2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT) | 979-8-3503-0984-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/GCAIOT57150.2022.10019058

A Deep Analysis of Textual Features Based Cyberbullying Detection Using Machine Learning

Md Ishtyaq Mahmud
College of Science and Engineering
Central Michigan University
Mount Pleasant, MI 48858, USA
mahmu4m@cmich.edu

Muntasir Mamun
Department of Computer Science
University of South Dakota
South Dakota, USA
muntasir.mamun@coyotes.usd.edu

Ahmed Abdelgawad
College of Science and Engineering
Central Michigan University
Mount Pleasant, MI 48858, USA
abdel1a@cmich.edu

***Abstract*—Today's internet advancements boost our electronic connectivity to one another through the use of social media platforms. Using social media has facilitated us in many ways, but it has also negatively impacted us. One of the negative repercussions of utilizing social media is cyberbullying, which harms our reputation, privacy, and feelings, or harasses us. Cyberbullying can be controlled by early detection and legal action. By using machine learning and natural language processing (NLP), it is possible to automatically identify tweets, images, and videos that contain offensive language associated with bullying. In this study, we analyzed five distinct machine learning models, including LightGBM, XGBoost, Logistic Regression, Random Forest, and AdaBoost, to detect cyberbullying using the textual feature-based tweeters dataset. We used more than 47,000 tweets from our dataset, which were divided into six classes. We analyzed the machine learning model and observed that LightGBM performed significantly better than other models, reaching accuracy rates of 85.5%, precision rates of 84%, recall rates of 85%, and an F-1 score of 84.49%.**

***Index Terms*—Cyberbullying, Twitter, Machine Learning, Text Classification.**

I. Introduction

Social media networks are a huge aspect of our daily lives today. We use social media in a variety of circumstances, including the entertainment, educational, personal growth, and professional spheres [1]. We depend increasingly on social media in our daily lives due to the internet's and technology's rapid advancements. Since the internet is widely accessible and growing rapidly, all types of people can utilize social media platforms on their smartphones, tablets, and desktops [2]. There are no age restrictions on the majority of social media platforms, which is a poor policy that negatively affects the lives of our children and teenagers [3]. Teenagers are not fully developed due to their limitless use of social media since they prefer online contact over face-to-face interaction with their friends and family [4]. Using social media platforms, people can also exchange vital files, photos, and videos [5]. Among the most popular and extensively used media platforms globally are Facebook, YouTube, Twitter, Instagram, TikTok, WeChat, Telegram, and WhatsApp [6].

When we use social media platforms to exhibit our negative and antisocial behaviors, cyberbullying develops. Anybody, even older citizens, can be a victim of cyberbullying in today's world; it is no longer restricted to any one community or age group [7]. When people disagree with each other's opinions, they may sometimes bully the other person. Bullying is the term for hostile behavior that can be expressed verbally, by text, physically, or in public. Sometimes those people unwittingly intimidate people who don't even know the benefits of social media and how to use it [8]. The most popular social

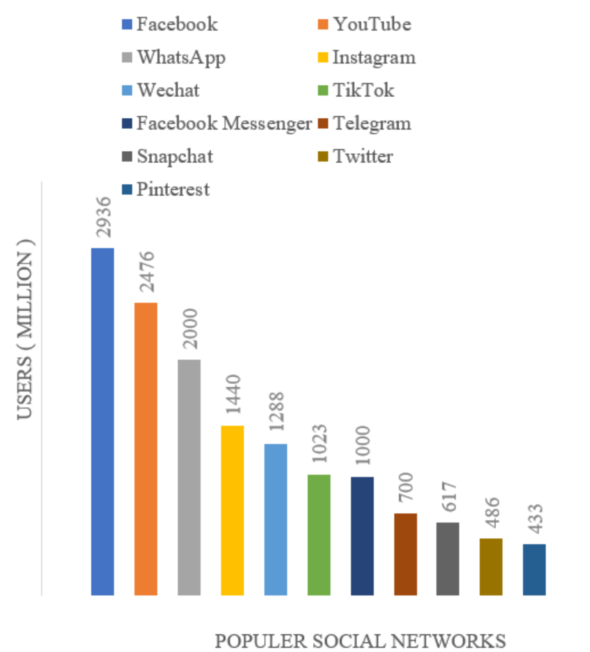


Figure 1: Popular Social Media Platform Globally in 2022

media network globally in 2022 is depicted in Figure 1. In 2022, Facebook had about 2936 million users, making it the most popular social networking platform. In 2022, there will be 7.98 billion people around the world, 5.03 billion internet users, 5.34 billion mobile phone users, and 4.70 billion active social media users [9]. An increase in social media usage and a significant number of cases of cyberbullying have been reported in the USA as a result of the COVID-19 lockdown [10].

Many nations adopt laws and regulations to prevent and restrict cyberbullying [11] [12]. Nevertheless, because it can occur on a variety of social media platforms, it is difficult to identify cyberbullying when laws are applied [13]. Cyberbullying behaviors are intended to intimidate and mock their victims. It also includes things like posting false information on social media, spreading personal videos, spreading private photos, and offensive commentary, among other things [14]. One of the common forms of cyberbullying is a text comment. The most efficient technique to identify cyberbullying and manage these social issues is through machine learning model [15]. Additionally, artificial intelligence (AI), particularly NPL, can be used to stop text-based bullying [16].

The remainder of the article is structured as follows: Previous works are listed in Section II. In Section III, we described the approach and included details on dataset collecting, dataset pre-processing, feature extraction, and algorithm selection. Section IV distributes the machine learning model's performance and results; section V summarizes the findings and offers suggestions for future study.

II. Literature Reviews

The literature review is described in depth in this part, and Table I summarize previous researchers work about cyberbullying detection.

Ali et al. [17] presented different types of classifier of machine learning where SVM classifier performed better consider others classifiers for detecting cyberbullying. They used publicly available dataset for analyzing machine learning model. They also utilized the ensemble approach, although SVM performed better for their dataset, with a detection accuracy of about 79% for cyberbullying on social media.

Alsubait et al. [14] suggested machine learning methods, including Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), and Linear Regression (LR), where LR outperformed taking into account other ML models for identifying cyberbullying on YouTube comments for Arabic users. The dataset was collected from Arabic users' posted video comments between 2015 and 2017. More than 15000 comments with 14 features are included in the dataset. The researchers employed the count vectorizer and the tfidf as two separate approaches for feature extraction. They received a 78.6% F1 score for their LR model after analyzing their presented ML model for identifying cyberbullying.

Muneer et al. [6] used different types of machine learning model namely, random forest (RF), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Naive Bayes (NB), and LR for detecting cyberbullying using unique twitter dataset. They tried their best to obtain the recall, precision, F1 score, and accuracy matrices for the model they were using. They received the highest F1 score and accuracy for LR, 90.57% and 92.8%, respectively, whereas SGD received a precision score of 96.8%.

Dharani et al. [18] suggested two machine learning classifiers, Naive Bayes (NB) and logistic regression (LR), for the automatic detection of bullying in live chat. The dataset was collected from kaggale and is based on textual conversations. The dataset, which includes 2000 question-and-answer exchanges, is based on live social media conversations. The researchers got around 89.79% accuracy for NB classifiers for detecting live chat cyberbullying which is better than other LR classifier.

Balakrishnan et al. [19] proposed two ensemble classifiers such as RF and AdaBoost for detecting personality and emotional based cyberbullying in YouTube comments. The study examined 5151 English-language YouTube comments, with 2576 of those comments falling under the heading of ”bullying” and 2576 falling under the heading of ”non-bullying”. They examined two machine learning (ML) classifiers for identifying YouTube cyberbullying, with RF outperforming the other classifier with a 95% accuracy rate. Dewani et al. [20] presented an LSTM and CNN model for identifying text patterns associated with cyberbullying in Romanian Urdu. They used online available dataset in Roman Urdu language for analyzing their ML models. When compared to the CNN model, the long short-term memory networks (LSTM) ML model performs better for their dataset in terms of identifying cyberbullying. They got 85% accuracy and 70% F1 score for LSTM.

Agarawal et al. [21] proposed traditional machine learning models and deep neural networks for detecting cyberbullying on various social media platforms. They used 3 different online platform for collecting the dataset such as FormSpring, Twitter and Wikipedia. 100k, 12k, and 16k posts in Wikipedia, Twitter, and FormSpring, respectively, are included in the dataset. The BLSTM model, which has an F1 score of 94% over the entire dataset, performs better than the other machine learning models they examined.

Zhang et al. [22] presented machine learning approach, including puncuation-based CNN, SVM, CNN random, and Random forest, to identify cyberbullying using twitter dataset. They have used 23,243 sentences, and1,623, or almost 7%, of them have been identified as harassing tweets on Twitter. Among the machine learning models they examined, a CNN model based on punctuation was suggested. For detecting cyberbullying in the Twitter dataset, after evaluating the PCNN, the results were accuracy of 98.9%, recall of 97%, precision of 99.1%, and F1 score of 98%.

In this article we have analyzed textual feature based Tweeters dataset for detecting cybervullying using five different machine learning models. For our dataset, we used a sizable number of comments from Twitter. More than 47,000 tweets comprise the dataset, which is divided into six different classes: age, religion, gender, ethnicity, not cyberbullying, and others cyberbullying. For feature extraction, text-based features like TF-IDF are used. We applied and analyzed five different machine learning models, including LightGBM, XGBoost, Logistic Regression, Random Forest, and AdaBoost, in order to detect cyberbullying. In order to determine the optimal machine learning model for detecting cyberbullying in the Twitter dataset, we also took into consideration the accuracy, precision, recall, and F1 score matrices. Considering the metrics performance of every machine learning model for detecting cyberbullying, LightGBM perform better than any other models.

Table I: Summary of Literature Reviews.

Reference	Dataset Collection	Models	Performance
[17], 2020	publicly available dataset	Random Forest, SVM(Proposed), Naive Bayes	SVM: accuracy 79%
[14], 2021	15000 Youtube comments dataset (Only Arabic users)	MNB, CNB, and LR(Proposed)	LR: F1 score 78.6%
[6], 2020	37,373 Twitter dataset (Only Arabic users)	SGD, SVM, NB, RF, and LR(Proposed)	LR: accuracy is 90.57%, 92.8% F1 score, SGD: 96.8% precision.
[18], 2022	Kaggle (Live chat Dataset, around 2000 conversation)	NB (Proposed), and LR	NB: accuracy is 89.79%
[19], 2022	5152, Youtube comments	Random Forest(RF) (Proposed) and AdaBoost	RF: 95% accuracy
[20], 2021	Roman Urdu(Social Media Dataset)	CNN, and LSTM (Proposed)	LSTM: 85% accuracy
[21], 2018	FormSpring, Twitter, and Wikipedia dataset	Deep neural network, and Transfer learning model	BLSTM: 94% F1 score
[22], 2016	23,243 Twitters dataset	PCNN (Proposed), Random forests, SVM, CNN random	PCNN: accuracy 98.9%, recall, 97%, precision 99.1%, and F1 score 98%
Our work	47,000 Twitters textual dataset	LightGBM (Proposed), XGBoost, Logistic Regression, Random Forest, and AdaBoost	LightGBM: 85.5% accuracy, 84% precision, 85% recall and F-1 score 84.49%

III. Methodology

Prior to pre-processing, the methodology begins by collecting dataset from easily accessible online sources. Tokenization, text reducing, stop word removal, stemming, feature extraction, and dataset labeling and cleaning are all done during the preprocessing and feature extraction stage. Then, in order to identify cyberbullying, we applied and analyzed five different machine learning models such as LightGBM, XGBoost, Logistic Regression, Random Forest, and AdaBoost. We also have considered the matrics of accuracy, precission, recall and F1 score to identify the best machine learning model for cyberbullying detection in a Twitter dataset. Figure 2 demonstrated the overall procedure of our study.

A. Dateset Collection

The dataset used in this study to identify cyberbullying was gathered from the internet resource Kaggle (www.kaggle.com). The dataset includes more than 47,000 tweets that were gathered during COVID-19. According to UNICEF statistics, 36.5% of people experienced cyberbullying during that time, and 87% had seen it [23]. At that time, social media was the primary means for connecting people worldwide. We use 80% of data for training and 20% of data for testing purposes. The six classifications of the dataset are listed in Table 2, along with the number of tweets for each class.

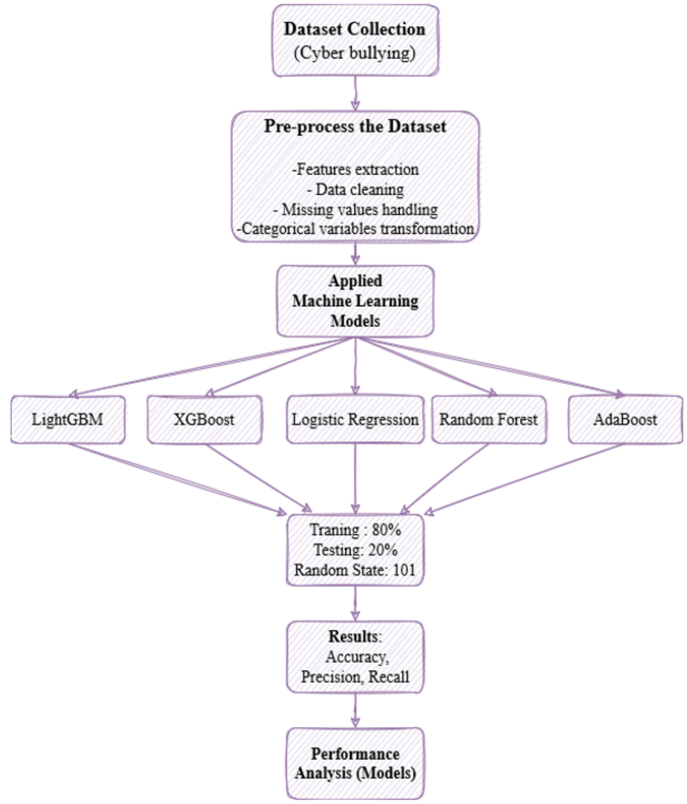


Figure 2: Overall Study

Table II: Dataset

Cyberbullying Types	Values Count
Religion	7998
Age	7992
Gender	7973
Ethnicity	7961
Not Cyberbullying	7945
Other Cyberbullying	7823

B. Dateset Pre-processing and Feature Extraction

Data balance is a significant factor of the preprocessing stage. Approximately 47,000 tweets were labeled into six categories by 7992 tweets, 7973 tweets, 7961 tweets, 7998 tweets, 7945 tweets, and 7823 tweets, based on age, gender, ethnicity, religion, and whether the tweets involved cyberbullying or not. The technique of separating or chopping apart each word that makes up a document or dialogue is known as tokenization,

Figure 3: Images Cluster of Dataset

which is applied after it. Every twitter exchange with English vocabulary has redundant words removed using the Stop World filter. For the purpose of turning a text discussion into simple words, the Porter Stemmer algorithm is used. In order to extract features, text-based features like TF-IDF are used [24]. In order to create a list of features, features from the data are extracted. Figure 3 shows the image cluster of the dataset.

C. Algorithms Selection

While there are many machine learning models for detecting cyberbullying on social media platforms, we have only examined five of them: LightGBM, XGBoost, Logistic Regression, Random Forest, and AdaBoost. We have used same dataset for all of the machine learning model for evaluating the models performance.The metrics of accuracy, precision, recall, and F-1 score are calculated while evaluating the model.

Precision = TP / (TP+FP)

Recall =TP/(TP+FN)

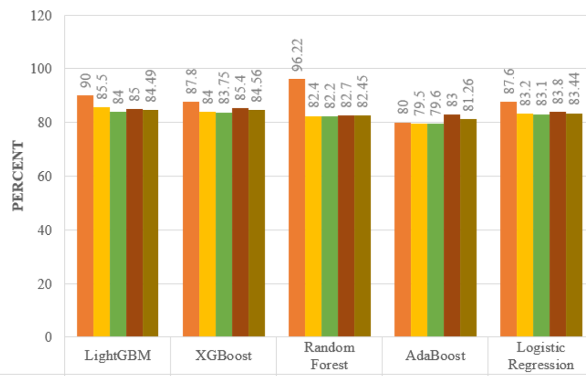
F- 1 Score = 2*(Precision*Recall) / (Precision + Recall)

Where TP = True positive numbers

TN = True negative numbers

FN = False negative numbers

FP = False positive numbers



	LightGBM	XGBoost	Random Forest	AdaBoost	Logistic Regression
Training Accuracy	90	87.8	96.22	80	87.6
Testing Accuracy	85.5	84	82.4	79.5	83.2
Precision	84	83.75	82.2	79.6	83.1
Reccall	85	85.4	82.7	83	83.8
F-1 Score	84.49	84.56	82.45	81.26	83.44

Figure 4: Results of different measures for different machine learning models for detecting cyberbullying.

IV. Results and Discussions

The performance of LightGBM, XGBoost, Logistic Regression, Random Forest, and AdaBoost models was evaluated to see which one performed best at detecting cyberbullying based on Twitter dataset. We evaluated the accuracy, precision, recall, and F-1 score metrics for each machine learning model to determine which one performed the best at detecting cyberbullying on social media. Fig. 4 presents the results of the accuracy, precision, recall, and F-1 score for the performance observation of the models.

For the purpose of evaluating a model's performance, accuracy by itself is insufficient. In order to choose the best model out of the five models for identifying cyberbullying, precision, recall, and F-1 score must also be taken into consideration. A formula that requires precision and recall can be used to determine the F-1 score.

Figure 4 displays the total effectiveness of five various machine learning models based on the particular metrics. After examining every metrics result, we find that all of the models performed essentially similarly, with LightGBM considerably outperforming other models. In addition to outperforming the competition in terms of accuracy, LightGBM also obtained outstanding precision, recall, and F-1 score, which are 84.0%, 85.0%, and 84.49%, respectively. The random forest model has a training accuracy that is around 96.22%, which is better than any other model. Considering the metrics of precision, LightGBM shows better performance which is 84%. LightGBM and

XGBoost performed similarly to one another in terms of recall and F-1 score, scoring about 85% and 84.5%, respectively. When utilizing the Twitter dataset to analyze five machine learning models for the detection of cyberbullying, LightGBM outperformed the other models, and AdaBoost performed the worst of the five algorithms.

V. Conclusion and Future work

The use of a machine learning model has resulted in a sizable amount of work on the topic of cyberbullying for the Twitter dataset. Although in our work, we used a large amount of Twitter's dataset and five different machine learning models, including LightGBM, XGBoost, Logistic Regression, Random Forest, and AdaBoost, to identify cyberbullying. With accuracy of 85.5%, precision of 84%, recall of 85%, and F1 score of 84.49%, LightGBM performed better than any other models when performance measures were taken into consideration. In the future, we might think about using a transfer learning model to analyze the detection of cyberbullying. To analyze the machine learning model, we can also take into account the datasets from various social media platforms.

References

[1] K. Maity, S. Saha, and P. Bhattacharyya, "Emoji, sentiment and emotion aided cyberbullying detection in hinglish," *IEEE Transactions on Computational Social Systems*, 2022.

[2] S. Suleiman, P. Taneja, and A. Nainwal, "Cyberbullying detection on twitter using machine learning: A review."

[3] P. Yi and A. Zubiaga, "Session-based cyberbullying detection in social media: A survey," *arXiv preprint arXiv:2207.10639*, 2022.

[4] R. Bayari and A. Bensefia, "Text mining techniques for cyberbullying detection: state of the art," *Adv. Sci. Technol. Eng. Syst. J*, vol. 6, pp. 783–790, 2021.

[5] R. R. Dalvi, S. B. Chavan, and A. Halbe, "Detecting a twitter cyberbullying using machine learning," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2020, pp. 297–301.

[6] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for cyberbullying detection on twitter," *Future Internet*, vol. 12, no. 11, p. 187, 2020.

[7] M. I. Mahmud, A. Abdelgawad, V. P. Yanambaka, and K. Yelamarthi, "Packet drop and rssi evaluation for lora: An indoor application perspective," in *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*, 2021, pp. 913–914.

[8] M. Arif, "A systematic review of machine learning algorithms in cyberbullying detection: future directions and challenges," *Journal of Information Security and Cybercrimes Research*, vol. 4, no. 1, pp. 01–26, 2021.

[9] "D. chaffey, "global social media statistics research summary 2022 [june 2022]", smart insights, 2022. [online]. available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [accessed: 17- sep- 2022]."

[10] L. M. Al-Harigy, H. A. Al-Nuaim, N. Moradpoor, and Z. Tan, "Building towards automated cyberbullying detection: A comparative analysis," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[11] M. I. Mahmud, A. Abdelgawad, and V. P. Yanambaka, "A deep analysis of hybrid-multikey-puf," *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, 2022.

[12] M. Mamun, S. B. Shawkat, m. S. Ahammed, M. M. Uddin, M. i. Mahmud, and A. M. islam, "Deep learning based model for alzheimer's diseasedetection using brain mri images," *IEEE 13th Annual Ubiquitous Computing Electronics Mobile Communication Conference (UEMCON)*, 2022, (Preprint).

[13] C. Evangelio, P. Rodriguez-Gonzalez, J. Fernandez-Rio, and S. Gonzalez-Villora, "Cyberbullying in elementary and middle school students: A systematic review," *Computers & Education*, vol. 176, p. 104356, 2022.

[14] T. Alsubait and D. Alfageh, "Comparison of machine learning techniques for cyberbullying detection on youtube arabic comments," *International Journal of Computer Science & Network Security*, vol. 21, no. 1, pp. 1–5, 2021.

[15] M. Mamun, M. I. Mahmud, H. md Iqbal, A. M. Islam, M. S. Ahammed, and M. M. Uddin, "Vocal feature guided detection of parkinson's disease using machine learning algorithms," *IEEE 13th Annual Ubiquitous Computing Electronics Mobile Communication Conference (UEMCON)*, 2022, (Preprint).

[16] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 187–193.

[17] A. Ali and A. M. Syed, "Cyberbullying detection using machine learning," *Pakistan Journal of Engineering and Technology*, vol. 3, no. 2, pp. 45–50, 2020.

[18] M. N. Dharani, "Cyberbullying detection in chat application," *Journal homepage: www. ijrpr. com ISSN*, vol. 2582, p. 7421.

[19] V. Balakrishnan and S. K. Ng, "Personality and emotion based cyberbullying detection on youtube using ensemble classifiers," *Behaviour & Information Technology*, pp. 1–12, 2022.

[20] A. Dewani, M. A. Memon, and S. Bhatti, "Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for roman urdu data," *Journal of big data*, vol. 8, no. 1, pp. 1–20, 2021.

[21] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European conference on information retrieval*. Springer, 2018, pp. 141–153.

[22] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, "Cyberbullying detection with a pronunciation based convolutional neural network," in *2016 15th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2016, pp. 740–745.

[23] J. Wang, K. Fu, and C.-T. Lu, "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 1699–1708.

[24] M. Mamun, A. Farjana, M. A. Mamun, M. S. Ahammed, and M. M. Rahman, "Heart failure survival prediction using machine learning algorithm: am i safe from heart failure?" in *2022 IEEE World AI IoT Congress (AIIoT)*, 2022, pp. 194–200.