


Identification and Detection of Cyberbullying on Facebook Using Machine Learning Algorithms

Nureni Ayofe AZEEZ, Department of Computer Sciences, University of Lagos, Nigeria

Sanjay Misra, Ostfold University College, Halden, Norway*

 <https://orcid.org/0000-0002-3556-9331>

Omotola Ifeoluwa LAWAL, Department of Computer Sciences, University of Lagos, Nigeria

Jonathan Oluranti, Covenant University, Nigeria

ABSTRACT

The use of social media platforms such as WhatsApp, Facebook, Instagram, and Twitter has facilitated efficient, effective, and frequent communication amongst people. Despite the numerous benefits associated with these social media platforms, they have also resulted in cyberbullying, which frequently occurs while using these networks. Cyberbullying is known to be the cause of some serious health, emotional, psychological, and social issues among social media users. With damages done globally with this social media threat, creating a way to identify and detect it is very significantly important. Against this backdrop, this paper takes a look at unique features obtained from the Facebook dataset and utilized machine learning algorithms to identify and detect cyberbullying posts and subsequently notify the internet users of some undesirable features they should desist from when they are being harassed or bullied in cyberspace. The algorithms used are naïve Bayes and k-nearest neighbor. The study also uses a feature selection algorithm, namely the χ^2 test (chi-square test) to select important features leading to improvement in the classification performance. The result of the study indicates the detection of cyberbullying on Facebook with a high degree of accuracy with the selected machine learning algorithms along with the chosen metrics for performance evaluation. Specifically, the k-nearest neighbor performed better when compared to naïve Bayes classifier with much improvement in the performance and classification time.

Keywords Accuracy, Algorithms, Classifiers, Cyberbullying, Features, Performance

1. INTRODUCTION

The continuous usage of social media platforms such as WhatsApp, Facebook, Instagram, Twitter, etc., has made it possible for many people to communicate effectively, efficiently and frequently with other people. With this development, the usage of social media has undoubtedly created an opening for some users to intimidate users with mean and nasty comments. They also go a step further by

DOI: 10.4018/JCIT.296254

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

posting derogative messages intending to belittle other users on those same platforms. This scenario is known as Cyberbullying (Al-Garadi et al., 2016; Fridh et al., 2015; Dadvar et al., 2013).

Cyberbullying is the harassment or insulting of an individual caused by global sharing and sending messages of hurting, hostile, aggressive or threatening nature with Information and Communication Technology (ICT) infrastructure as a platform. Cyberbullying poses a significant challenge and the immeasurable threat to the physical and mental health of the victims.

Cyberbullying is known to be the cause of some serious health issues among social media users (Nixon, 2014; Olweus, 2012). It is, therefore, necessary to create a way to identify and detect this threat to prevent its reoccurrence (Rosa et al., 2018a; Chatzakou et al., 2017; Chavan & Shylaja, 2015). Cyberbullying is a form of online bullying with resultant negative impacts on victims (Salawu et al., 2017; Gahagan et al., 2016; Ptaszynski et al., 2016).

Facebook is a popular social media network that allows users to post different and unique comments and share these comments among various social platforms. On Facebook, over 2.32 billion monthly users are active on this platform as of December 2018 (Balakrishnan et al., 2020; Souza et al., 2018). There is an increase of 9% in Facebook monthly active users (MAUs) over the years (Zhao & Mao, 2016a; Zhang et al., 2016). There are more occurrences or instances of cyberbullying on Facebook and more social media platforms these days than before (Kowalski, Limber & McCord, 2019; Dredge, Gleeson & Garcia, 2014). This increase is as a result of the fact that cyberbullying has been found to be easier since the victims are bullied without direct contact or confrontation by the bullies. At the same time, they make use of communication devices like handsets, tablets, computer systems, etc. On the other hand, the traditional way of bullying is harder to practice as it sometimes involves physical contact between the parties and does not involve the use of a phone or computer system. The features of a social media network such as Facebook have enabled cyberbullies to expand their reach to different locations around the world that were not reachable before in different cities and countries (Veiga-Simao, et al., 2018; Facebook, 2015; Dadvar et al., 2012).

After registering on the network, users can create their customized profile displaying the information of their choice. Users can post images, comments, texts, use multimedia and share with other users known as friends on the network after accepting their friend requests. Users can use different types of embedded applications and be alerted on notifications about their friends' posts or activities. Users can also create and join groups of similar interest. Facebook has been involved in many controversies over the years such as cyberbullying and user privacy infringement which have had psychological effects on its users (Rosa et al., 2018b; Wright, 2017; Singh Huang, & Atrey, 2016). The organization has faced much pressure on censorship, and as a result of contents or feeds, users did not find satisfactory. Other issues include its addictive tendency to users, excessive retention of user information and facial recognition software.

Facebook has also been condemned for enabling users to release illegal, offensive or harmful materials to the public. Some of these posts violate copyright laws and constitute an infringement to intellectual property. Other issues with Facebook include hate speeches, rape incitements, acts of terrorism, crimes, fake news, live streams of violent incidents and killings (Nahar, Li, & Pang, 2014). Some statistics indicate that boys of age 19 who are users of Facebook are the most common victims of cyberbullying (Gahagan et al., 2016). According to the report, 49% of victims were bullied offline, while 65% of teenagers were harassed and victimized online, and only 37% of victims reported to the social network when it happened. Thus, the issue of cyberbullying has played a significant impact on online social network and society at large.

The size and velocity of the sharing technique allow unfriendly comments or embarrassing images to go online and viral within a few seconds. Several approaches have been adopted to prevent cyberbullying. The popular measures are: filtering unwanted messages from internet users whose profile pictures could not be uploaded or satisfactorily verified and activating a time-out feature that prevents users using rude and offensive language. Despite these efforts, the cyberspace is not adequately protected from cyberbullying.

Against these backdrops, cyberbullying, therefore, requires urgent attention in order to curb it and avoid a situation where it gets out of hand in terms of negative impact on social network users. There were some effects of cyberbullying discovered which include; depression, lack of sleep, substance and drug abuse, suicide or mental breakdown, harm on oneself (Facebook, 2015).

Detecting cyberbullying is very valuable because it assists in identifying and classifying cyberbullying activities, decreases the problem after it has been identified and helps internet users to take action before becoming a victim of cyberbullying.

This study aims to identify and detect cyberbullying from Facebook posts using machine learning algorithms. The objectives of the study to include a selection of unique features from a Facebook dataset from which the relevant features and samples will be derived and used as input to the machine learning algorithms. The paper also demonstrated a comparison of the performance of two machine learning classifiers to pick the most suitable for detecting cyberbullying-related posts and determine which of the algorithms can detect cyberbullying posts with a high degree of accuracy and efficiency. The paper further provides satisfactory suggestions and usable recommendations regarding how the result findings can be utilized to mitigate cyberbullying.

The paper has five major sections. In section 2, a detailed review of related and relevant literature is presented while in paragraph 3, the proposed method and tools for the study are presented. Implementations and evaluations are presented in section 4, while section 5 presents the conclusion and recommendations made for future work.

2. LITERATURE REVIEW

In this section, a review of related journals, conference proceedings and other documents is presented. Several approaches used in the detection and identification of cyberbullying in Facebook using machine learning algorithms are discussed.

Ducharme (2017) employed Support Vector Machine (SVM) and N-grams in the automated identification of cyber-harassment and cyberbullying. The study utilized these algorithms and was able to classify comments on Youtube with an overall accuracy of 81.8%. It later increased to 83.9% when the misclassified comments were added to the training set, and retraining was carried out. In achieving this feat, a-350 comment balanced training set with 7% high entropy, 3 length n-grams and a polynomial with C- error factor of 1, degree 2 and co-efficient of 1 were used in the LibSVM implementation of the SVM algorithm. K-Nearest algorithm was also used in trimming the comments where k was taken to be 4% of the training set size. The algorithm is a multi-threading algorithm and can be run on multiple servers, and the system efficiently calculated the accuracy by classifying three comments per second.

Amarasinghe et al. (2018) studied the integration of machine learning algorithms in a system for the detection of fraud. Their study employed machine learning algorithms such as SVM, Naïve Bayes, Hidden Markov Model (HMM), Fuzzy Logic, K-Nearest Neighbor, and K- Means clustering to achieve this purpose. Most existing fraud detection approaches focus on discrete data points (IP addresses, user accounts). Nowadays, an analytical approach is used to address the limitations and inefficiency of the existing traditional methodologies. The pros and cons of these algorithms are discussed thoroughly, and it is noticed that some of them are good, and some had drawbacks.

Also, accuracy alone is not a good metric to compare performance, especially where the dataset appears imbalanced. More metrics like recall, precision may also be necessary for the evaluations.

Kowalski et al. (2019) proposed a scalable approach by using Twitter to detect aggressive and cyberbullying demeanors. They adopted Naïve Bayes, Decision tree, Neural Networks and Random forests for classification by obtaining text, network and user-based features, after that, learning the characteristics of cyberbullies and the features that differentiate them from normal users. The study made use of 1.6 million tweets from Twitter posted over 3 months and showed how machine algorithms could be employed to accurately (over 90% AUC) classify the tweets. It was observed that random

forest classifier could differentiate effectively well between non-cyberbullies and cyberbullies with an accuracy of 91%.

The work of Van et al. (2018) focused on the automatic detection of cyberbullying using social media text by modeling posts written by bullies, victims, and spectators of online bullying. The system was evaluated on a manual annotated cyberbullying collection of writings for English and Dutch. It showed that their approach is applicable to different languages as long as the required data is available and usable. A set of two classification experiments were performed to how possible it is to detect cyberbullying automatically on social media. Following the optimization of features and hyperparameters of the models, a maximum score for F1 of 64.32% and 58.72% were reported for English and Dutch respectively. The classification algorithms, therefore, greatly surpassed the performance of the keyword and unoptimized N-gram baseline. The drawback of this research was that a qualitative breakdown of the outputs showed the presence of some false positives, an indication that cyberbullying or offences through irony.

Sugandhi et al. (2016) presented a survey of techniques for detecting cyberbullying. In their work, they discussed some of the most common forms of cyberbullying such as harassment, flaming(Heated online arguments and fights using vulgar languages), denigration(whereby the secrets of a person are exposed with the intention of destroying his/her image or reputation), impersonation, trickery and last but not the least, interactive gaming. The authors also talked about the unavailability of existing datasets that have prevented more studies on cyberbullying detection and their unreliable and inaccurate results on the studies currently done on them. The major drawback of this paper is the presence of few or no labelled datasets that future researchers can work on instead of gathering new datasets.

Nandakumar et al. (2018) worked on cyberbullying detection in email application using the Naïve Bayes classification algorithm. The system involved the identification and filtering of spams in emails; then applying the Naïve Bayes classification algorithm to classify the denoised messages. The feature extraction method is first applied to the messages. The paper also revealed that the Naïve Bayes classification algorithm and SVM were plotted and the efficiency factor was compared among the two classifier algorithms. Modules for the suggested system include GUI designing, dataset training, classification and analysis of Twitter messages for the occurrence of spam content and the classification technique used was Naïve Bayes Classifier algorithm. This paper also identified email-based cyberstalking as a big issue, and it involves two phases; the first is to identify and detect cyberstalking emails and the second is substantiate the proof for finding out cyberstalkers as a detection mechanism. The main method for cyberbullying revelation is web-based mining technologies. The output of the suggested system is promising, and a decent level of precision can be obtained through the system. For future works, the proposed system can be altered and improved for cyberbullying revelation in non-English applications.

Haidar et al. (2017) focused their research on the mitigation and detection of cyberbullying by developing a multilingual system. They covered cyberbullying detection in Arabic language content through the utilization of the machine learning approach. The system made use of a dataset (Twitter and Facebook) which was prepared for testing and training the system. The two machine learning classifiers used were Naïve Bayes and Support Vector Machine (SVM) algorithms. When the system was trained using Naïve Bayes algorithm, the results obtained show a higher degree of accuracy when compared to SVM as a significant amount of cyberbullying instances were detected, an indication that cyberbullying in Arabic can also be detected.

The drawback of this research is that the output achieved by this system is not perfect when compared to cyberbullying detection in the English language even though it met its aim. In the results obtained by using the Naïve Bayes algorithm, it showed that at a minimum one-third of cyberbullying is detectable in Arabic language using the system. Gomez-Adorno et al. (2018) focused on detecting aggressive tweets in Spanish on twitter using machine learning approaches. The researchers collated tweets, of which 75% was non-aggressive, and 25% was aggressive. After training, the distribution

was 35.42% aggressive tweets and 64.58% non-aggressive tweets; this analysis proved that the number of aggressive tweets was half the non-aggressive tweets.

The algorithms adopted were the logistic regression, SVM and multinomial Naïve Bayes. It was discovered that the logistic regression showed better results and performance than the SVM and multinomial Naïve Bayes algorithms. The work of Mangaonkar et al. (2015) has to do with the collaborative detection of cyberbullying behavior in Twitter data. This research is essential in detecting cyberbullying behavior accurately by analyzing Twitter tweets in real-time as much as possible. This research introduces a new method known as the distributed collaborative approach for detecting cyberbullying. The approach comprises of a network of nodes for detection and is also capable of classifying tweets supplied to it. The nodes work together in situations where they need assistance in classifying a tweet. The research evaluates different patterns of collaboration and measures the performance of each pattern to detail.

The results obtained represent an improvement in precision and recall values of the method employed for detection when compared with the stand-alone method. This research further measures the scalability of the process by adding to the number of network nodes.

Nandihini and Sheeba (2015) aimed at detecting and classifying cyberbullying using information retrieval algorithm. They proposed a system for taking note of and classifying cyberbullying activities such as racism, harassment, terrorism, and flaming in social media networks. The authors adopted Naïve Bayes classifier for classifying the cyberbullying activity and Lavenshte algorithm for cyberbullying detection. The following parameters, namely; F-measure, accuracy, recall, and precision, were used in the evaluation of the system. The mean accuracy gotten from fornspring.me showed 93.79% accuracy and myspace.com showed 94.59% accuracy. One major drawback of this project is the inability of it to be used in detecting and identifying cyberbullying activities in other social network systems.

Zhao, Zhou, and Mao (2016b) proposed a system that detected cyberbullying automatically on social networks using bullying features. Bullying contents from the dataset were detected by the use of machine learning and natural language processing techniques. The study employed the embedding-enhanced Bag-of-Words (eBoW) learning method that focused on word embedding and expansion of a collection of pre-defined derogatory words. Different weights were assigned in order to arrive at bullying features which are then concatenated with BoWs and semantic latent features to produce the final representation. The outcome was then inputted into a linear Support Vector Machine classifier. The experiment was conducted on Twitter tweets, and the procedure was compared with several existing and related models. Five-fold cross-validation technique was employed in which the dataset was partitioned into five parts comprising of four parts for training and one part for testing.

Sanchez-Medina et al., in 2020, attempted to carry out similar research to enhance knowledge and skills concerning cyberbullying that is related to sexual behaviour and personality. They proposed a very vital tool to navigate through this scenario. The adopted methodology focused mainly on the application of ensemble classification tree and structural equation modelling to evaluate and analyze how traits, specifically, those associated with the Dark Triad can impact this attitude and behaviour. The results obtained revealed that high levels of Machiavellianism and psychopathy are very likely to be linked to cyberbullying that is sexual-related.

Balakrishnan et al. (2019), developed a model for detecting cyberbullying on the user's personality, which was determined by the Dark Triad models and Big Five., in 2019. The sole aim of the model was to identify bullying styles and patterns among Twitter users. They further adopted Random Forest for classifying cyberbullying. The results obtained revealed that consideration of the user's personality could significantly enhance the detection of cyberbullying.

Because discovering cyberbullying "hotspots" on the global network is very crucial for safeguarding against cyber-victimization, Ho, et al. (2020), attempted to come up with a unique prediction model for cyberbullying "hotspots" identification. This was achieved by studying, analyzing and investigating charged and emotional languages on Twitter. The results show that specific charged

and emotional languages in tweets can signify a very high potential for cyberbullying cases (Ho, et al. 2020).

Automatic detection of cyberbullying on Twitters using features such as emotion, sentiment and personalities were carried out by Balakrihnan et al., 2020. The authors used Dark Triad models and Big Five to determine user's personalities. They further utilized machine learning technique to categorize the tweets into four sections. Finally, the results revealed that there was an improvement in cyberbullying detection when sentiment and personalities were considered. A different case was, however, observed for emotion. A real-world corpus was used to verify the effectiveness of the model experimentally. It was also observed that semantic BoW could produce a slightly better performance than ordinary BoW. Four performance metrics, namely F-measure, accuracy, recall, and precision, were used in evaluating the system. The results of the experiments conducted showed that the system has a 76.8% precision value, 79.4% recall value and 78% F1-measure score.

A summary of the reviewed articles is presented in Table 1.

3. METHODOLOGY

The entire process for cyberbullying detection and identification could be achieved by using the six stages of architecture, as revealed in Figure 1. The stages are data gathering, data splitting, feature extraction, a matrix of token counts and weights, classification, and identification and classification.

3.1. An Architectural Model for Cyberbullying Identification and Detection

This method tends to classify Facebook comments or posts as cyberbullying or non-cyber bullying. The implementation of the Cyberbullying identification and detection system on Facebook is based on a system that uses features that are derived from the Facebook dataset(a sample of the dataset is given in Figure 2) by applying two machine learning algorithms (Naives Bayes and K-Nearest Neighbor).

3.2 Data Gathering

The trained data was gotten from Joshmiloni Github on Cyber Bullying Detection. The csv data was uploaded on the Github early May 2018. Github is a git repository hosting service with different features that enables programmers and software developers to store and access open-source projects and files. It also allows modification and download of the same files and projects. The Facebook comments were obtained using Facebook API service. The data set contains over 8,818 labeled Facebook comments and was provided in the form of text document for each comment. These comments were extracted from Facebook publicly available content. Each comment contains labels that indicate if a comment is a cyberbullying comment or not.

3.2.1. Labelling the Data

The dataset used in the study was labelled, and efforts were made to extract the question text and the answer text from a randomly chosen subset.

The following questions were asked to aid the labelling process:

1. Does the current post contain cyberbullying (Yes / No)?
2. How terrible or bad is the cyberbullying in this post (enter 0 for no cyberbullying)? On a scale of 1 (mild) to 10 (severe)
3. What phrases or words in the post(s) are indicative of the cyberbullying?
4. Please enter any additional information you would like to share this post.

With this questions approach, the dataset was labelled within a few hours.

Table 1. Summary of the reviewed articles

AUTHOR	YEAR	APPROACH	STRENGTHS	WEAKNESSES
Balakrishnan et al.	2020	Dark Triad models and Big Five to determine user's personalities with machine learning technique	improvement in cyberbullying detection when sentiment and personalities	A few features were considered. It is not verified on other social media platforms apart from Twitter.
Ho, et al.	2020	prediction model for cyberbullying "hotspots" identification	It incorporates charged and emotional languages on Twitter	It is applicable only on Twitter
Balakrishnan et al.	2019	They adopted Dark Triad models and Big Five as well as Random Forest for classifying cyberbullying	The approach enhances the detection of cyberbullying.	It doesn't incorporate other cases of detecting cyberbullying.
Sanchez-Medina et al.	2020	application of ensemble classification tree and structural equation modelling	It is very useful in the analysis of traits and how they can impact on attitude and behavior for detecting cyberbullying	Its narrow methodology has undoubtedly limited its application
Nandhini & Sheeba	2016	Naïve Bayes classifier for cyberbullying classification and Lavenstein distance algorithm for cyberbullying detection	Showed high accuracy of over 90% for cyberbullying detection and classification, an efficient approach	Cannot be used for other social network platforms, restricted to only formspring.me and myspace.com
Chatzakou et al	2017	Detection of aggressive and cyberbullying demeanor on twitter with the use of Machine learning Classification Algorithms (classifiers like Naïve Bayes, Decision tree, Random forests).	It gives room for scalability.	Not an effective method in curbing cyberbullying accounts.
Ducharme	2017	Support Vector Machine Model & N-grams in the automated identification of cyberbullying & cyberharassment on youtube.	Ability to classify comments on Youtube with a high degree of accuracy	Presence of non-cyberbullying comments classified as cyberbullying.
Haidar, Charmoun, & Serhrouchni.	2017	Multilingual system for cyberbullying detection(Arabic language)	Ability to detect cyberbullying in Arabic language.	Result obtained was not as perfect as that of detection in the English language.
Mangaonkar	2017	Collaborative detection of cyberbullying behavior in twitter data	Results show an improvement in recall and precision of the mechanism for detection over the stand-alone paradigm.	Inability to examine the possibility of picking a node depending on previous tweets suggestions
Shrivastava	2017	Usage of individual topic sensitive classifiers. Usage of machine learning classifier algorithms. Usage of time series modeling for detection of cyberbullying	It is very efficient and reliable.	Limited word-set of negative words reduces efficiency in cyberbullying detection. Presence of few or no labeled datasets
Zhao, Zhou, & Mao,	2017	Representation learning framework specific to cyberbullying detection. The learning method is named Embedding-enhanced Bag-Of-Words	The approach here was able to achieve a significant performance improvement compared to sBoW over all three evaluation measures.	The parameters evaluated in this model are not of high accuracy; therefore, making it difficult to correctly detect cyberbullying efficiently.
Amarasinghe, Aponso, & Krishnarajah	2018	Use of Machine learning algorithms (Naïve Bayes, Support Vector Machine, KNN Nearest Neighbor, e.t.c) for fraud detection in financial transactions.	Low complexity, robust and simple to understand, a requirement of little data preparation, ability to be used for real-time fraud detection, automatic decision and training processes.	Identification of inefficient results by the classification algorithms. it did not perform well for fraud detection because normal and anomalous classes are imbalanced.
Gomez-Adorno et. Al.	2018	Logistic regression algorithm(detecting aggressive tweets in Spanish)	Use of SMOTE to solve un-evened data which produced better output in the corpus.	Inability to solve un-evened data problem with deep analysis of SMOTE process. No utilization of linguistic patterns
Nandakumar, V., Kovoov, B. C., & Sreeja, U. M.	2018	Identification and filtration of spams in emails using Naïve Bayes Classifier Algorithm.	It Identified email-based cyberstalking.	Inability to apply it to non – English applications.
Van Hee, et al	2018	Automatic detection of cyberbullying in social media text by modelling posts written by bullies, victims and spectators of bullying online	A promising method for the detection of cyberbullying signals on social media automatically.	Difficulty in identifying victims of cyberbullying.

Figure 1. A System Architecture for Cyberbullying Identification and Detection

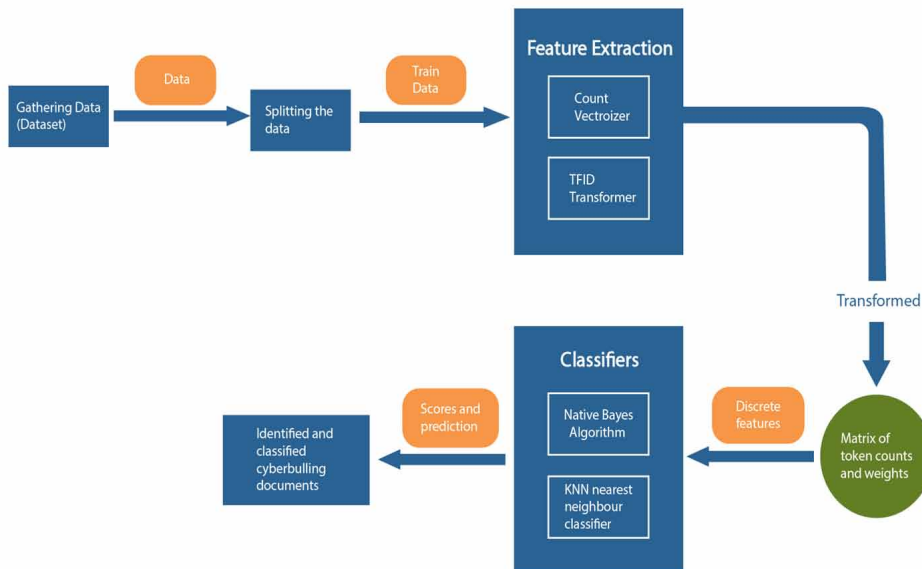


Figure 2. Sample of Dataset

label_bullying	comments
0	yeah I got 2 backups for all that. I just hate when that happen. I been strugglin for a week now...handle that tho
0	I hate using my BB but love my iPhone. Haven't tried the new BB. My BB is provided by my corp. I don't get to pick which model
1	Get fucking real dude.
1	She is as dirty as they come and that crook Rengel the Dems are so fucking corrupt it's a joke. Make Republicans look like ...
1	why did you fuck it up. I could do it all day too. Let's do it when you have an hour. Ping me later to sched writing a book here.
1	Dude they dont finish enclosing the fucking showers. I hate half assed jobs. Whats the reasononing behind it? Makes no sense.
1	WTF are you talking about Men? No men thats not a menage that's just gay.
1	Ill save you the trouble sister. Here comes a big ol fuck France block coming your way here on the twitter.
1	Im dead serious.Real athletes never cheat don't even have the appearance of at his level. Fuck him dude seriously I think he did
0	wow lol sounds like a lot of piss then hehehe
0	not a damn thang..the typical rap beef. one person worrying about what the next is doing and the other respondin etc etc
1	...go absolutely insane.hate to be the bearer of bad news..LoL..dont shoot the messenger (cause we all know you bought that pistol
0	well damn!! where have you been when i have needed you 'Mother Time'
0	watching without a trace too...hate when i miss the 1st 5 minutes when the person disappears!!!!
0	which they do most of the time:-P I don't hate them either I wanted to be one once..but I figured out I couldn't live w/ myself
1	Lmao im watching the same thing ahaha. The gay guy is hilarious! "Dede having a good day and I dont want anyone to mess it up."
1	LOL no he said What do you call a jail cell to a gay guy? Paradise! ahaha.
1	truth on both counts that guy is an ass and their product is sub par. I tell people try Dalesandros orJim's
1	Shakespeare nerd!
1	you are SUCH a fucking dork
1	Heh. Fuck 'em WHERE?!?
1	damn it i totally forgot that one!
0	haha fuck i wish i was there :(
0	paranoid is wack as fuck...the best song on 80s is DEFINITELY Bad News
1	wow damn I would have been pissed @ that...
1	nigga u geigh lmao! fuck yo finals beeeeeeitch
0	in London I hate thee :(ENJOY YOURSELVES!!!! <33333
0	lolipop lolipop...oh loli loli loli...duh duh dum dum dum *Pop* Damn im so bad at cheeering people up <3
1	that sucks :(
1	read that this morning my fuck how they just straight up say "fuck chet"

An attempt was made to use this information to train the machine learning algorithms in detecting the F1 score, precision, accuracy, and recall of cyberbullying in the comments.

The rationale for the use of the dataset in the study is premised on the distinctive and unique features obtained from Facebook showing most well documented, large-scale instances of bullying, harassment and aggressive behaviours. What is more, the objective is also to create a dataset with hundreds of millions of chat messages over many years which are significantly well related to the issues of this research.

3.3 Splitting of the Data

With Scikit learn library on Jupyter Notebook, the data was split using a function called ‘train_test_split’ into a training and testing sets. The training set consists of the Comments data, which is called ‘X_train’ and the label data, which is called ‘y_train’. These data are fitted into the Machine learning classifiers so that they learn how to predict the detection of cyberbullying. The data is split into a training set of 70% per cent and a test set of 30%. The test set consists of the Comments data, which is called ‘X_test’ and the label data, which is called ‘y_test’.

3.4 Data Reprocessing and Feature Extraction

For feature extraction, a Scikit learn feature extraction function which is Count Vectorizer was used. Count vectorizer is a feature extraction function that converts a package or collection of text documents into a matrix of token counts. It takes each unique word in the document and counts it based on the number of times it occurs in a document which is a row in a data. It represents the features in a vectorized form (well-arranged format). It selects unique words in the entire document and the number of times the word occurs. A feature selection approach was used: test (CHI-SQUARE TEST) to determine the discriminative power of each feature and also improve the performance of each classifier algorithm.

3.5 Feature Selection Algorithm

This is an essential part of building machine learning models because training the models with insignificant features will affect the performance of the classifiers.-

3.5.1. Chi – Square Test (X^2 Test)

This feature selection algorithm is used with text data to test for dependence between two events. It is specifically used to test if the occurrence of a specific term (feature) represented as frequencies and the occurrence of a specific class (labels) are dependent. Therefore the use of this feature selection algorithm is to remove the features that do not correlate with the class and not useful for classification. In this work, a feature selection function that was used is called SelectKBest. It ranks the features with the statistical test and selects the top k features (the features that are more important for classification) where k is a number that can be changed.

$$x^2(D, t, c) = \sum_{et \in} \{0,1\} \sum_{ec \in} \{0,1\} \frac{(Ne_t e_c - Ee_t e_c)^2}{Ee_t e_c} \quad (1)$$

Where N = observed frequency, E = expected frequency.

$E_t = 1$ if the document contains term t and $E_t = 0$ otherwise.

$E_c = 1$ if the document is in class c and $E_c = 0$ otherwise.

3.6 Data Training and Testing

The data was trained after splitting them by using two machine learning classifiers, namely Naïve Bayes algorithm and K-Nearest Neighbor algorithm. The features derived were used to detect cyberbullying on Facebook.-

3.6.1. Naïve Bayes Algorithm

Naïve Bayes algorithm is a machine learning algorithm that can be used to predict the likelihood that an event will occur provided evidence that is present in the data. A multinomial Naïve Bayes algorithm was used because it works better on features that describe discrete frequency counts which is similar to the features of the data in this work. The Naïve Bayes algorithm is as described in Equation ii.

$$P(y|x_1, x_2, x_3, \dots, x_n) = \frac{P(y) P(x_1, x_2, x_3, \dots, x_n | y)}{P(x_1, x_2, x_3, \dots, x_n)} \quad (2)$$

Where $P(y)$ = labels

$P(x_1, x_2, x_3, \dots, x_n)$ = Comments

$P(y|x_1, x_2, x_3, \dots, x_n)$ = the probability of hypothesis(labels) given the observed evidence

$P(x_1, x_2, x_3, \dots, x_n|y)$ = the probability of evidence given that hypothesis is true

$P(y)$ = the probability of hypothesis before observation

$P(x_1, x_2, x_3, \dots, x_n)$ = the probability of new evidence given all possible hypotheses

3.6.2. K-Nearest Neighbor Algorithm (K-NN)

K-NN is an example of a supervised machine learning algorithm. This algorithm makes use of similarity in observations or samples to make predictions for new ones. The assumption is that the more similar samples are, the more likely they belong to the same category or class. The K in the K-NN represents the number of nearest neighbors for which the decision of whether or not the new sample belongs to the same class. In this work, there are two main categories for which group of neighboring samples can be classified, namely: Cyberbullying or non-cyber bullying. Usually, a distance function is applied to determine K nearest neighbors of a new sample or observation $K=1$ is the simplest of the cases of the K-NN algorithm in which the sample is simply assigned to the class or category of the single nearest neighbor. To get the best K, a method known as the elbow method was employed to determine the K with the lowest error rate, and this was achieved by calculating the mean of errors obtained in the model's prediction when compared to the labeled test data as indicated in Equation iii.

$$\text{Distance Function (Euclidean distance)} = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (3)$$

This implies that the square root of the sum of the difference between the new point x and an existing point y.

3.7 Evaluation Metrics

The metrics for the evaluation are:

- i. **True Positive Rate (TP):** This refers to the ratio of the documents or samples that are cyberbullying and are actually classified as such.
- ii. **True Negative Rate (TN):** This refers to the ratio of documents that are not cyberbullying and are actually classified as such.

- iii. **False Positive Rate (FP):** This refers to the ratio of non-cyber bullying documents that are wrongly classified as cyberbullying.
- iv. **False Negative Rate (FN):** This refers to the ratio of cyberbullying documents that are wrongly classified as non-cyber bullying
- v. **Accuracy (Acc):** This is the measure of overall percentage or ratio of classified documents in relation to the sum of the actual or correctly classified cyberbullying and non-cyber bullying documents. It is denoted as:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

where TP, TN, FP, FN are as already described.

Precision: This is the ability of the classifier to correctly label a sample according to its correct class. In this case, it is the ability of the classifiers to correctly label the cyberbullying sample as cyberbullying and not something different. It is denoted as:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

TP = True Positive value, FP = False Positive.

Recall: This is the ability of the classification algorithm to correctly locate all the positive samples in the pool of samples. In this case, it is the ability of the classifiers to find all cyberbullying comments in the pool of data. It is denoted as:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Where TP = True Positives and FN = False Negatives for cyberbullying

F1 Score: The F1 can be referred to as the weighted average of precision and recall values. The best F1 score is 1, while the worst is 0. It is denoted as:

$$\text{F1 score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}).$$

4. RESULTS AND EVALUATION

To evaluate the effectiveness of the methodologies used in this research concerning the identification and detection of cyberbullying in Facebook, different experiments were carried out on the data by applying machine learning algorithms and the results obtained were promising, efficient and reliable. Four metrics were used for evaluation. The metrics are F1-score, accuracy, recall, and precision. The execution time and confusion matrix were also used to measure the performance of these classifiers.

4.1 Analysis of Results

4.1.1. Results Obtained When using Naïve Bayes

0 = Non-Cyberbullying

Table 2. Results obtained when using Naïve Bayes

CLASSIFIERS	PRECISION	RECALL	F1 - SCORE
0	0.72	1.00	0.84
1	0.57	0.01	0.01

1 = Cyberbullying
Accuracy score = 0.72

From Table 2, it was observed that with Naïve Bayes classifier, there's an accuracy of 0.72, which implies that the model correctly predicted 72% of the comments labeled as cyberbullying or non-cyber bullying. The model also exhibits a precision score of 0.72 for non-cyber bullying indicating that 72% of those comments the model predicted as non-cyber bullying are non-cyberbullying, and 0.57 for cyberbullying indicating that 57% of those comments the model predicted as cyberbullying are even cyberbullying. The model also exhibits a recall score of 1.00, and this shows that the model was able to discover 100% of the comments the model predicted as non-cyber bullying in the pool of comments. Also, 0.01 implies that the model is able to discover 1% of the comments the model predicted as cyberbullying in the pool of comments. The model also shows a F1-score of 0.84 (84%) as the weighted average of the precision and recall for non-cyber bullying, and 0.01(1%) as the s the weighted average of the precision and recall for cyberbullying. Confusion Matrix of the classification result for Naïve Bayes is demonstrated in Figure 3. Naïve Bayes Confusion Matrix Bar Graph is also shown in Figure 4.

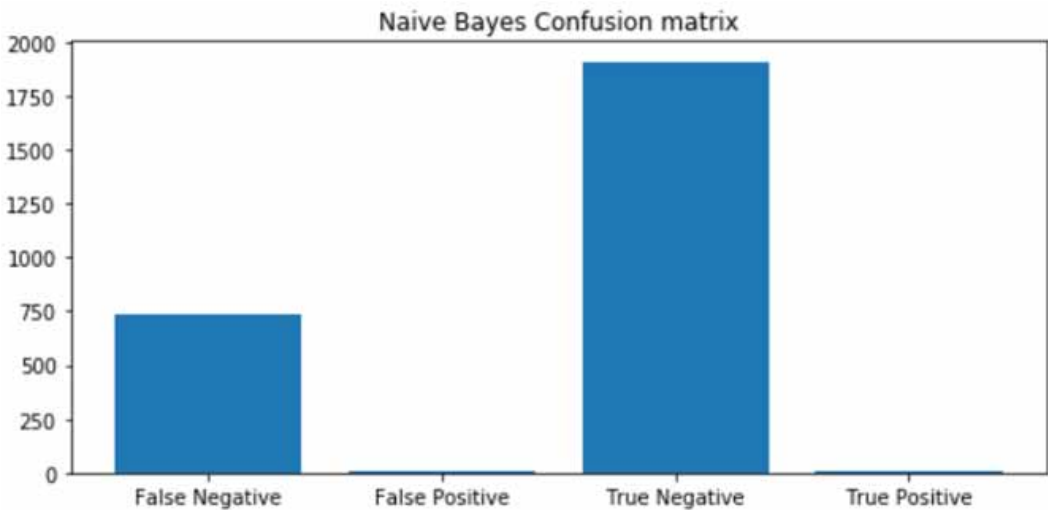
From the confusion matrix, the following observations were recorded:

- i. True Positive (TP) = 4 (4 out of 735 documents)
- ii. True Negative (TN) = 1908 (1908 out of 2646 documents)
- iii. False Positive (FP) = 3 (3 out of 2646 documents)
- iv. False Negative (FN) = 731 (731 out of 735 documents)

Figure 3. Confusion Matrix of the classification result for Naïve Bayes

$$\begin{bmatrix} [1908 & 3] \\ [731 & 4] \end{bmatrix}$$

Figure 4. Naïve Bayes Confusion Matrix Bar Graph



$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 4/(4+3) = 0.57$
 $\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) = 4/(4+731) = 0.01$
 $\text{F1-Score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$
 $= 2 * (0.57*0.01)/(0.57+0.01)$
 $= 0.01$
 Accuracy score
 $= (\text{TP}+\text{TN}) / (\text{TP}+\text{FP}+\text{FN}+\text{TN})$
 $= (4+1908) / (4+3+731+1908)$
 $= 0.72$
 Execution time for learning the model for Naive Bayes classifier is: 0.006987s

ii) K-Nearest Neighbor

Table 3. Results obtained when using K-Nearest Neighbor Classifier

CLASSIFIERS	PRECISION	RECALL	F1 - SCORE
0	0.73	0.98	0.84
1	0.56	0.07	0.12

From Table 3, it was observed that with the K-Nearest Neighbor classifier, there is an accuracy of 0.73. This implies that the model correctly predicted 73% of the comments labeled as cyberbullying or non-cyber bullying. The model also exhibits a precision score of 0.73 for non-cyber bullying which implies that 73% of those comments the model predicted as non-cyber bullying are actually non-cyber bullying, and 0.56 for cyberbullying indicating that 56% of those comments the model predicted as cyberbullying are actually cyberbullying.

The model also exhibits a recall score of 0.98. This indicates that the model was able to find 98% of the comments the model predicted as non-cyber bullying in the pool of comments. Also, 0.07 indicates that the model was able to find 7% of the comments the model predicted as cyberbullying in the pool of comments. The model also produced F1-score of 0.84 (84%) as the weighted average of the precision and recall for non-cyber bullying and 0.12(12%) as the weighted average of the precision and recall for cyberbullying. Confusion Matrix of the classification result for KNN is shown in Figure 5. KNN Confusion Matrix Bar Graph is also shown in Figure 6.

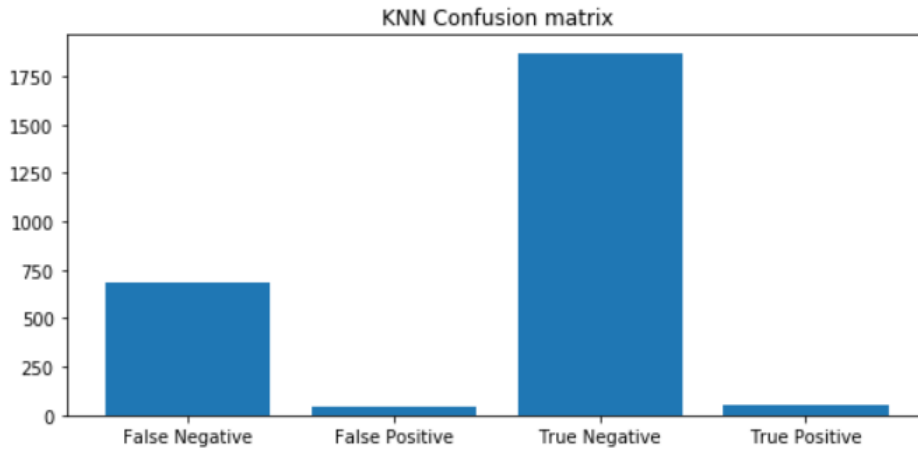
From the confusion matrix, the following observations were recorded:

Figure 5. Confusion Matrix of the classification result for KNN

$$\begin{bmatrix} 1871 & 40 \\ 685 & 50 \end{bmatrix}$$

- i. TP = 50 (50 out of 735 documents)
- ii. TN = 1871 (1871 out of 2646 documents)
- iii. FP = 403 (403 out of 2646 documents)
- iv. FN = 685 (685 out of 735 documents)

Figure 6. KNN Confusion matrix Bar graph



$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 50 / (50 + 40) = 0.56$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) = 50 / (50 + 685) = 0.07$$

$$\begin{aligned} \text{F1-Score} &= 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) \\ &= 2 * (0.56 * 0.07) / (0.56 + 0.07) \\ &= 0.12 \end{aligned}$$

$$\begin{aligned} \text{Accuracy} &= (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN}) \\ &= (50 + 1871) / (50 + 40 + 685 + 1871) \\ &= 0.73 \end{aligned}$$

Execution time for learning the model for K-Nearest Neighbor is classifier is: 0.004247s

Table 4. Classification result of Naive Bayes Algorithm Classifier with Chi-Square Test

CLASSIFIERS	PRECISION	RECALL	F1 - SCORE
0	0.73	1.00	0.84
1	0.79	0.03	0.06

4.2. Results Obtained by Using Machine Learning Classifiers with Chi-Square Test

4.2.1. Naïve Bayes Algorithm

With the use of Naïve Bayes classifier, there is an accuracy of 0.73, which implies that the model correctly predicted 73% of the comments labeled as cyberbullying or non-cyber bullying. This is shown in Table 4. The model also exhibits a precision score of 0.73 for non-cyber bullying which indicates that 73% of those comments the model predicted as non-cyber bullying are actually non-cyber bullying, and 0.79 for cyberbullying indicating that 79% of those comments the model predicted as cyberbullying are actually cyberbullying.

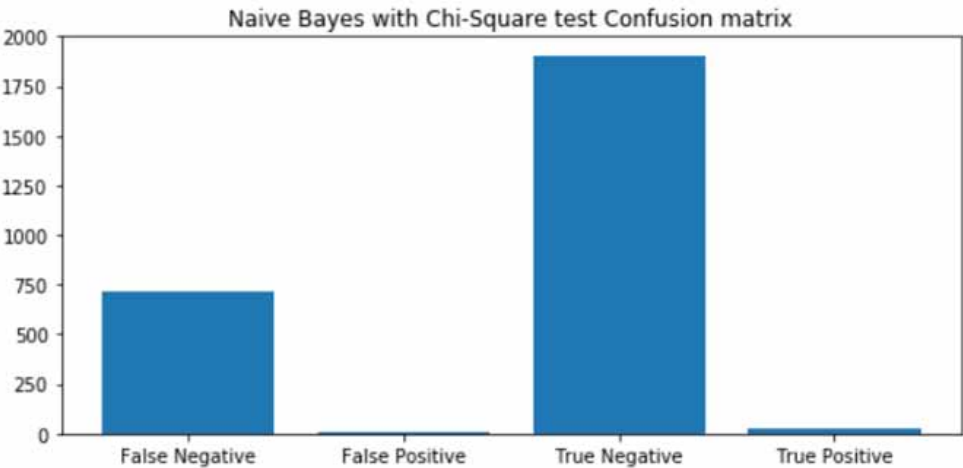
The model also exhibits a recall score of 1.00. This simply implies that the model was able to find 100% of the comments the model predicted as non-cyber bullying in the pool of comments. Also, 0.03 depicts that the model was able to find 3% of the comments the model predicted as cyberbullying in the pool of comments. The model also exhibited a F1-score of 0.84 (84%) as the weighted average of the precision and recall for non-cyberbullying and 0.06 (6%) as the weighted average of the precision and recall for cyberbullying. Confusion Matrix of the classification result for Naïve Bayes with Chi-Square Test is demonstrated in Figure 7. Naïve Bayes with Chi-Square

Figure 7. Confusion Matrix of the classification result for Naïve Bayes with Chi-Square Test

[[1905		6]
	[713		22]]]

Bar Graph is also shown in Figure 8.

Figure 8. Naïve Bayes with Chi-Square test



From the confusion matrix, the following observations were recorded:

- i. TP = 22 (22 out of 735 documents)
- ii. TN = 1905 (1905 out of 2646 documents)
- iii. FP = 713 (713 out of 2646 documents)
- iv. FN = 6 (6 out of 735 documents)

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 22 / (22 + 713) = 0.79$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) = 22 / (22 + 6) = 0.03$$

$$\text{F1-Score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

$$= 2 * (0.79 * 0.03) / (0.79 + 0.03)$$

$$= 0.06$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

$$= (22 + 1905) / (22 + 713 + 6 + 1905)$$

$$= 0.73$$

Execution time for learning the model for Naive Bayes classifier with Chi Square Test is: 0.001995s

4.2.2. K-Nearest Neighbor Algorithm

Table 5. Classification result of K-Nearest Neighbor Classifier with Chi-Square

CLASSIFIERS	PRECISION	RECALL	F1 - SCORE
0	0.73	0.99	0.84
1	0.69	0.07	0.13

From Table 5, it was observed that, with the use of K-Nearest Neighbor classifier, there is an accuracy of 0.73 which implies that the model correctly predicted 73% of the comments labeled as cyberbullying or non-cyber bullying. The model also exhibits a precision score of 0.73 for non-cyberbullying indicating that 73% of those comments the model predicted as non-cyberbullying are actually non-cyber bullying, and 0.69 for cyberbullying indicating that 69% of those comments the model predicted as cyberbullying are actually cyberbullying. The model also exhibits a recall score of 0.99. This simply implies that the model was able to find 99% of the comments the model predicted as non-cyber bullying in the pool of comments. Also, 0.07% indicates that the model was able to find 7% of the comments the model predicted as cyberbullying in the pool of comments. The model also exhibited a F1-score of 0.84 (84%) as the weighted average of the precision and recall for non-cyberbullying and 0.13(13%) as the weighted average of the precision and recall for cyberbullying. Confusion Matrix of the classification result for KNN classifier with Chi-Square is shown in Figure 9. KNN with Chi-Square Bar Graph is also shown in Figure 10.

From the confusion matrix, the following observations were recorded:

- i. TP = 53 (53 out of 735 documents)
- ii. TN = 1887 (1887 out of 2646 documents)
- iii. FP = 24 (24 out of 2646 documents)
- iv. FN = 682 (682 out of 735 documents)

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) = 53 / (53 + 24) = 0.69$$

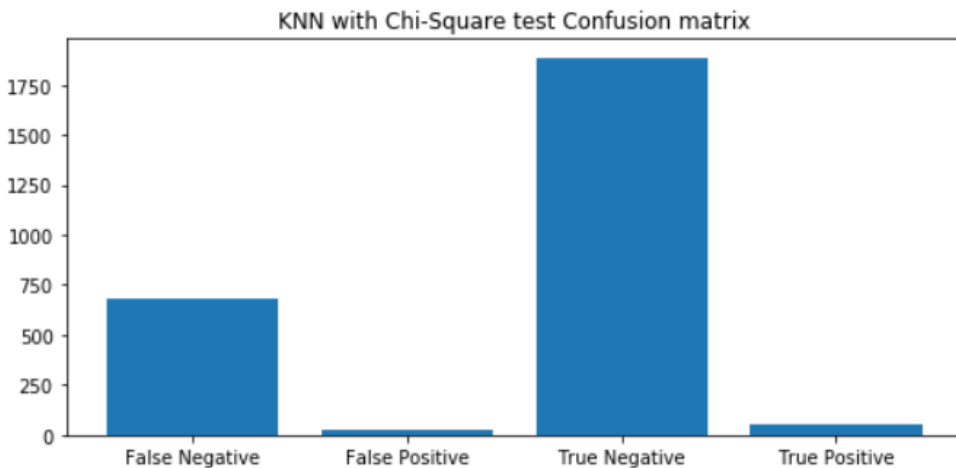
$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) = 53 / (53+682) = 0.07$$

$$\text{F1-Score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

Figure 9. Confusion Matrix of the classification result for K-Nearest Neighbor Classifier with Chi-Square

Confusion matr
[[1887 24]
[682 53]]

Figure 10. KNN with Chi Square test Confusion Matrix Bar Graph



$$= 2 * (0.69*0.07) / (0.69+0.07)$$

$$= 0.13$$

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{FN}+\text{TN})$$

$$= (53+1887) / (53+24+682+1887)$$

$$= 0.73$$

Execution time for learning the model for KNN classifier with Chi Square Test is: 0.001032

Using the classification report scores, execution time and confusion matrix of both machine learning classifiers, K-Nearest Neighbor Classifier provides an efficient and more effective way of detecting Cyberbullying in Facebook using this dataset.

5. CONCLUSION

Cyberbullying is becoming a thing of great concern in the cyberspace. People no longer use information and communication technology for normal communication only but also to harass and bully other users. This development has undoubtedly created an avenue to some internet users to intimidate and make mean and nasty comments, post derogative messages aimed at downgrading and also to belittle

internet users. Efforts have been made to utilize two machine learning algorithms (Naïve Bayes and K-Nearest Neighbor) to identify and detect cyberbullying on Facebook. These classifiers were tested to determine the most suitable amongst the two for detecting cyberbullying in the posts with high degree accuracy and precision.

Chi-Square test was used to find the most convenient and important features that improve the performance of the classifiers with reduced classification time. The K-Nearest Neighbor performed better over the Naïve Bayes classifier with much improvement in the performance metrics and classification time. Its proposed real-time implementation will assist in no small measure to protect the cyberspace from cyberbullying.

One of the main benefits of this work is in its ability to provide the report of cyberbullying after identifying features related to it to internet users. It will also avail users the opportunity to be completely protected from any comment identified as instances of cyberbullying. This approach is a reliable method to detect cyberbullying activities on Facebook as the detection method can identify the presence of cyberbullying terms, classify them accordingly and identify potential risks automatically. It provides reliable information to the internet users and creates awareness so that they can be guided from being cyberbullied.

REFERENCES

- Al-Garadi, M., Varathan, K. D., & Ravana, S. D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63, 433–443. .2016.05.051<ALIGNMENT.qj></ALIGNMENT>10.1016/j.chb
- Amarasinghe, T., Aponso, A., & Krishnarajah, N. (2018). Critical Analysis of Machine Learning Based Approaches for Fraud Detection in Financial Transactions. *Proceedings of the 2018 International Conference on Machine Learning Technologies - ICMLT '18*. doi:10.1145/3231884.3231894
- Balakrihnan, V., Khan, S., & Arabnia, H.R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, 90(2020), 1-11.
- Balakrishnan, V., Khan, S., & Arabnia, R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security*, 90, 1–11. doi:10.1016/j.cose.2019.101710
- Balakrishnan, V., Khan, S., Fernandez, T., & Arabnia, H. R. (2019). Cyberbullying detection on twitter using Big Five and Dark Triad features. *Personality and Individual Differences*, 141, 252–257.
- Chatzakou, D., Kourtellis, N., Blackburn, J., Cristofaro, E.D., Stringhini, G., & Vakali, A. (2017). *Mean Birds: Detecting Aggression and Bullying on Twitter*. ArXiv, abs/1702.06877.
- Chavan, V. S., & Shylaja, S. S. (2015, August). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In *Proceedings of the international conference on advances in computing, communications and informatics* (pp. 2354–2358). ICACCI. doi:10.1109/ICACCI.2015.7275970
- Dadvar, M., de Jong, F., Ordelman, R. J. F., & Trieschnigg, R. B. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the 12th Dutch-Belgian information retrieval workshop* (pp. 23-25). Retrieved from <https://eprints.eemcs.utwente.nl/21608/>
- Dadvar, M., Trieschnigg, D., Ordelman, R., & de Jong, F. (2013). Improving cyberbullying detection with user context. In *Proceedings of the European conference on information retrieval* (pp. 693–696). Berlin: Springer. doi:10.1007/978-3-642-36973-5_62
- Dredge, R., Gleeson, J., & Garcia, X. (2014). Presentation of Facebook and risk of cyberbullying victimization. *Computers in Human Behavior*, 40, 16–22. doi:10.1016/j.chb.2014.07.035
- Ducharme, D. N. (2017). *Machine Learning for the Automated Identification of Cyberbullying and Cyberharassment*. Academic Press.
- Facebook. (2015). *Facebook reports second quarter 2015 results*. Retrieved from <https://investor.fb.com/releasedetail.cfm?ReleaseID=924562>
- Fridh, M., Lindström, M., & Rosvall, M. (2015). Subjective health complaints in adolescent victims of cyber harassment: Moderation through support from parents/friends - a Swedish population-based study. *BMC Public Health*, 15(1), 949. Advance online publication. doi:10.1186/s12889-015-2239-7 PMID:26399422
- Gahagan, K., Vaterlaus, M. J., & Frost, R. L. (2016). College student cyberbullying on social networking sites: Conceptualization, prevalence, and perceived bystander responsibility. *Computers in Human Behavior*, 55, 1097–1105. doi:10.1016/j.chb.2015.11.019
- Gomez-Adorno, H., Bel-Enguix, G., Sierra, G., Sanchez, O., & Quezada, D. (2018). *A Machine Learning Approach for Detecting Aggressive Tweets in Spanish*. Academic Press.
- Haidar, B., Charmoun, M., & Serhrouchni, A. (2017). A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. *Advances in Science, Technology and Engineering Systems Journal*, 2(6), 275–284. doi:10.25046/aj020634
- Ho, S.M., Kao, D., Chiu-Huang, M., & Li, W. (2020) Detecting Cyberbullying “Hotspots” on Twitter: A Predictive Analytics Approach. *Forensic Science International: Digital Investigation*, 32(2020), 1-3.
- Kowalski, R. M., Limber, S. P., & McCord, A. (2019). A developmental approach to cyber- bullying: Prevalence and protective factors. *Aggression and Violent Behavior*, 45, 20–32. doi:10.1016/j.avb.2018.02.009

- Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015). Collaborative detection of cyberbullying behavior in Twitter data. *2015 IEEE International Conference on Electro/Information Technology (EIT)*, 611-616. doi:10.1109/EIT.2015.7293405
- Nahar, V., Li, X., & Pang, C. (2014). An effective approach for cyberbullying detection. *Communications in Information Science and Management Engineering*, 3(5), 238–247.
- Nandakumar, V., Kovoov, B. C., & Sreeja, U. M. (2018). Cyberbullying Revelation In Twitter Data Using Naïve Bayes Classifier Algorithm. *International Journal of Advanced Research in Computer Science*, 9(1), 510–513. doi:10.26483/ijarcs.v9i1.5396
- Nandhini, B. S., & Sheeba, J. I. (2015). Cyberbullying Detection and Classification Using Information Retrieval Algorithm. *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) - ICARCSET '15*. doi:10.1145/2743065.2743085
- Nixon, C. (2014). Current perspectives: The impact of cyberbullying on adolescent health. *Adolescent Health, Medicine and Therapeutics*, 5, 143–158. doi:10.2147/AHMT.S36456 PMID:25177157
- Olweus, D. (2012). Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology*, 9(5), 520–538. .682358<ALIGNMENT.qj></ALIGNMENT>10.1080/17405629.2012
- Ptaszynski, M., Masui, F., Nitta, T., Hatakeyama, S., Kimura, Y., Rzepka, R., & Araki, K. (2016). Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization. *International Journal of Child-Computer Interaction*, 8, 15–30. doi:10.1016/j.ijcci.2016.07.002
- Rosa, H., Carvalho, J. P., Calado, P., Martins, B., Ribeiro, R., & Coheur, L. (2018a). Using fuzzy fingerprints for cyberbullying detection in social networks. In *Proceedings of the IEEE International Conference on Fuzzy Systems* (pp. 56–62). doi:10.1109/FUZZ-IEEE.2018.8491557
- Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. P. (2018b). A deeper look at detecting cyberbullying in social networks. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 323–330). .2018.848921110.1109/IJCNN
- Salawu, S., He, Y., & Lumsden, J. (2017). *Approaches to automated detection of cyberbullying: A survey*. Transactions on Affective Computing.
- Sanchez-Medina, A.J., Galvan-Sanchez, I., & Fernandez-Monroy, M. (2020). Applying artificial intelligence to explore sexual cyberbullying behaviour. *Heliyon*, 6, 1-9.
- Singh, V. K., Huang, Q., & Atrey, P. K. (2016). Cyberbullying detection using probabilistic socio-textual information fusion. In *Proceedings of the international conference on advances in social networks analysis and mining* (pp. 884–887). ASONAM.
- Souza, S. B., Veiga Simão, A. M., Ferreira, A. I., & Ferreira, P. C. (2018). University students' perceptions of campus climate, cyberbullying and cultural issues: implications for theory and practice. *Studies in Higher Education*, 11(43), 2072–2087.
- Sugandhi, R., Pande, A., Agrawal, A., & Bhagat, H. (2016). Automatic monitoring and prevention of cyberbullying. *International Journal of Computers and Applications*, 8, 17–19. <https://pdfs.semanticscholar.org/eb09/>
- Van, H. C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS One*, 13(10), e0203794. doi:10.1371/journal.pone.0203794
- Veiga Simão, A. M., Ferreira, P., Francisco, S. M., Paulino, P., & Souza, S. B. (2018). Cyberbullying: Shaping the use of verbal aggression through normative moral beliefs and self-efficacy. *New Media & Society*, 14, 1–20. 10.1177/
- Wright, M. F. (2017). Cyberbullying in cultural context. *Journal of Cross-Cultural Psychology*, 48(8), 1136–1137. <https://doi.org/10.1177/0022022117723107>
- Zhang, S., Yu, L., & Wakefield, R. L., & Leidner, D. (2016). Friend or Foe: Cyberbullying in Social Network Sites. *The Data Base for Advances in Information Systems*, 47(1), 51–71.
- Zhao, R., & Mao, K. (2016b). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, 8(3), 328–339. <https://doi.org/10.1109/taffc.2016.2531682>

Zhao, R., Zhou, A., & Mao, K. (2016a). Automatic detection of cyberbullying on social networks based on bullying features. In Proceedings of the 17th international conference on distributed computing and networking. ACM Press. <https://doi.org/10.1145/2833312.2849567>.

Nureni Ayofe AZEEZ obtained his B. Tech. (Hons.) from the Federal University of Technology, Akure, Nigeria in 2005, MSc from the University of Ibadan, Oyo State, Nigeria in 2008, and Ph.D. from University of the Western Cape, South Africa in 2013, all in Computer Science. His areas of research include Security & Privacy, Access Control, Grid and Cloud Computing, Sensor Networks, E-Health and ICT4D. He is a recipient of The Young Scientist Award at the 22nd International CODATA Conference that was held in Cape Town, South Africa in October 2010. He is currently a Senior Lecturer in the Department of Computer Sciences, University of Lagos, Nigeria.

Sanjay Misra is a Professor at Ostfold University College(HIOF), Halden Norway. Before coming to HIOF he was a professor in Covenant University (400-500 ranked by THE(2019)) Nigeria for 9 yrs. He is PhD. in Inf. & Know. Engg (Software Engg) from the Uni of Alcala, Spain & M.Tech.(Software Engg) from MLN National Institute of Tech, India. As per SciVal (SCOPUS- Elsevier) analysis (on 01.09.2021)- He is the most productive researcher(Number 1) <https://t.co/fBYnVxbmiL> in Nigeria since 2017 (in all disciplines), in comp science no 1 in the country & no 2 in the whole of Africa. Total more than 500 articles (SCOPUS/WoS) with 500 coauthors worldwide (-110 JCR/SCIE) in the core & appl. area of Soft Engg, Web engineering, Health Informatics, Cybersecurity, Intelligent systems, AI, etc. He got several awards for outstanding publications (2014 IET Software Premium Award(UK)), and from TUBITAK-Turkish Higher Education and Atilim University). He has delivered more than 100 keynote/invited talks/ public lectures in reputed conferences and institutes (traveled to more than 60 countries). He is one editor in 58 LNCSS, 4 LNEEs, 1 LNNs, 3 CCISs & 10 IEEE proceedings, six books, and Editor in Chief of 'IT Personnel and Project Management, Int J of Human Capital & Inf Technology Professionals -IGI Global & editor in various SCIE journals.

Omotola Ifeoluwa LAWAL graduated with a B.Sc. (Hons) in Computer Science with Second Class Upper Division from the University of Lagos, Nigeria in 2019. His areas of interest include Cyber Security, Cloud Computing and Distributed Systems.

Jonathan Oluranti is a lecturer in Covenant University, Ota, Ogun State, Nigeria. His area of research is e-government, ICT, artificial intelligence.