 2011 10th International Conference on Machine Learning and Applications and Workshops. https://doi.org/10.1109/icmla.2011.152

[12] Muneer, A., & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. Future Internet, 12(11), 187–207. https://doi.org/10.3390/fi12110187

[13] Mouheb, D., Albarghash, R., Mowakeh, M. F., Aghbari, Z. A., & Kamel, I. (2019). Detection of Arabic Cyberbullying on Social Networks using Machine Learning. 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). https://doi.org/10.1109/aiccsa47632.2019.9035276

[14] Alsubait, T., & Alfageh, D. (2021). Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments. International Journal of Computer Science and Network Security, 21(1), 1–5. https://doi.org/10.22937/IJCSNS.2021.21.1.1

[15] Barlett, C. P., Rinker, A. & Roth, B. (2021). Cyberbullying perpetration in the COVID-19 era: an application of general strain theory. The Journal of Social Psychology, 161(4), 1–11.

[16] Z. Villines, "Cyberbullying: A Global Trend," IDG Connect, May 22, 2014. https://www.idgconnect.com/article/3576618/cyberbullying-a-global-trend.html

[17] A. Sengupta and A. Chaudhuri, "Are social networking sites a source of online harassment for teens? Evidence from survey data," *Children and Youth Services Review*, vol. 33, no. 2, pp. 284–290, 2011, [Online]. Available: https://ideas.repec.org/a/eee/cysrev/v33y2011i2p284-290.html

[18] UNICEF, "UNICEF poll: More than a Third of Young People in 30 Countries Report Being a Victim of Online Bullying," *Unicef.org*, Sep. 03, 2019. https://www.unicef.org/press-releases/unicef-poll-more-third-young-people-30-countries-report-being-victim-online-bullying

[19] A. Skrba, "FirstSiteGuide team," *FirstSiteGuide*, Nov. 13, 2019. https://firstsiteguide.com/cyberbullying-stats/

[20] Enough is Enough, "Enough Is Enough: Cyberbullying," *Enough.org*, 2015. https://enough.org/stats_cyberbullying

[21] UNICEF, "Cyberbullying: What is it and how to stop it," *www.unicef.org*, Feb. 2022. https://www.unicef.org/end-violence/how-to-stop-cyberbullying (accessed Jul. 12, 2022).

[22] Md. T. Ahmed, M. Rahman, S. Nur, A. Islam, and D. Das, "Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study," *IEEE Xplore*, Feb. 01, 2021. https://ieeexplore.ieee.org/document/9392608 (accessed Apr. 30, 2022).

[23] A. Dewani, M. A. Memon, and S. Bhatti, "Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data," *Journal of Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00550-7.

[24] Wolke D, Lee K, Guy A. Cyberbullying: a storm in a teacup?. *Eur Child Adolesc Psychiatry*. 2017;26(8):899-908. doi:10.1007/s00787-017-0954-6

[25] Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 2(01), 20–28. https://doi.org/10.38094/jastt20165

[26] S. Gordon. "What is cyberbullying?". *Verywellfamily.com*. March 19, 2022.

[27] "List of suicides attributed to bullying," *Wikipedia*, May 04, 2022. https://en.wikipedia.org/wiki/List_of_suicides_attributed_to_bullying

[28] "Featured Content on Myspace," *Myspace*, 2014. https://myspace.com/

[29] A. Lenhart, "Cyberbullying," *Pew Research Center: Internet, Science & Tech*, Jun. 27, 2007. https://www.pewresearch.org/internet/2007/06/27/cyberbullying/ (accessed Jul. 12, 2022).

[30] Security.org Team, "Cyberbullying Prevalence and Factors in 2020," *Security.org*, Jul. 24, 2020. https://www.security.org/digital-safety/cyberbullying-covid/#references (accessed Jul. 12, 2022).

[31] C. L. Nixon, "Current perspectives: the impact of cyberbullying on adolescent health," *Adolescent Health, Medicine and Therapeutics*, Aug. 01, 2014. https://www.dovepress.com/current-perspectives-the-impact-of-cyberbullying-on-adolescent-health-peer-reviewed-fulltext-article-AHMT (accessed Jul. 12, 2022).

[32] Assistant Secretary for Public Affairs (ASPA), "What Is Cyberbullying," *StopBullying.gov*, Sep. 24, 2019. https://www.stopbullying.gov/cyberbullying/what-is-it (accessed Jul. 12, 2022).