ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 14 (2024) No. 8, pp. 262-284
DOI:10.33168/JSMS.2024.0817

Examining Machine Learning Approaches for Detecting Cyberbullying in Social Media Content

Esraa C. Hussein Omran [1], Jamal Al Qundus [2], Hussain Al Sharoufi [1], Kosai Dabbour [3]

[1]Computer Science Department, Center of Applied Mathematics and Bioinformatics (CAMB), and Linguistics English Department Gulf University for Science and Technology, Kuwait City, Kuwait

[2]Business Intelligence and Data Analytics, German Jordanian University, Amman, Jordan

[3]EVA Electronics Co., Hawally, Kuwait

hussain.i@gust.edu.kw, jamal.alqundus@gju.edu.jo, alsharoufih@gust.edu.kw, qusai@evakw.com

Abstract. This study investigates machine learning approaches for detecting cyberbullying across textual social media data. Three models - Extra Trees, Random Forest, and XGBoost are evaluated on a labeled dataset of 20k tweets. Results indicate Extra Trees achieves highest accuracy (90%) and AUC-ROC (95%) for classifying cyberbullying vs non-cyberbullying posts. Additionally, lexical analysis of 2000 YouTube comments expands existing knowledge of terms and phrasing markers of online harassment. The research contributes both methodological and practical advances in applying ML to combat rising social media hostility. We make the code and the generated data freely available at https://github.com/jamalalqundus/code-paper-cyberbullying-detection.git for further research.

Keywords: Cyberbullying, Hate Speech Detection, Machine learning Model, User-generated Content

1. Introduction

Affecting not only adults, the Annapolis Police Department[1] reports that nearly 42% of children and 9/10 of middle school students have been cyberbullied or had their feelings hurt online. The psychological consequences of cyberbullying are similar to those in real life, with the difference being that cyberbullying takes place over the internet, which is available 24/7. Different time periods and eras have different linguistic vocabularies. There are no limits to the use of words and the creation of new words for cyberbullying. New and hidden paraphrases need to be tracked and databases updated accordingly.

Cyberbullying refers to words or acts whose intent is to induce hatred towards a selected group of people, which could also be a community, religion, or race (Chandrasekaran, Singh Pundir, and Lingaiah 2022). This speech may or might not be meaningless but is probably going to end in violence. Hate speech online has been linked to a worldwide increase in violence toward minorities, including mass shootings, lynching, and group action.

Moreover, due to the huge rise of user-generated content including personal data and information sharing (Fu et al. 2020; Wakefield and Wakefield 2023), the number of hate speech encounters is steadily increasing. Particularly on social media networks that have major impact on several areas e.g. business(Rahman et al. 2022; Alalwan et al. 2020) and even politics (Grover et al. 2021). Due to the increasing prevalence of cyberbullying on social media and the resulting harmful effects, especially on the younger generation, research on cyberbullying detection has increased recently (Muneer and Fati 2020). There is a growing body of work on automated approaches to cyberbullying detection (e.g. (Lozano-Blasco, Cortés-Pascual, and Latorre-Martínez 2020; Bozyiğit, Utku, and Nasibov 2021) (Marabelli, Vaast, and Li 2021)). Such experiments use machine learning (ML) technologies, including natural language (NLP) processing, to detect and automatically identify the characteristics of a cyberbullying exchange by matching text data with the detected features.

However, simple word filters do not provide sufficient remediation for detecting expressions classified as hate speech. Since these can be blurred by aspects such as the context, utterance domain and discourse context of the media objects (e.g., video, audio, image), specialized language processing is required to overcome this challenge.

The need is twofold: (1) Data sets are always in demand, especially qualitative and intellectualized data sets that provide segments and patterns related to cyberbullying. Such datasets serve as a basis for improving the performance of machine learning algorithms. (2) Selecting a suitable algorithm is always a challenge, as algorithms in this context are case-dependent. The cases are related to the dataset acting as input that influences the performances. The machine learning algorithms need to be continuously compared in different cases to facilitate the selection of such algorithms in subsequent similar cases. These two targets are the research objectives of this work.

To this end, existing works applying Machine learning techniques to address cyberbullying detection in social media follow either Deep Learning approaches e.g. (Chandrasekaran, Singh Pundir, and Lingaiah 2022) or Shallow approaches e.g. (Bozyiğit, Utku, and Nasibov 2021) achieved good results. However, research suffers from the lack of focus on the input data investigating the combination of machine learning and natural language processing data preprocessing pipelines to compare the performance of several models in detection of cyberbullying. Moreover, the use of free text leads to the formulation of phrases that are difficult to recognize by machine learning techniques. Keywords related to cyberbullying can be found in the literature, but not manually extracted key phrases, which are very important for various use cases, e.g., to overcome ambiguity due to lack of context, or to classify cyberbullying segments in order to select appropriate combat, etc. Such exploration, in conjunction with the provision of enriched datasets, is still hard to find or difficult to access in previous work. Moreover, manually processed datasets are always valuable because they involve human decisions, unlike

[1] https://www.annapolis.gov/908/Facts-About-Cyberbullying [accessed:12.12.2023]

machine-generated datasets, which in turn offer fewer performance improvements when training algorithms. To our knowledge, this is the first study to address this research gap.

This work presents a comprehensive and structured tour of automatic cyberbullying detection and compares machine learning algorithms in a very methodical way, along with insightful coverage of several published research papers on cyberbullying detection techniques. This study investigates the forms of cyberbullying using a focused research question: To what extent can cyberbullying on social media be detected using machine learning techniques augmented with manually selected keywords and phrases? In order to address it, this work applies combined data preprocessing on short text from social media twitter feeding the resulting data into a set of selected Machine learning models and comparing their performances. Frequency-Inverse Document Frequency (TF-IDF) and Dimensionality Reduction were applied to highlight the limitations and advantages of Machine learning classification models towards cyberbullying.

In addition, the current study examined the cyberbullying-labeled dataset of 2,000 YouTube comments collected by (Ashraf, Zubiaga, and Gelbukh 2021) and expanded it by manually extracting 3,044 keywords, of which 932 are unique, and 1,106 key phrases related to cyberbullying. Comments on YouTube are more likely to be directed at groups than elsewhere, as our preliminary examination of various data showed. This is more in line with our research, which is why we decided to manually explore and extend the YouTube dataset for our work. We made this dataset freely available for further research[2]. It can serve as a training set for a machine learning model or a neural network, or as a pattern for building a dictionary for recognizing cyberbullying in social networks, e.g., by a multiagent system. The manually extracted keywords and especially key phrases can be used to identify new cyberbullying phrases, group them according to their toughness or impact on victims, and rank them, for example, to select an appropriate response or reaction.

The contributions of the current work are summarized in the following list:

1) Investigate the performance of various machine learning algorithms for detecting cyberbullying to reduce selection overhead in subsequent work.
2) Exploring datasets generated from various online communities, i.e. Twitter and YouTube
3) Providing pruned and supervised datasets as additional enrichment that can be of great value in supporting machine learning models.
4) Providing unique sets of keywords and key phrases, including infrequently used expressions, that were manually identified to facilitate the detection of cyberbullying phrases.

Provide an ordering of these expressions to support further research in the area of combating or responding appropriately to cyberbullying.

This paper is organized as follows: Section 2 presents related work that motivated the current work. In Section 3 includes the machine learning models applied. Section 4 represents the results that are further discussed in Section 5. Conclusion, Limitation and Future Scope of Research take place in Section 6.

2. Literature Review

Myriad approaches have been developed to detect cyberbullying, especially machine learning approaches for classification of text data by feature extraction applying TF-IDF, sentiment analysis, dimensionality reduction, etc., which have achieved considerable accuracy. In related work, different machine learning algorithms have been applied to multiple languages with different features, e.g. text length, to detect cyberbullying on social media. In (Akhter et al. 2023), a robust hybrid ML model was developed for detecting cyberbullying in Bengali language on social media. It incorporates effective text preprocessing to convert the Bengali text data into a usable text format. Feature extraction using

[2] https://github.com/jamalalqundus/code-paper-cyberbullying-detection.git

TfidfVectorizer (TFID) was used to extract the useful information from the text data. In a very similar work also investigating the same language, (Ahmed et al. 2023) used multiple machine learning classifiers to recognize bully text. Bangla texts were collected from the comment section of political Facebook posts and then classified as bullied or non-bullied. Random Forest provided an accuracy of 91.08%. Other challenging and strong languages such as Arabic were also considered by (Khairy et al. 2023; Alzaqebah et al. 2023). In (Khairy et al. 2023), three machine learning classifiers were applied to three Arabic datasets, which are publicly available, to detect cyberbullying. Their voting principle results achieved accuracy values of 71.1%, 76.7% and 98.5% for each of the three datasets used. While in (Alzaqebah et al. 2023),the focus was on imbalanced short texts and different dialects in the Arabic text data, good results are obtained using different machine learning algorithms. These works and our work have similar investigations and comparable results. However, the novelty of our work lies in the manual extraction of cyberbullying phrases as well as the more realistic performance comparison due to the intellectual effort involved in preprocessing, which provides a solid input for the machine learning models under investigation.

According to (Alduailaj and Belghith 2023) a limited amount of work has investigated the detection of cyberbullying on Arabic social media platforms. The authors explored this direction by applying Arabic language machine learning for automatic cyberbullying detection. The authors used datasets from YouTube and Twitter to train a support vector machine model to detect cyberbullying with 95.74% accuracy.

The focus of (Casavantes et al. 2023) was on detecting abusive, aggressive, hostile, and hateful messages. To this end, the authors extended a set of Twitter benchmark datasets and evaluated different learning models considering classical (Bag of Words), advanced (Glove), and modern (BERT) text representations. They used a mixture of text and meta-data and showed significant differences between the classification results of all methods, pointing to the importance of context as the most important factor to consider.

(Murshed et al. 2022) built and compared the model efficiency of a deep learning approach (DEA-RNN) on a data set of 10000 tweets using Bi-directional long- short-term memory (Bi-LSTM), Recurrent Neural Network (RNN), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB) and Random Forests (RF) based algorithms. Their experimental results show improved performance of DEA-RNN. An average accuracy of 90.45% was achieved for CB detection. Following a related Deep Learning approach, (Kumari and Singh 2021) proposed a neural network that uses multimodal information to detect hate speech in social media and processes text and image features. The authors used a pre-trained network and an RNN for feature extraction and applied a genetic algorithm for optimization, which achieved a performance of 78% of the F1-score.

(Zhu et al. 2021) explored the Reddit comments dataset and trained a word embedding model based on the word2vec Skip-Gram model to identify cyberbullying. Using properties of the new word embedding model for training a random forest model to classify cyberbullying comments, four pre-trained word embedding models and manual feature extraction methods were proposed. While, (Silva, Hall, and Rich 2018) suggested a model for detecting cyberbullying based on psychological research. They described the development of an app called Bully Blocker to notify a user's parents when cyberbullying is detected. They analyzed the user's social media data in a traditional way, examining comments and messages to classify it as warning sign or as bullying. The application, designed particularly for teens, employs Facebook's older detection methods, but could be expanded as an app for data collection and ML classification.

(Yao, Chelmis, and Zois 2019) conducted a survey of social media users to gather examples of nastiness and used machine learning techniques to classify these examples into categories of nastiness. The authors propose a machine learning-based approach to detecting cyberbullying by analyzing the content and context of social media posts. The system would identify patterns and features of cyberbullying, such as the use of derogatory language or repeated harassment, to accurately detect and

classify instances of cyberbullying. They discussed the ethical considerations of such a system, including the need for user privacy and the potential for false positives.

(Samghabadi et al. 2017) explored ways to recognize meanness on social networks applying NLP approaches to identify as well as prevent cyberbullying. These approaches were applied in a manner that they can also detect whether bad language is used in the content in an offensive or neutral manner. The data were obtained from posts in English-language on social media websites, including semi-anonymous social media websites like ask.fm. A modified linear SVM method was taken to detect negative words in an incidental manner, and several additional features that might be overlooked were accounted for, like: questions and answer posts as well as emojis. This model achieves an F1-score of 0.59, considering that this study did not use user-defined data but a real data set.

(Abarna, Sheeba, and Devaneyan 2023) greatly decreased the number of characteristics used for categorization while keeping high accuracy by using a sequential hypothesis test design. This takes into consideration the recurrent character of online bullying. Cyberbullying is defined as a series of abusive messages sent by a bully with the intention of harming the victim. The main objectives of this strategy are high accuracy, scalability, as well as timeliness. They applied pre-trained word embedding language models for feature selection and a knowledge-based frequent pattern method. The model is trained with unsupervised approaches on the Instagram dataset, which was assembled by snowball sampling and partially manually labeled by a team of domain experts. Limitations of this technique were the use of a single dataset unique to Instagram, the lack of a way to verify the accuracy of the labeling, and the time required to collect annotation-based labels.

(Abaido 2020; Cao et al. 2020) examined the impact of cyberbullying on individuals and society, as well as ways to prevent and address it. The authors explicate various forms that cyberbullying can take, including cyberstalking, trolling, and sexting, and the methods in which it can differ from traditional bullying. They also delve into the psychological and social factors that contribute to cyberbullying, such as anonymity, lack of face-to-face communication, and the ease of spreading information online. The authors provide practical strategies for addressing and preventing cyberbullying, including educating young people about responsible online behavior, establishing clear policies and procedures, and involving parents and other adults in the process.

3. Methodology

This section describes the approach conducted during this study, which begins with data preparation processing followed by investigation and comparison of several ML models.

3.1. Dataset

The experiment conducted relies on a balanced and high-quality dataset provided by Kaggle in a JSON format. This dataset[3] is manually labeled and containing totally 20,001 instances distributed over 2 attributes (Tweet Text, and Label) where the Label corresponds to non-cyberbullying and cyberbullying. In addition, we consider the cyberbullying-labeled dataset of 2,000 YouTube comments collected by (Ashraf, Zubiaga, and Gelbukh 2021) and expanded it by manually extracting 3,044 keywords, of which 932 are unique, and 1,106 key phrases related to cyberbullying.

The dataset has the potential to explore narratives related to cyberbullying and the impact of social media on users' sharing behavior on social media platforms. Tracking back to a specific topic is difficult due to the proliferation of information on most of these platforms. This is different from YouTube, where posts follow a specific structure and are divided into the main topic of the video, the reaction to it, and the comments on it. Also, comments on YouTube are more likely to be directed at groups than elsewhere, as our preliminary examination of various data showed. This is more in line with our research, which is why we decided to use the YouTube dataset for our work.

[3] https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls

3.2. Data preparation

As known, the most ML model cannot deal with dependent variables of string-format, that's why the first step was encoding the target variable into integer, where 0 replaces non-cyberbullying and 1 replaces cyberbullying labels as shown in Table 1.

Table 1: Data sample showing the dependent variable (annotation) being encoded

ID	Content	Annotation
0	Get fucking real dude.	1
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
2	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
3	Dude they dont finish enclosing the fucking showers. I hate half assed jobs. Whats the reasoning behind it? Makes no sense.	1
4	WTF are you talking about Men? No men thats not a menage that's just gay.	1
…	…	…
19996	I dont. But what is complaining about it going to do?	0
19997	Bahah yeah i&;m totally just gonna&; get pissed at you for talking to you. Mhm thats just how i am! ;D r Ha noooot ss much.	0
19998	hahahahaha >:) im evil mwahahahahahahahaha	0
19999	What&;s something unique about Ohio? :)	0
20000	Who is the biggest gossiper you know?	0

The dataset includes a total of 20001 instances, distributed between 7822 instances labeled 1 and 12179 instances labeled 0, which is slightly unbalanced as illustrated in Figure 1.

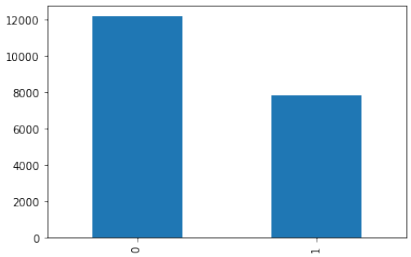


Fig. 1: Labels vs. instances distribution

The Youtube dataset includes 2000 cyberbullying-label rows consisting of the columns id, comment, replies, title, label (having value 1), political, religion, other, time, date. This work considered the columns id, comment and label for the investigation.

3.3. Data cleaning

Using the NLTK[4] library along with REGEX[5] library the following NLP preprocessing steps have been applied: (1) Word Tokenization to convert each tweets text in each instance into a list of words/tokens to facilitate further operation on a single word. (2) English stop words Removal to eliminate all words from each instance's tweet text as (in, the, on, a, an, etc.). This process has been followed directly with a punctuation removal process using a predefined regular expression designed and implemented to remove all punctuation marks like (?, !, :, emojis, etc.). (3) Digit Removal has been applied to remove all numbers (integer and real) from the instance's tweet text. (4) Stemming is very useful by reducing the context grammars in any list of tokens for further operations such as word-counting. Each comment

[4] https://www.nltk.org
[5] https://pypi.org/project/regex/

of the Youtube dataset has been manually examined and individual keywords as well as key phrases have been extracted into a separate column as shown in Table 2.

Table 2: shows samples of keywords and key phrases manually extracted from comments

video-id	**comment**	**Comment keyphrases**
UgjsJ8edAikfNngCoAEC	butthurt old fuck, someone ought to teach him a lesson. pathetic swine.	butthurt fuck pathetic swine
Ugw01F2oXDD_in0pKcN4AaABAg	When someone presents you with new info that flies in the face of your world view, maybe listen and make sure there isn't something to it before you knee jerk. If they are wrong, they are wrong, but as progressive, this is what the trump supporters are crying about and they're right.	knee-jerk
UgjOYOFZD27I9XgCoAEC	CORRUPTION. raises bail because shes not in a bad mood, he just had 2 try n break her down. fuck that judge	CORRUPTION fuck
UghINqGgHMsSpngCoAEC	you do realize if trump doesn't win, we are screwed.	screwed
UgyJOdnd7DjvH2vkunN4AaABAg	Poor impulse control, parents. Some black kids spring out of their seats. After all, she done called a spade a spade.	black-kids
UgxlIltpsnx4vF5EpTN4AaABAg	Spoiled twat	Spoiled-twat
Ugz0SwNENi8UDiXl9XJ4AaABAg	What shameful pieces of shit treating that man like that.	shameful pieces-of-shit
Ugx-buppYtj532FUsoZ4AaABAg	I LOVE THIS Girl even if I am Gay, Oh My God, is this America and American Dream, God Thanks I do believe in Freedom like this whore in New York, State of Liberty:-))	whore
UgxcwBvbnbKbdycjglh4AaABAg	Speak face to face , they will kill you	will-kill-you
UghCv3GQHBAku3gCoAEC	Only stupid people mess with a person who is alone with their food.	stupid-people
UgwrgAXtFv0g8vUPUHp4AaABAg	the difference in between islam and hinduism in india are,, islam came from arab,, and who raped, killed, forced to convert for islam,, they are so called muslims in india,, and hindu's are somuch toleranted people,, who gave their life but never converted themself for islam,,	raped killed
UggZNMwzgKw7tHgCoAEC	You can't fix stupid.. so if you're stupid... well... you get about a month's worth of striped tanning	you're-stupid
UgyKlR5Uo7QWsbxNxax4AaABAg	wtf was wrong with the cop smfh thats what happens when give mentally unstable people a badge	wtf mentally-unstable
UgwW1gL61uvkLLM6sLp4AaABAg	Whoever did the b role for this video should be fired its fuckin hoeeible	fuckin-hoeeible

Extracted keywords (3044) are separated by spaces, while words of the key phrases (1106) are connected with '-' character

3.4. Data representation

The previous preprocessing phase converts the twitter data into single stemmed tokens, as a consequence, a feature extraction process is needed. To this end, the following data representation phase covers Term Frequency Inverse Document Frequency (TF-IDF) and Dimensionality Reduction.

3.5. TF-IDF transformation

TF-IDF is a statistical process measures the importance of a word/token by calculating the multiplication between the number of times that a word appeared in a specific document (Term Frequency TF) with the word distribution in inverse documents (Inverse Document Frequency IDF). The higher the TF-IDF score, the more relevant that word is in that particular document.

Applying this statistical model on the dataset results in a matrix containing 20,001 rows representing instances and 14,783 columns representing TF-IDF features for each tweet text instance. A vectorization process on the TF-IDF matrix was conducted to get the scores for each individual token and highlight tokens having the highest scores indicating the token importance as illustrated in Figure 2.

```
hate            Score: 533.8157298036014
fuck            Score: 503.76150769255435
damn            Score: 482.3875012051478
suck            Score: 407.37790877127185
ass             Score: 337.54089621427744
that            Score: 311.6250930420745
lol             Score: 298.0085779872157
im              Score: 296.0216055277791
like            Score: 287.8183474868775
you             Score: 284.7850587424088
it              Score: 254.75722294501585
get             Score: 253.19747902607998
what            Score: 221.43673623523864
know            Score: 211.53595900888456
would           Score: 202.5073882820925
bitch           Score: 193.08800391463464
ye              Score: 182.22364463196365
love            Score: 181.49014270754344
go              Score: 180.2588319545915
haha            Score: 179.29466045019018
think           Score: 178.9039058038677
one             Score: 174.16019276608517
do              Score: 160.57524593088053
time            Score: 160.1100301847739
gay             Score: 159.5820454915121
peopl           Score: 151.04499856119287
```

Fig. 2: Sample tokens with related TF-IDF scores

3.6. Features space reduction using PCA

Each individual token represents a feature. The large number of tokens creates a high-dimensional space that makes it difficult to identify the relevant independent variables. This high-dimensional space must be reduced to the most important dimensions (features). Principal Component Analysis (PCA) is a technique used in machine learning to reduce the dimensionality of a dataset. It is done by identifying the directions in which the data varies the most and projecting the data onto a lower-dimensional space along these directions.

In this phase, PCA was applied to the sparse TF-IDF matrix to reduce the feature space. First, the documents were transformed into a matrix of TF-IDF features. Each row of the matrix represents a document, and each column represents a word. The value in each cell is the TF-IDF weight for that word in that document. Once the TF-IDF matrix is created, PCA can be applied as follows:

- Standardize the data by subtracting the mean and scaling to unit variance.
- Calculate the covariance matrix of the data.
- Calculate the eigen-vectors and eigen-values of the covariance matrix. The eigen-vectors are the principal components, and the eigen-values are their corresponding variances.
- Sort the eigen-vectors by their corresponding eigen-values in decreasing order.
- Select the k eigen-vectors with the highest eigen-values, where k is the desired number of dimensions in the reduced space.

Project the data onto the space spanned by the k eigen-vectors to obtain the reduced data. Figure 3 illustrates a sample of the transformed dataset based on PCA and TF-IDF.

	A	B	TARGET
0	-0.029823	-0.219646	1.0
1	-0.027610	-0.084035	1.0
2	-0.031197	-0.091155	1.0
3	-0.067740	0.006192	1.0
4	0.052153	0.023154	1.0
...	...	...	...
19996	-0.021709	-0.002466	0.0
19997	0.002002	-0.001112	0.0
19998	-0.004288	-0.013867	0.0
19999	-0.009142	-0.011963	0.0
20000	-0.009088	-0.008184	0.0

Fig. 3: Dataset sample after PCA

3.7. Data splitting

The last step in the data preparation process is data splitting. Table 3 represents the split of the dataset into 80% training instances in order to train different machine learning models and 20% testing in order to evaluate the trained models and derive the best and most suitable classification model.

Table 2: Overview of training and test data instances

	Training	**Test**
Cyberbullying	9750	2429
Non-Cyberbullying	6250	1572
Total	16000	4001

3.8. Youtube dataset processing

After extracting keywords and key phrases stemming and Bag-of-Word approaches have been applied on the keywords. The key phrases were compared among each other using the language model distilbert-base-uncased from huggingface.io[6] to calculate the sentence similarity (cosine similarity) resulting in a new dataset phrase_similarity.csv[7] consisting of the columns phraseI, phraseIplusK, cosineScore.

4. Experimental Results

The Machine learning models investigated during this study are Naïve Bayes Model, Decision Tree Model, Random Forest Model, Extra Tree Model, XGBoost and Logistic Regression. The selection of these models was based on literature review. Researchers initially select this kind of model due to their sensibility towards preprocessing the input. It is generally known that they achieve good performance in several tasks, which is another reason for their selection for this investigation.

Figure 4 summarizes the approach taken in this work to process the datasets under consideration, resulting in a new dataset of manually extracted features. Starting from the datasets collected from Kaggle and YouTube, the ML/NLP preprocessing phase consisting of data preparation, data cleaning and data representation begins. A set of models is used for performance comparison. Keyword extraction results in a new unique dataset that is made freely available for further research.

[6] https://huggingface.co/distilbert-base-uncased
[7] Included in the repository freely available

Table 4: Naïve Bayes Model Performance Measure

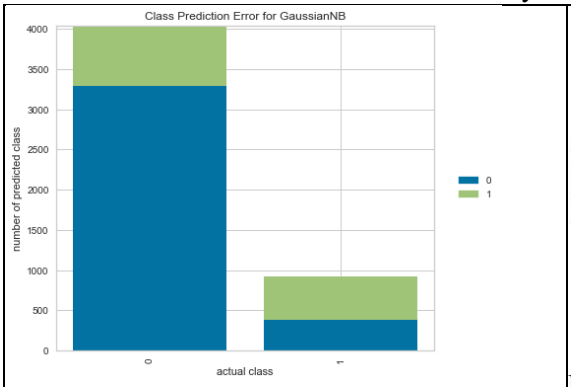


Fig 4: Class Prediction Error

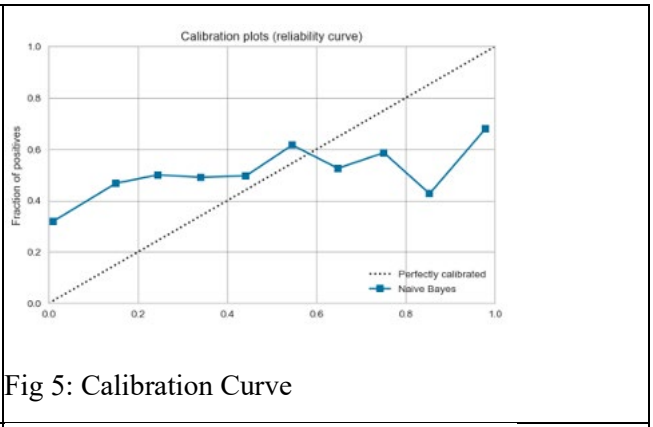


Fig 5: Calibration Curve

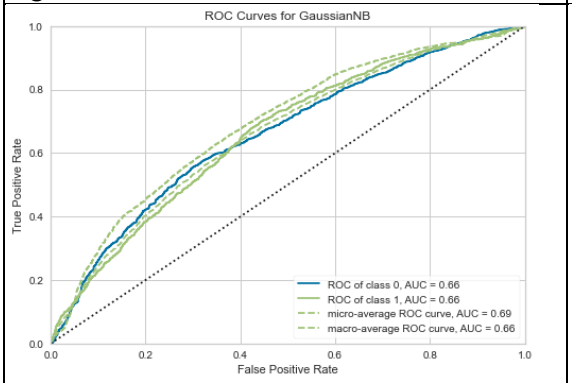


Fig. 6: ROC Area

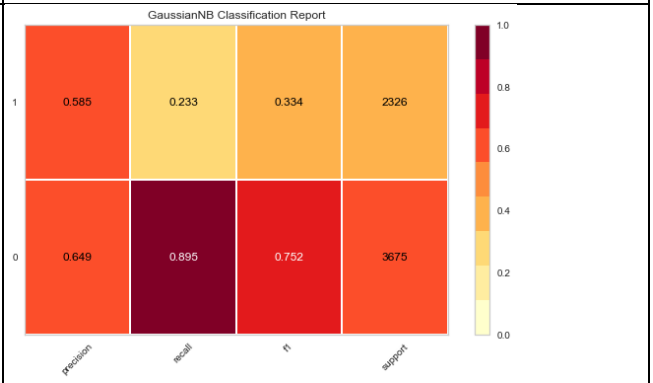


Fig 7: Class Report

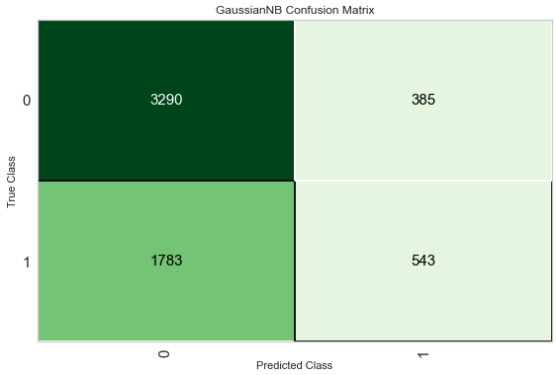


Fig 9: Confusion Matrix

Table 5 includes the performance measures of the Decision Tree (DT) model. The DT model is quite good (better than NB model) at correctly predicting cyberbullying and good at correctly predicting non-cyberbullying.

Figure 10 represents the class prediction error for DT on the number of predicted classes to the actual class. The bar chart in Figure 10 shows the support (number of training samples) for the non-cyberbullying (0) and cyberbullying (1) classes in the fitted DT classification. The two bars are divided into the proportion of predictions (thereof FP and FN) for the two classes (0 and 1). Figure 11 illustrates the calibration of the DT model having irregular flow and the calibration suffered indicating its need to be retrained to improve its accuracy and calibration.

The plotted True Positive Rate (TPR) and False Positive Rate (FPR) variables in the ROC curve in Figure 12 show the response of the DT model for five thresholds. Figure 13 and Figure 14 provide the class report and confusion matrix of the DT classification, respectively. Both confirm the observation provided by the bars in Figure 10.

Table 5: Decision Tree Model Performance Measure

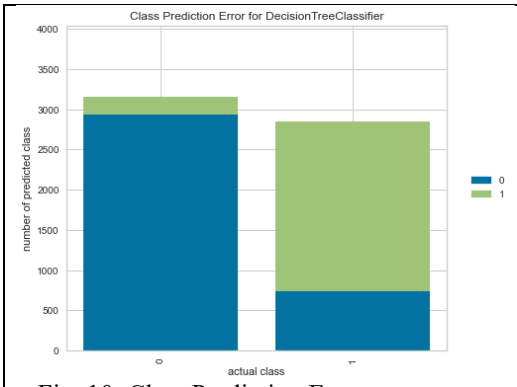


Fig. 10: Class Prediction Error

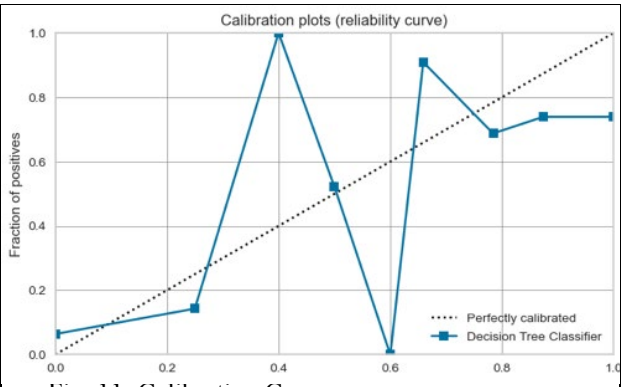


Fig. 11: Calibration Curve

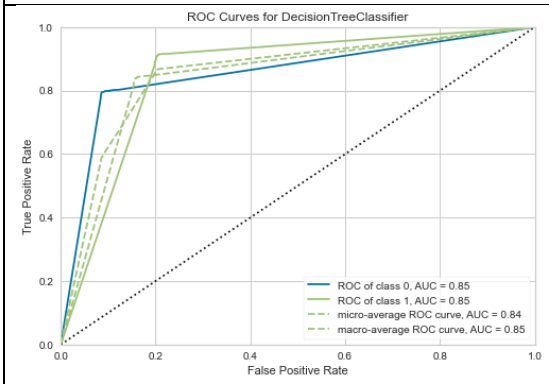


Fig. 12: ROC Area

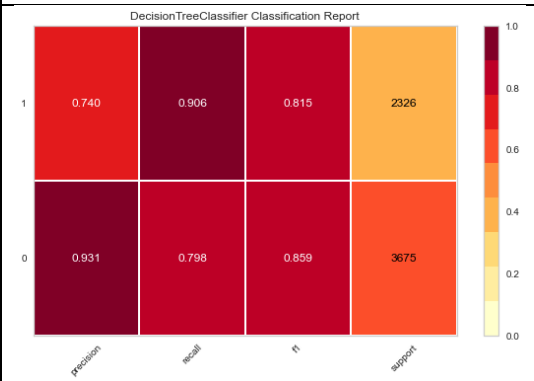


Fig. 13: Class Report

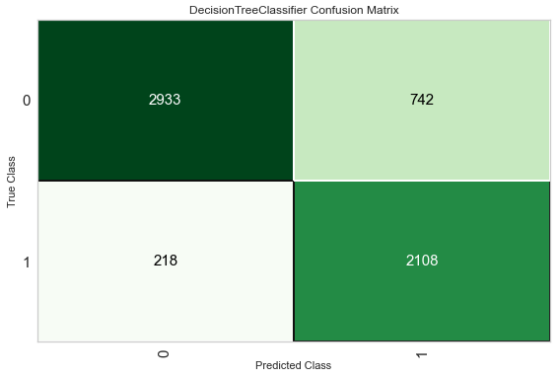


Fig. 14: Confusion Matrix

Table 6 contains the performance measures of the Random Forest (RF) model. The bar chart in Figure 15 includes the support (number of training samples) for the non-cyberbullying (0) and cyberbullying (1) classes in the fitted RF classification. The two bars are divided into the proportion of predictions (thereof FP and FN) for the two classes (0 and 1).

Figure 15 illustrates the class prediction error for RF on the number of predicted classes to the actual class. While the RF model seems to be quite good (better than DT model) at correctly predicting non-cyberbullying and good at correctly predicting cyberbullying.

Figure 16 represents the calibration of the RF model having a better flow indicating improved accuracy and calibration. The plotted True Positive Rate (TPR) and False Positive Rate (FPR) variables in the ROC curve in Figure 17 show the response of the RF model for five thresholds. Figure 18 and Figure 19 conclude the class report and confusion matrix of the RF classification, respectively. Both confirm the observation shown in the bars in Figure 15.

Table 6: Random Forest Model Performance Measure

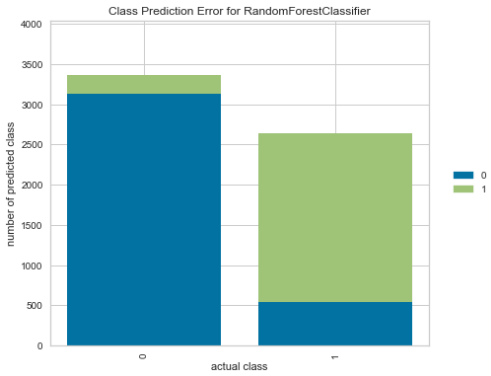


Fig. 15: Class Prediction Error

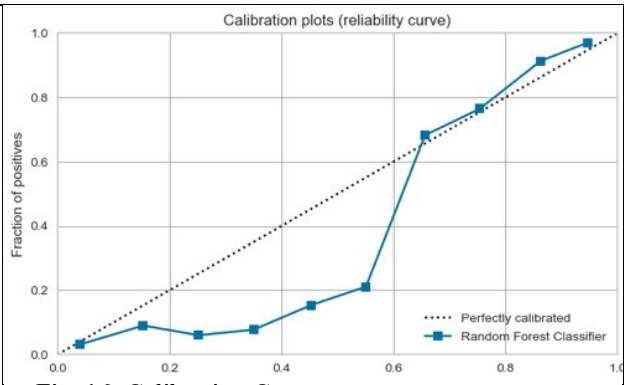


Fig. 16: Calibration Curve

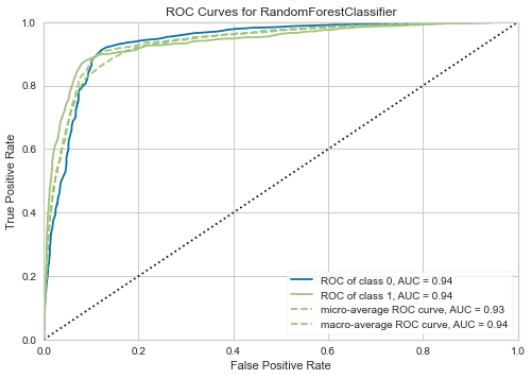


Fig. 17: ROC Area

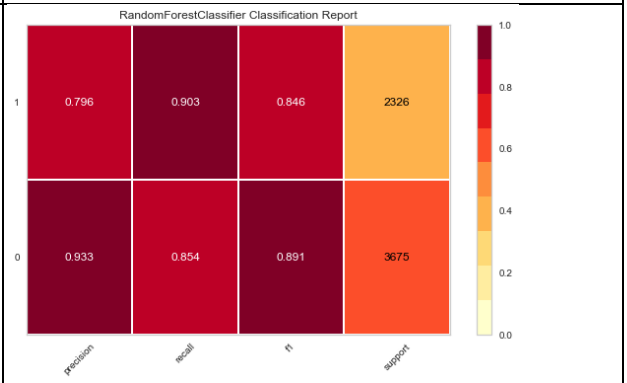


Fig. 18: Class Report

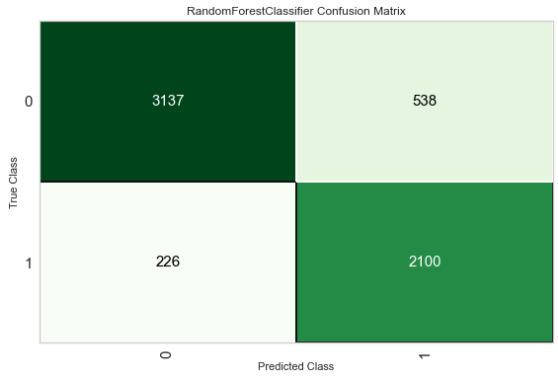


Fig. 19: Confusion Matrix

Table 7 covers the performance measures of the Extra Tree (ExF) model. Figure 20 shows the class prediction error for ExF on the number of predicted classes to the actual class. The two bars are divided into the proportion of predictions (there of FP and FN) for the two classes (0 and 1). They include the support (number of training instances) for the non-cyberbullying (0) and cyberbullying (1) classes in the fitted ExF classification. While the ExF model seems to be quite good (comparable to DT model) at correctly predicting cyberbullying and good at correctly predicting non-cyberbullying.

Figure 21 illustrates the calibration of the ExF model having a good flow indicating good accuracy and calibration. The plotted True Positive Rate (TPR) and False Positive Rate (FPR) variables in the ROC curve in Figure 22 show the response of the ExF model for five thresholds. Figure 23 and Figure 24 represent the class report and confusion matrix of the ExF classification, respectively. Both confirm the observation provided by the bars in Figure 20.

Table 7: Extra Tree Model Performance Measure

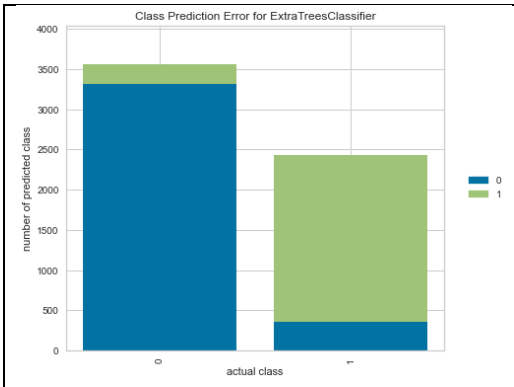


Fig. 20: Class Prediction Error

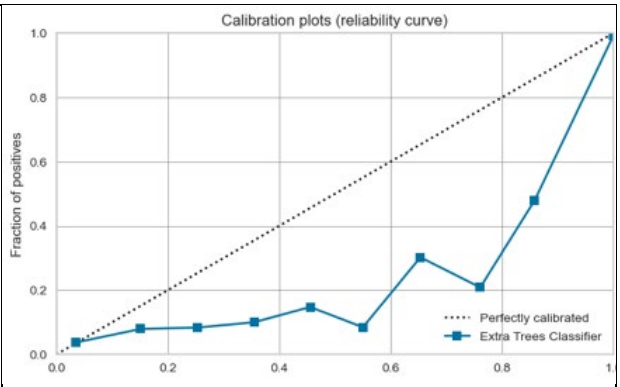


Fig. 21: Calibration Curve

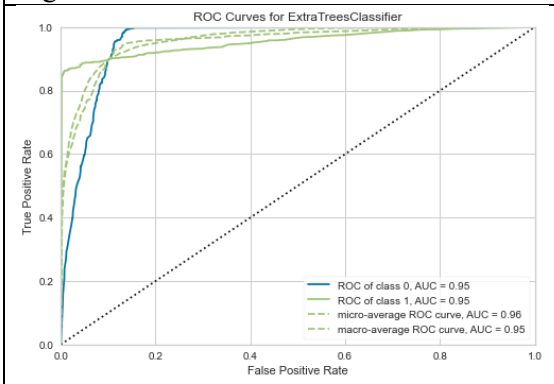


Fig. 22: ROC Area

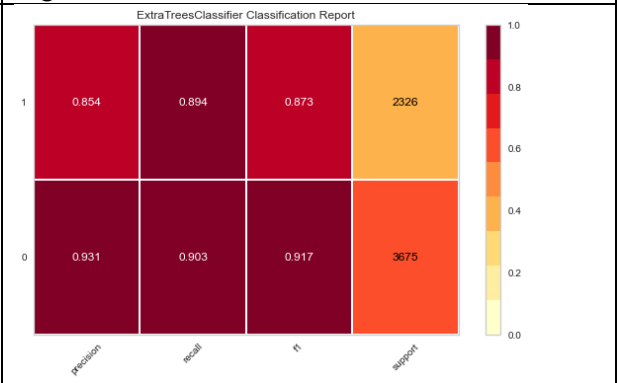


Fig. 23: Class Report

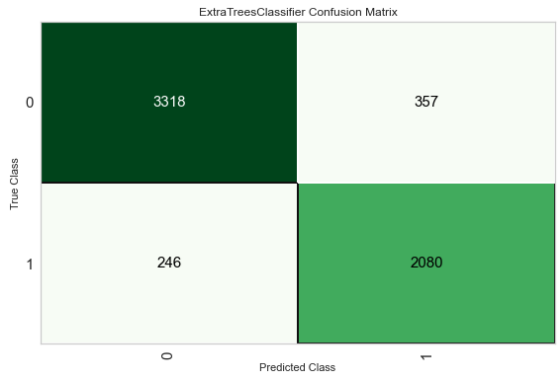


Fig. 24: Confusion Matrix

Table 8 contains the performance measures of the XGBoost model. The bar chart in Figure 25 shows the support (number of training samples) for the non-cyberbullying (0) and cyberbullying (1) classes in the fitted XGBoost classification. The two bars are divided into the proportion of predictions (including FN and FP) for the two classes (0 and 1).

Figure 25 also represents the class prediction error for XGBoost on the number of predicted classes to the actual class. While the XGBoost model seems to be good (comparable to ExF model) at correctly predicting cyberbullying and good at correctly predicting non-cyberbullying. The calibration of the XGBoost model as shown in Figure 26 has a nearly perfect flow indicating quite good accuracy and calibration. Figure 27 shows the plotted True Positive Rate (TPR) and False Positive Rate (FPR) variables in the ROC curve and the response of the XGBoost model for five thresholds. Figure 28 and Figure 29 represent the class report and confusion matrix of the XGBoost classification, respectively.

Both confirm the observation provided by the bars in Figure 25.

Table 8: XGBoost Model Performance Measure

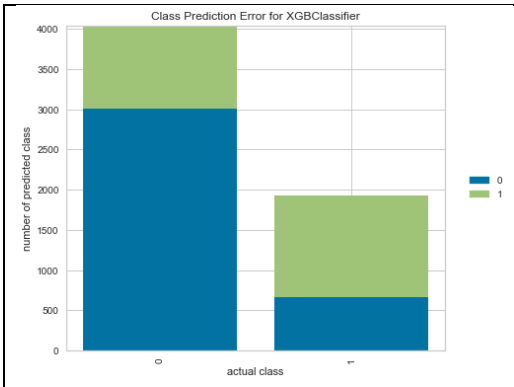


Fig. 25: Class Prediction Error

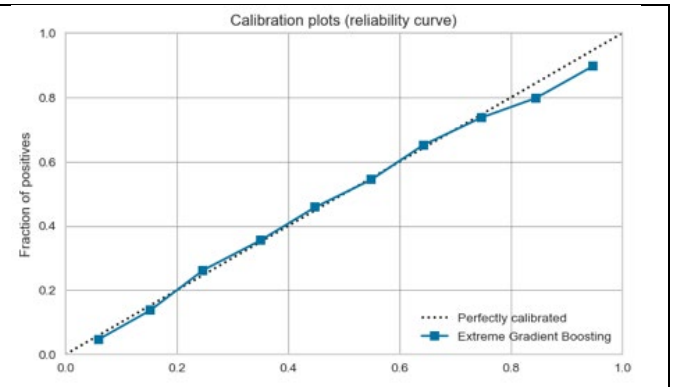


Fig. 26: Calibration Curve

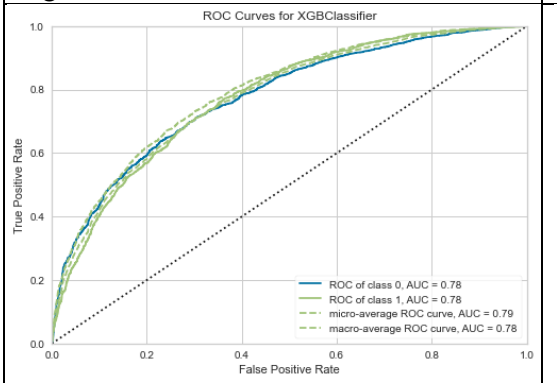


Fig. 27: ROC Area

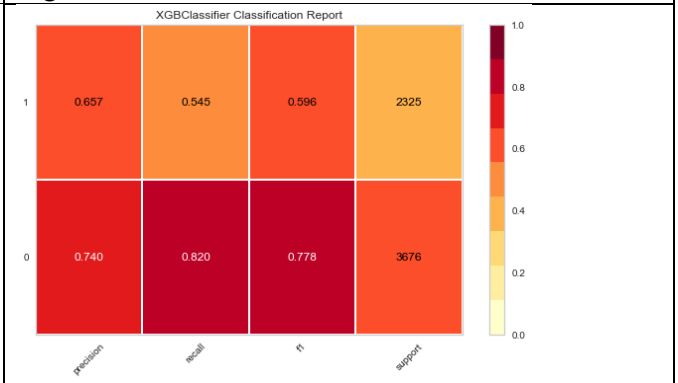


Fig. 28: Class Report

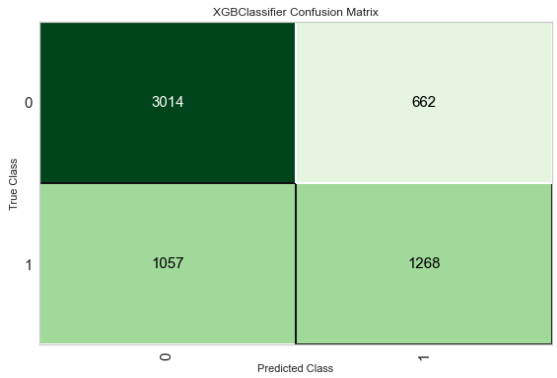


Fig. 29: Confusion Matrix

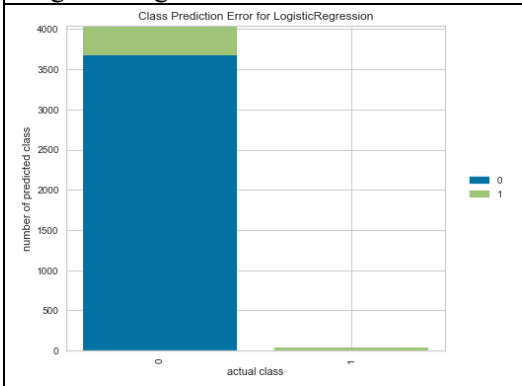
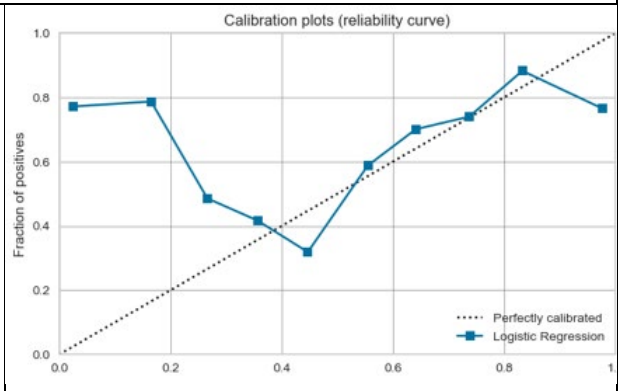
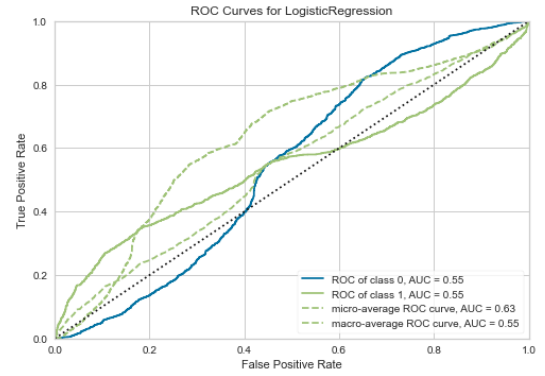
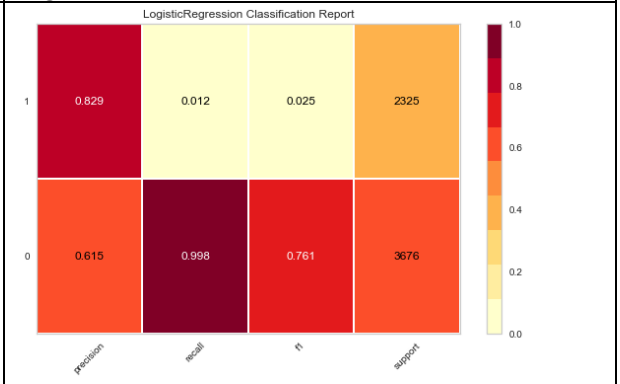
Table 9 contains the performance measures of the Logistic Regression (LR) model. Figure 30 illustrates the class prediction error for LR on the number of predicted classes to the actual class. The bar chart in Figure 30 shows the support (number of training instances) for the non-cyberbullying (0) and cyberbullying (1) classes in the fitted LR classification.

The two bars are divided into the proportion of predictions (thereof FP and FN) for the two classes (0 and 1). While the LR model seems to be quite good (comparable to XGBoost model) at correctly predicting non-cyberbullying and poor at correctly predicting cyberbullying.

Figure 31 represents the calibration of the LR model. It suffers in performance from a low number of samples, but gains quite good accuracy and calibration when the number of samples increases. The plotted True Positive Rate (TPR) and False Positive Rate (FPR) variables in the ROC curve in Figure 32 show the response of the LR model for five thresholds. Figure 33 and Figure 34 illustrate the class

report and confusion matrix of the LR classification, respectively. Both confirm the observation provided by the bars in Figure 30.

Table 9: Logistic Regression Model Performance Measure

Logistic Regression Model	
 Fig. 30: Class Prediction Error	 Fig. 31: Calibration Curve
 Fig. 32: ROC Area	 Fig. 33: Class Report

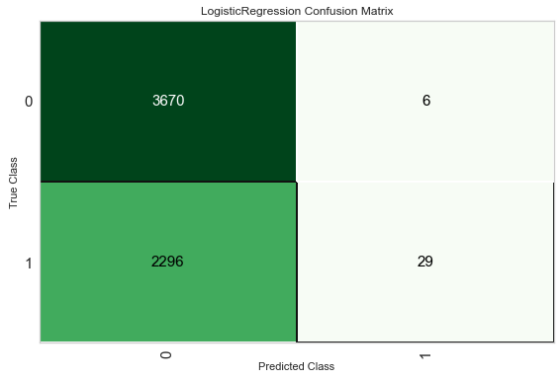


Fig. 34: Confusion Matrix

Table 10 summarizes the achieved performance measures of the applied model during this experiment. Based on all previous results and plots, it can be seen that Extra Tree model has performed with highest performance parameters compared to the other models.

Table 10: Performance summary of the ML model developed

	Naïve Bayes	Decision Tree	Random Forest	Extra Tree	XGBoost	LR
Accuracy	0.64	0.84	0.88	0.9	0.7	0.6
Precision	0.6	0.74	0.8	0.85	0.65	0.82
Recall	0.23	0.9	0.9	0.9	0.54	0.02
F1-Score	0.33	0.81	0.85	0.87	0.6	0.02
ROC-Area	0.66	0.85	0.93	0.95	0.55	0.55

In addition, we make use of the cyberbullying-labeled dataset provided by (Ashraf, Zubiaga, and Gelbukh 2021) on Youtube comments during the year 2016 about the USA election. The dataset has been explored and extended by manual extraction of keywords and key phrases that led to classify the comments as cyberbullying. The extracted keywords provide a distribution, as shown in Figure 35, that illustrates the frequency of words used to formulate cyberbullying.

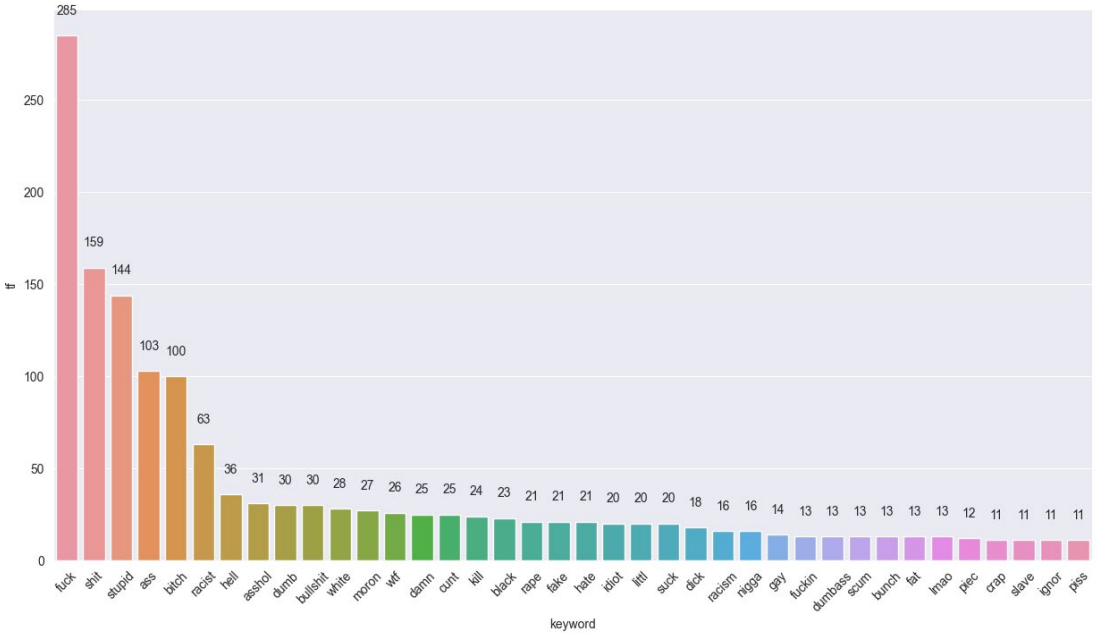


Fig. 35 illustrates the frequencies (>10) of the extracted keywords related to cyberbullying

Similarly, the frequencies of the extracted key phrases have been illustrated, as shown in Figure 36.

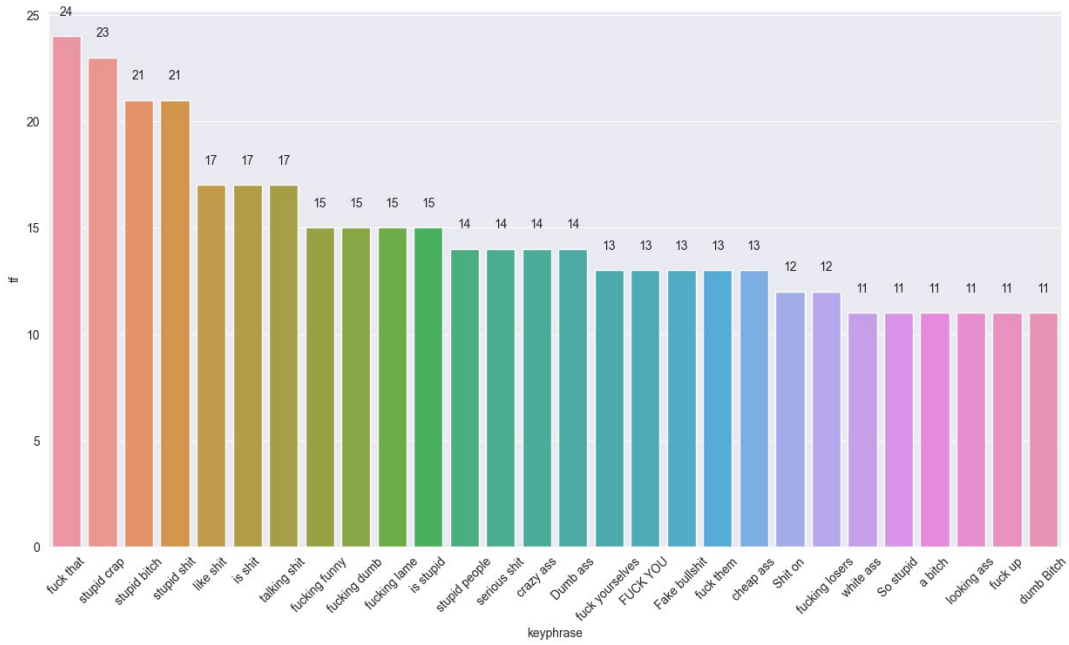


Fig. 36 represents the frequencies (>10) of the extracted key phrases related to cyberbullying

For the key phrases (1106), we applied the language model distilbert-base-uncased provided by huggingface.io to calculate the cosine similarity among all key phrases. The distribution of the similarity scores illustrates Figure 37.

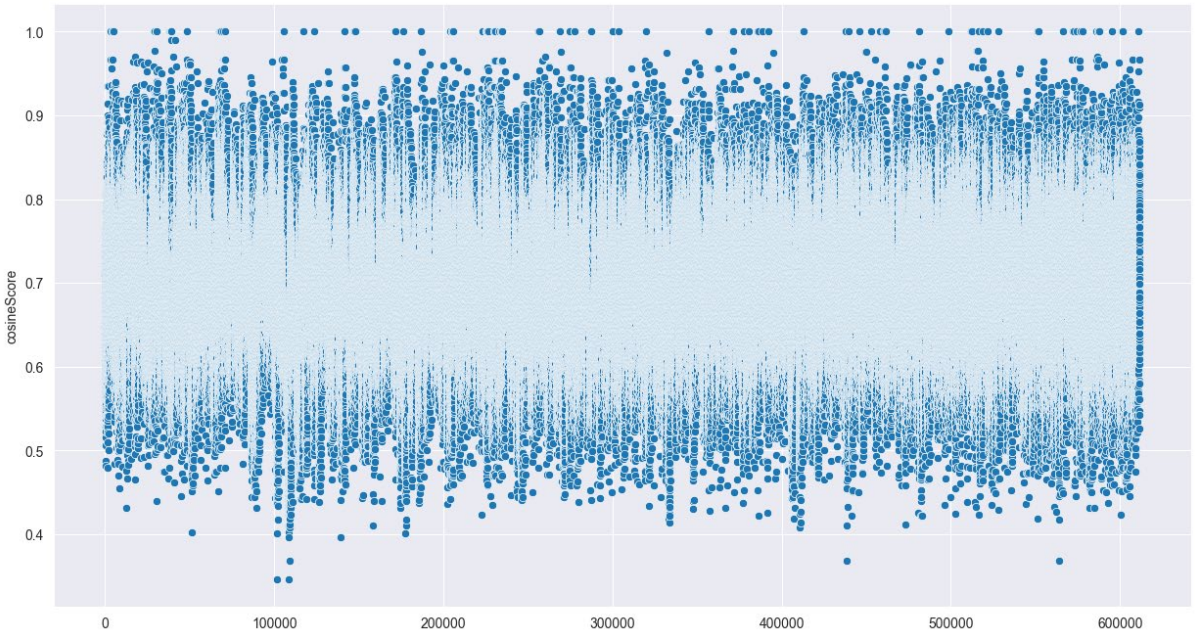


Fig. 37 represents the similarity scores of extracted key phrases

The similarity scores range from 12% to 100%. Table 11 includes the distribution of the key phrases over the scores starting with the similarity score of ≤ 45%.

Table 11 gives an overview on extracted key phrases similarity scores

# of key phrases (sum= 610,034)	x = Similarity score range in %	Examples	
		Key phrase sample 1	Key phrase sample 2
208	≤ 0.45	„fuck the shit out of“	„fiction false“
1654	0.45 < x ≤ 0.5	„you're stupid“	„hammer their sqauare pegs into the round hole”
42876	0.5 < x ≤ 0.6	„human Pig hybrid“	„fuckin hoeeible“
231356	0.6 < x ≤ 0.7	„thug gangster“	„you're stupid“
278246	0.7 < x ≤ 0.8	„knee jerk“	„mentally unstable“
55412	0.8 < x ≤ 0.9	„knee jerk“	„a jerk“
1584	0.9 < x ≤ 0.95	„black kids“	„black guy“
156	0.95 < x < 1.0	„you fools“	„you worthless“
126	=1.0	„pieces of shit“	„Piece of shit“

5. Discussion

Two datasets from twitter and Youtube comments were considered towards different approaches for the same purpose of detecting cyberbullying. Based on Youtube dataset a new dataset has been generated, manually evaluated and made available freely at github for further investigation.

5.1. Twitter dataset

On twitter dataset, different evaluation metrics were used and compared including accuracy, F1 score, ROC-AUC, and recall as summarized in Table 3. Each of these metrics provides different insights into the performance of the models. Accuracy is the proportion of correctly classified instances over the total number of instances, and it provides a general overview of the performance of the model. However, accuracy alone may not be sufficient in situations where the class distribution is imbalanced, or when false negatives or false positives have serious consequences. While, recall, also known as sensitivity, is the proportion of true positive instances that were correctly classified, and is particularly important in situations where false negatives have a high cost. On other hand, the F1-score is a balance between precision and recall and provides a measure of the overall accuracy of any model. It is particularly useful in situations where there is a high cost associated with false negatives or false positives. Finally, the ROC-AUC, or the area under the receiver operating characteristic curve, is a summary of the performance of a binary model, taking into account the trade-off between true positive rate and false positive rate. A high ROC-AUC value indicates a high degree of separation between the positive and negative classes. In terms of current study, recall was considered as the most important evaluation metric, as its target is to minimize the number of cyberbullying instances that could be undetected or false classified. However, depending on the specific requirements and constraints of the problem, a combination of these factors was applied to provide a more comprehensive evaluation of the model in consideration.

This section discusses the algorithms that achieved the highest performance in terms of accuracy, recall and ROC-AUC. Tree based models (Decision Tree in Figures 11-15, Random Forest in Figures 16-20, Extra Tree in Figures 21-25 and XGBoost in Figures 26-30) provided the highest performance and evaluation. The results of the proposed design and implementation indicate that both Extra Tree and Random Forest models performed well in classifying tweets towards the classes cyberbullying and non-cyberbullying, with accuracy scores up to 90% as shown in Figure 21, Figure 25 and Figure 16, Figure 20, respectively.

Furthermore, considering Figure 18 and Figure 23 it can be seen that both models achieved high ROC values; Random Forest and Extra Tree reached 94% and 95%, respectively, within both classes. A ROC value of 95% in Figure 23 of the Extra Tree model indicates its ability to distinguish between cyberbullying and non-cyberbullying tweets very effectively showing a low rate of false positives and

false negatives. On other hand, both tree-based models Decision Tree and XGBoost reached ROC values 85% and 55% in Figure 13 and in Figure 28 respectively, which leads to the statement that Random Forest and Extra Tree are more suitable for this kind of classification issues.

Considering the recall values, which provide information on the number of true positive cases (i.e., cyberbullying cases) that were correctly identified by the model. A high recall value indicates that the model is able to detect a high proportion of the actual cyberbullying incidents. Decision Tree in Figure 14 and Figure 15, Random Forest in Figure 19 and Figure 20, Extra Tree in Figure 24 and Figure 25, and XGBoost in Figure 29 and Figure 30 provide recall values of 90% indicating good performance in detecting cyberbullying incidents. This indicates that the models are able to accurately identify a high proportion of the actual cyberbullying incidents.

The superior performance of Random Forest and Extra Tree compared to XGBoost and Decision Tree models can be attributed to several factors, in which the Random Forest and Extra Tree are ensemble-based learning methods, meaning that they make predictions by combining the outputs of multiple decision trees. This allows for more robust prediction and often results in improved performance compared to a single decision tree. Random Forest uses a random subspace method, where at each split of a tree, a random subset of features is selected from the available features. This helps to prevent overfitting, which is a common problem in decision trees, by adding diversity to the trees in the forest. The Extra Tree method takes the randomization in Random Forest a step further by randomly selecting the threshold for splitting the data at each node. This further increases the diversity of the trees and helps to prevent overfitting. In comparison, XGBoost and Decision Tree models only make use of a single decision tree, which may not be as effective in capturing complex relationships in the data, and are more prone to overfitting. So, the combination of ensemble learning and the randomization methods used in Random Forest and Extra Tree help these models to improve their performance compared to XGBoost and Decision Tree models.

By putting everything together, we can conclude that the Extra Tree model is the best model in our study. The accuracy, recall, and ROC-AUC values of 90% in Figure 24, 90% in Figure 24, and 95% in Figure 23 of the Extra Tree model provide an overall evaluation of the performance of the model on the cyberbullying detection task. With an accuracy of 90% meaning that the model is correctly classifying 90% of the instances. With a recall of 90%, this means that the model is correctly identifying 90% of the actual cyberbullying incidents. A ROC-AUC of 95% indicates that the model is able to distinguish between the positive (cyberbullying) and negative (non-cyberbullying) classes with a high degree of separability. In summary, the Extra Tree model is performing well on the cyberbullying detection task, with high values of accuracy, recall, and ROC-AUC. This suggests that the model is able to accurately classify a large proportion of instances and identify a large proportion of the actual cyberbullying incidents, while also providing a high degree of separability between the positive and negative classes. One more important point is the combination of TF-IDF and PCA. This investigated combination has several benefits:

1) Text representation: TF-IDF is a common method for representing text data, as it takes into account the frequency and importance of individual words in a document. This is a crucial step in converting unstructured text data into a numerical representation that can be used by machine learning algorithms.
2) Dimensionality reduction: PCA is a dimensionality reduction technique that helps to reduce the number of features in a dataset, by identifying and retaining the most important variables. This can help to simplify the data and reduce the computational complexity of the machine learning algorithms applied.
3) Improved performance: By reducing the number of features in the dataset, PCA can help to eliminate noise and irrelevant information that might negatively impact the performance of the machine learning algorithms. It can result in improved accuracy and reduced overfitting.

In the context of our cyberbullying detection system, the combination of TF-IDF and PCA has likely helped to improve the performance of the models by transforming the raw text data into a more manageable representation that can be effectively processed by the machine learning algorithms. Therefore, the combination of TF-IDF and PCA has played a critical role in the success of the developed cyberbullying detection model.

On the contrary, the conducted experiment revealed that Naïve Bayes and Logistic Regression show limited performance as provided in Figures 6-10 and in Figures 31-35, respectively. This result is remarkable since Naive Bayes in particular is generally known for its good performance. However, the various metrics of Naïve Bayes in Figure 6 to Figure 10 and Logistic Regression in Figure 31 to Figure 35 show weak performance compared to the other models.

5.2. Youtube comments dataset

From cyberbullying-labeled Youtube comments dataset, a set of keywords and another set of key phrases have been manually extracted. Figure 36 represents the frequency distribution of the extracted 3,044 keywords (932 unique). Different from (Brandwatch and Ditch The Label 2016), who reported the most used cyberbullying key-terms as (Bitch, Dumb, Fucking, Fat/Ugly, Idiot, Stupid, LMAO, Moron, Piece of shit, Punk, Hate, Nigga), our list of most frequently used cyberbullying keywords are: fuck, shit, stupid, ass, bitch, racist, hell, asshol, dumb, bullshit, white, moron, wtf, damn, cunt etc. as illustrated in Table 12.

Table 12 keywords frequency comparison with list by (Brandwatch and Ditch 2016)

(Brandwatch & Ditch 2016)	*Bitch*	*Dumb*	*Fucking*	*Fat/ Ugly*	*Idiot*	*Stupid*	*LMAO*	*Moron*	*Piece of shit*	*Punk*	*Hate*	*Nigga*
Our finding	*fuck*	*shit*	*stupid*	*ass*	*racist*	*hell*	*asshol*	*dumb*	*bullshit*	*white*	*Moron*	*wtf*

This variation is due to the context and date of the conversation. Depending on the topic keywords frequency varies. In addition, since language changes overtime, especially among youth, the usage of terms on social networks changes accordingly. As a consequence, a fixed list of keywords cannot be established and generally used for detecting cyberbullying.

Figure 37 and Figure 38 represent the key phrases usage frequencies and similarity scores, respectively. Thresholds can be determined, as shown in Table 4, which helps grouping the key phrases together to simplify detection of cyberbullying formulations. The smaller the value, the more interesting the phrases are. Indeed, these represent a true variation of rarely used expressions that have been manually identified. This enrichment can be of great value in supporting machine learning models. The ordering of these expressions can support further investigation in the area of combat or appropriate response. Considering only the expressions with a similarity degree less than or equal to 70%, we obtain a dataset of 276,094 phrase variations distributed over four sections that would be qualitatively and quantitatively sufficient for a deep learning approach. From the remaining sections, a number of phrases can additionally be selected as representative.

The major observations can be summarized as follows:

1) Tree-based models provided the best performance and evaluation.
2) The construction of a single decision tree is not effective in capturing complex relationships that form bullying patterns.
3) No fixed list of keywords can be created and generally used to detect cyberbullying.

6. Conclusion

The number of users spending more and more time online is constantly growing, and the amount of user-created content is very high. Users have neither the same intentions nor the same moral values. As a result, the phenomenon of cyberbullying through electronic messages has emerged. A special type of

user acts as a bully and uses social networks as a platform to attack victims in the form of bullying. In many cases, the consequences are very dramatic and pose a real threat to society as a whole. This results in a great responsibility to find and prevent or combat this content. To this end, a lot of work has been done using different approaches such as rule-based approaches, NLP or DL architectures. The resulting models have performed well, but there is still a need to capture as many types of user-generated content as possible.

The provided investigation applied a comparative study between various ML models. Extra Tree Model demonstrated the overall greatest performance, providing an accuracy of about 90%. The model Naive Bayes, Decision Tree, and Logistic Regression all fared worse than the ensemble techniques (Extra Tree and Random Forest). The least accurate model was the logistic regression, with only 64% accuracy. The absence of designated records and a non-holistic strategy to harassment are two very noteworthy issues that require attention. These two issues together pose significant obstacles to detecting cyberbullying. The inability to evaluate the efficacy of well-known models like support vector machines and multi-layer perceptron (neural network) models was another problem.

This study fills the gap of missing key phrases datasets related to cyberbullying by manually extracting keywords and key phrases used for cyberbullying online. 3,044 keywords (932 unique), 1,106 key phrases as well as similarity comparison among the key phrases have been generated and provided freely for further research. Such new generated datasets can be applied into different approaches towards detection of cyberbullying on social media.

7. Future Research Directions

To improve the performance of the cyberbullying detection model, several approaches can be considered as Feature extraction, other Classification models or Data pre-processing. In terms of feature extraction other techniques besides TF-IDF, such as word embedding (e.g., word2vec, GloVe) or document embeddings (e.g., doc2vec) to capture semantic relationships between words could lead to closer insights. Furthermore, one could consider incorporating additional features such as user information, hashtags, and URLs, which can provide additional context for cyberbullying detection. Classification models that use more advanced ML concepts, such as deep learning models (e.g. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs)) could help classify the tweets. Moreover, ensemble methods such as stacking or voting which could improve the performance of the model by combining the predictions of multiple models could be considered. Techniques such as data normalization, data cleaning, and data augmentation could help to pre-process the data. This could lead to improved performance of the model. Additionally, oversampling or under-sampling could be considered to balance the classes in the data preventing class imbalance issues. These research directions would be out of the scope of our approach and require further investigation. Future work could also consider connection of multiple databases or a greater data platform.

Acknowledgements

This paper is funded by internal seed funding grant no 235055 from Gulf University for Science and Technology (GUST), Kuwait.

References

Abaido, Ghada M. 2020. “Cyberbullying on Social Media Platforms among University Students in the United Arab Emirates.” International Journal of Adolescence and Youth 25 (1): 407–20.

Abarna, S., J. I. Sheeba, and S. Pradeep Devaneyan. 2023. “A Novel Ensemble Model for Identification and Classification of Cyber Harassment on Social Media Platform.” Journal of Intelligent & Fuzzy Systems, no. Preprint: 1–24.

Ahmed, Md Tofael, Almas Hossain Antar, Maqsudur Rahman, Abu Zafor Muhammad Touhidul Islam, Dipankar Das, and Md Golam Rashed. 2023. "Social Media Cyberbullying Detection on Political Violence from Bangla Texts Using Machine Learning Algorithm." Journal of Intelligent Learning Systems and Applications 15 (4): 108–22.

Akhter, Arnisha, Uzzal Kumar Acharjee, Md Alamin Talukder, Md Manowarul Islam, and Md Ashraf Uddin. 2023. "A Robust Hybrid Machine Learning Model for Bengali Cyber Bullying Detection in Social Media." Natural Language Processing Journal 4: 100027.

Alalwan, Ali Abdallah, Raed Salah Algharabat, Abdullah Mohammed Baabdullah, Nripendra P. Rana, Zainah Qasem, and Yogesh K. Dwivedi. 2020. "Examining the Impact of Mobile Interactivity on Customer Engagement in the Context of Mobile Shopping." Journal of Enterprise Information Management.

Alduailaj, Alanoud Mohammed, and Aymen Belghith. 2023. "Detecting Arabic Cyberbullying Tweets Using Machine Learning." Machine Learning and Knowledge Extraction 5 (1): 29–42.

Alzaqebah, Malek, Ghaith M. Jaradat, Dania Nassan, Rawan Alnasser, Mutasem K. Alsmadi, Ibrahim Almarashdeh, Sana Jawarneh, Maram Alwohaibi, Noha A. Al-Mulla, and Nouf Alshehab. 2023. "Cyberbullying Detection Framework for Short and Imbalanced Arabic Datasets." Journal of King Saud University-Computer and Information Sciences 35 (8): 101652.

Ashraf, Noman, Arkaitz Zubiaga, and Alexander Gelbukh. 2021. "Abusive Language Detection in Youtube Comments Leveraging Replies as Conversational Context." PeerJ Computer Science 7: e742.

Bozyiğit, Alican, Semih Utku, and Efendi Nasibov. 2021. "Cyberbullying Detection: Utilizing Social Media Features." Expert Systems with Applications 179: 115001.

Brandwatch, and Ditch The Label. 2016. "Cyberbullying and Hate Speech Online." https://www.ditchthelabel.org/wp-content/uploads/2016/11/Cyberbullying-and-hate-speech.pdf.

Cao, Xiongfei, Ali Nawaz Khan, Ahsan Ali, and Naseer Abbas Khan. 2020. "Consequences of Cyberbullying and Social Overload While Using SNSs: A Study of Users' Discontinuous Usage Behavior in SNSs." Information Systems Frontiers 22: 1343–56.

Casavantes, Marco, Mario Ezra Aragón, Luis C. González, and Manuel Montes-y-Gómez. 2023. "Leveraging Posts' and Authors' Metadata to Spot Several Forms of Abusive Comments in Twitter." Journal of Intelligent Information Systems, 1–21.

Chandrasekaran, Saravanan, Aditya Kumar Singh Pundir, and T. Bheema Lingaiah. 2022. "Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media." Computational Intelligence and Neuroscience 2022.

Fu, Jindi, Rong-An Shang, Anand Jeyaraj, Yuan Sun, and Feng Hu. 2020. "Interaction between Task Characteristics and Technology Affordances: Task-Technology Fit and Enterprise Social Media Usage." Journal of Enterprise Information Management 33 (1): 1–22.

Grover, Purva, Arpan Kumar Kar, Shivam Gupta, and Sachin Modgil. 2021. "Influence of Political Leaders on Sustainable Development Goals–Insights from Twitter." Journal of Enterprise Information Management.

Khairy, Marwa, Tarek M. Mahmoud, Ahmed Omar, and Tarek Abd El-Hafeez. 2023. "Comparative Performance of Ensemble Machine Learning for Arabic Cyberbullying and Offensive Language Detection." Language Resources and Evaluation, 1–18.

Kumari, Kirti, and Jyoti Prakash Singh. 2021. "Identification of Cyberbullying on Multi-modal Social Media Posts Using Genetic Algorithm." Transactions on Emerging Telecommunications Technologies 32 (2): e3907.

Lozano-Blasco, Raquel, Alejandra Cortés-Pascual, and M. Pilar Latorre-Martínez. 2020. "Being a Cybervictim and a Cyberbully–The Duality of Cyberbullying: A Meta-Analysis." Computers in Human Behavior 111: 106444.

Marabelli, Marco, Emmanuelle Vaast, and Jingyao Lydia Li. 2021. "Preventing the Digital Scars of COVID-19." European Journal of Information Systems 30 (2): 176–92.

Muneer, Amgad, and Suliman Mohamed Fati. 2020. "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter." Future Internet 12 (11): 187.

Murshed, Belal Abdullah Hezam, Jemal Abawajy, Suresha Mallappa, Mufeed Ahmed Naji Saif, and Hasib Daowd Esmail Al-Ariki. 2022. "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform." IEEE Access 10: 25857–71.

Rahman, Mushfiqur, Erhan Aydin, Mohamed Haffar, and Uzoechi Nwagbara. 2022. "The Role of Social Media in E-Recruitment Process: Empirical Evidence from Developing Countries in Social Network Theory." Journal of Enterprise Information Management 35 (6): 1697–1718.

Samghabadi, Niloofar Safi, Suraj Maharjan, Alan Sprague, Raquel Diaz-Sprague, and Thamar Solorio. 2017. "Detecting Nastiness in Social Media." In Proceedings of the First Workshop on Abusive Language Online, 63–72.

Silva, Yasin N., Deborah L. Hall, and Christopher Rich. 2018. "BullyBlocker: Toward an Interdisciplinary Approach to Identify Cyberbullying." Social Network Analysis and Mining 8: 1–15.

Wakefield, Robin L., and Kirk Wakefield. 2023. "The Antecedents and Consequences of Intergroup Affective Polarisation on Social Media." Information Systems Journal.

Zhu, Chengyan, Shiqing Huang, Richard Evans, and Wei Zhang. 2021. "Cyberbullying among Adolescents and Children: A Comprehensive Review of the Global Situation, Risk Factors, and Preventive Measures." Frontiers in Public Health 9: 634909.