Cyberbullying Detection: An Investigation into Natural Language Processing and Machine Learning Techniques

Davy Viriya Chow   
Computer Science Department, School of Computer Science  
Bina Nusantara University  
Tangerang, Indonesia  
davy.chow@binus.ac.id

Karli Eka Setiawan  
Computer Science Department, School of Computer Science  
Bina Nusantara University  
Jakarta, Indonesia  
karli.setiawan@binus.ac.idFelicia Natania  
Computer Science Department, School of Computer Science   
Bina Nusantara University  
Tangerang, Indonesia  
felicia.natania@binus.ac.id

Muhammad Fikri Hasani  
Computer Science Department, School of Computer Science   
Bina Nusantara University  
Jakarta, Indonesia  
muhammad.fikri003@binus.ac.idOliverio Theophilus Nathanael  
Computer Science Department, School of Computer Science  
Bina Nusantara University  
Tangerang, Indonesia  
oliverio.nathanael@binus.ac.id

*Abstract*—Cyberbullying has been a concerning issue ever since the Internet and smartphones became very popular. There are several Cyberbullying types such as videos, images, text, and audio, but the most common occurrence of cyberbullying is through hate speech and offensive language via text message. Making it a suitable task for application of Natural Language Processing (NLP). Social media which includes various platforms like Instagram, Facebook, and Twitter plays a big part in this matter, as it is the medium for communication that has often been the home of hate speech. Many researchers have taken an initiative to generate a tool in detecting cyberbullying. This paper provides a comparative study using several deep learning algorithms such as BERT (Bidirectional Encoder Representations from Transformers), Bi-LSTM (Bidirectional Long Short-Term Memory), and Bi-GRU (Bidirectional Gated Recurrent Unit) in detecting tweets containing cyberbullying. Preprocessing is included in this study, followed by tokenization and embedding. The evaluation of these models focuses on accuracy, precision, recall and F1-Score, which measure their ability to correctly classify different types of cyberbullying from a dataset of tweets. Furthermore, this study also used a confusion matrix as part of the evaluation process. In our study, we found that the BERT model outperformed the BiGRU BiLSTM model in terms of accuracy, achieving about 96%. Followed by Bi-LSTM and Bi-GRU obtained accuracy of 95% and 94% respectively. The result of this comparative analysis indicates all three models exhibit strong performance, while the BERT model had a higher ability to correctly classify cyberbullying instances.

Keywords—Cyberbullying detection, Natural Language Processing, Machine Learning, BERT, Bi-LSTM , Bi-GRU

# Introduction

Cyberbullying has become a concern in recent years, the use of technology and social media has spread across the world due to technological advancement and need these past decades. Few Research on December 15, 2022 shows that nearly half (46%) of teens at the age of 13-17 have been a victim of cyberbullying [1]. This form of bullying takes place through digital platforms such as text messaged, emails, tweets, and many other forms on the internet. This kind of bullying requires the use of technology to intimidate other people, this could lead to mental health issues and threatening social relationships. According to JAMA Network Open on 2021 children between the ages of 10 and 16 who accessed photos of cyberbullying had up to a 50% higher risk for thoughts of suicide [2]. During the COVID-19 lockdown there has been a 70% increase in the amount of hate speech among children and teens by DigitalTrends.com [3]. Other than that emotional impact on the victim often leads to depression and anxiety this can occur among any age and contains threats, insults that intend to harm other people on the internet. In 2020, a poll released by UNICEF stated that 2,777 Indonesian teenagers aged 14-24 revealed that 45% reported they had experienced cyberbullying [4]. Cyberbullying can come through many types such as text, images, or even videos.

The most common textual occurrence is hate speech and offensive language. Using the application of Natural Language Processing (NLP) researchers have developed. It is required to use a robust NLP model because to detect hate speech we also need to understand the context and overall intention.

This paper provides a comparative study using several deep learning algorithms. This research used a dataset from kaggle as a training data for our models [5]. After gathering the data, it will be used to train deep learning models such as Bidirectional Long Short-Term Memory (LSTM) and also used to fine tune a pre-trained Bidirectional Encoder Representations from Transformers (BERT). The results were then examined using several classic classification evaluations such as accuracy, precision, recall, and F1-Score.

Cyberbullying detection is a crucial issue to be handled however there are only a few discussions related to this topic by harnessing the potential of the newest deep learning algorithm let alone fine tuning the state-of-the-art model for that task. This paper contributes by giving a comparison review of some of the state-of-the-art natural language classification models in the field of cyberbullying detection. Additionally, it also offers an insight on improving the NLP model for cyberbullying tasks.

The rest of the paper is organized as follows: Section 2 explains the related works developed for cyberbullying detection. Section 3 describes the methodology of the proposed work. Section 4 lists the experimental results obtained and comparison of their performances; finally, Section 5 concludes the contributions and future prospects.

# Literature Review

Cyberbullying has been frequently discussed in previous papers using different kinds of approaches and experimental methods. This issue has been experimented using machine learning algorithms such as Support Vector Machine (SVM), Naïve Bayes, XGBoost, and Logistic Regression [6]. Similarly, a paper also compared those traditional method results against Bidirectional RNNs and attention-based models and concluded that the newer models are going to bring further advances in cyberbullying detection. Bidirectional long short-term memory (BiLSTM) and CNN were also utilized by certain researchers to compare with a single BiLSTM in terms of their capacity to categorize social media posts into different categories of bullying. The performance of both classifiers was encouraging, with BiLSTM beating the CNN-BiLSTM model when used together [7]. By combining three deep learning architectures (a multichannel architecture made up of BiGRU, a transformer block, and CNN models), Munif Alotaibi et al. proposed a new method for detecting cyberbullying that outperformed state-of-the-art techniques [8]. Amgad Muneer et al. also conducted a comparative analysis to review the feature extraction methods TF-IDF and Word2Vec, as well as a number of conventional machine learning models, are compared by the cyberbullying machine learning classifiers [9]. This study also concluded that Logistic Regression techniques perform better as the data increases and obtains best prediction time compared to other classifiers used in this research.

This kind of research topic has also been researched by Md Manowarul Islam et al.’s paper, which discussed the automatic detection of social media posts using Bag-of-Word and TF-IDF as features and Decision Tree (DT), Random Forest, Support Vector Machine, Naive Bayes as There are four machine learning algorithms that can detect cyberbullying in comments on Facebook and Twitter [10]. Research has been conducted in 2022 to examine five different machine learning models. Results revealed that LightGBM performed better than other models while AdaBoost did the least well of the five algorithms [11]. The results show that the Logistic Regression technique produced the best F1 score and accuracy. When compared to other classifiers employed in this study, logistic regression techniques perform better as the amount of data grows and achieve the best prediction time.

A study of cyberbullying detection in Turkish tweets also compares the performance of nineteen different machine learning algorithms and acquired the best result from the Light Gradient Boosting Model (LGBM) [12]. Md Manowarul Islam et al.’s paper discussed automated identification of posts on social media using Bag-of-Word and TF-IDF as features and Decision Tree (DT), Random Forest, Support Vector Machine, Naive Bayes as Four Machine Learning Algorithm used to identify cyberbullying text on Facebook and Twitter comments [13]. John Hani et al. use tokenization, lowering text, stop words, encoding cleaning and word correction for the pre-processing step. Following the features extraction step using TF-IDF, the classification step used two different classifiers which are SVM and Neural Network. Both models receive high accuracy, and results show that NN outperforms SVM in terms of average accuracy and average f-score. [14].

A paper; aim to develop cyberbullying detection for English and Hindi with several machine learning algorithms combined with Count Vectorizer and TF-IDF for the features extraction process. This research found that the Decision Tree algorithms achieved the highest accuracy among other algorithms but had the worst training and prediction time. Additionally, the Count Vectorizer provides slightly better accuracy than TF-IDF [15]. In Andrea Perera et al. research aims to improve the word embedding for this topic as discussed in Andrea Perera et al. paper which claims that by using TF-IDF + Sentiment analysis as a feature can improve cyberbullying detection identification tasks [16].

In their research, Akshi Kumar et al. suggested a Bi-GRU-Attention-CapsNet (Bi-GAC) model to identify bullying in online language that gets superior F1-scores of 94.03 and 93.89 for the benchmark cyberbullying datasets from Formspring.me and MySpace, respectively. In this study, self-attention-based Bi-GRU encoder and capsule network benefits are combined, and ELMo contextual embeddings are employed as input [17]. On the majority of deep learning models, the ELMo word embedding performs better than alternative combinations of word embeddings, as stated in the Tina Yazdizadeh et al. research [18]. Additionally, Nabi Rezvani et al. suggested an attention-based model that enriches textual data by including contextual features and uses a state-of-the-art text classification algorithm [19]. This study demonstrates how LSTM and BERT-based models beat baseline approaches on practically all measures. By adopting modified BERT, Sayanta Paul and Sriparna Saha hope to improve its performance on tasks including the identification of cyberbullying. According to the experimental findings, the suggested BERT model performs quite well when compared to other machine learning models [20].. In Hind Saleh et al. paper, also make use of various word embedding methods as an option to optimize a BERT model [21].

In the realm of classic machine learning, it is shown that Random Forest performs better than SVM at 98.5% accuracy for identifying cyberbullying using over 16 thousand tweets [22]. Others found that Decision Tree and SVC combined with BoW perform the best when used to classify cyberbullying in Urdu [23].With the rise of transformers as the state-of-the-art method in Natural Language Processing it is also shown that transformers such as dBERT yield a better result even with imbalanced data compared to classic machine learning methods such as SVM [24]. Unlike classic machine learning, BERT won’t show a significant improvement when oversampling such as SMOTE is applied to the data [25]. BERT also proved to be able to locate the hate word because of its attention-based architecture [26]. Mengfan Yao et al. research also brings a new approach they introduce, CONcISE (Cyberbullying detectiON on Instagram media SEssions) and consistently outperforms State–of–the–art detection method when compared [27].

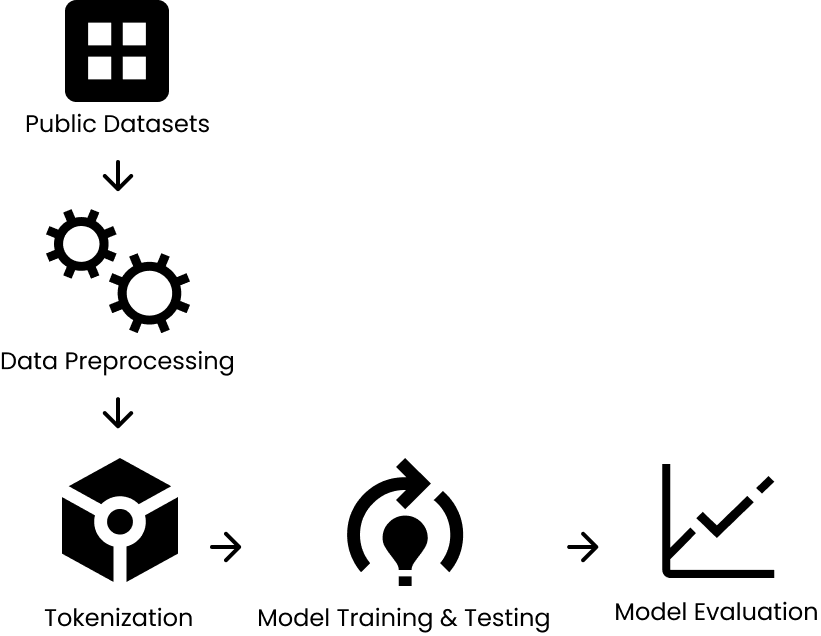
LSTM is one of the most popular and arguably the most powerful recurrent neural networks because of its ability to handle long sequences, it is proven that LSTM performs better than SVM when utilized to classify cyberbullying in the Bangla language [28]. Its bidirectional counterpart offers an even better evaluation of a long sequence by reading the sequence forward and backwards. Most often BiLSTM is compared with 1D convolution as a great alternative to evaluate long sequences or even detect cyberbullying. It is shown that BiLSTM performs slightly better than 1D convolution but 1D convolution is trained 65x faster with cyberbullying classification tasks [29]. However, both methods still underperformed compared to pre-trained BERT that were later trained for the same task [30]. In A Rishab Vanigotha et al. paper also proposed a new model called SparkNLP which uses a Deep Neural Network structure and achieved better results when compared to the BiLSTM model [31].

Other exploration including trying to detect cyberbullying in languages other than English, for this task often BERT is utilized for this task. Making an ensemble of mBERT and Indo-BERT has proven to be effective in detecting cyberbullying in a text mixed between Hindi and English [32]. Arabic language also has its own BERT counterpart such as AraBERT that yields a good performance but it later found that it is better to make an ensemble consist of AraBERT, AraELECTRA, Albert-Arabic, AraGPT2, mBERT, and XLM-RoBERTa using majority voting can yield an even better result [33]. Another ensemble approach was used in the research by Kazi Saeed Alam et al. to build voting-based models using Single Level Ensemble Model (SLE) and Double Level Ensemble Model (DLE) architecture, respectively. These models were built using Multinomial Naive Bayes (MNB), Logistic Regression (LR), Decision Tree Model (DT), Linear Support Vector Classifier (LSVC), Gradient Boosting Classifier (GBoost), and AdaBoost (AdB) [34].

Even though classification usually uses supervised learning techniques, Lu Cheng et al. proposed unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model [35]. Another variation in cyberbullying detection is to detect image-based cyberbullying. It is found that using pre-trained 2D CNN models such as InceptionV3 yields the best image even though it still can't classify a screenshotted cyberbullying text [36].

For this study, we compared the performance of BERT and LSTM in detecting tweets containing cyberbullying.

# Methodology

**

1. Workflow for Cyberbullying Classification model

Overall the methodology of this research was done as depicted in Fig. 1.

## Dataset

Dataset used in this experimental research is obtained from Kaggle which contains 47693 tweets data. These tweets consist of different types of cyberbullying types, such as age-related cyberbullying, ethnicity-related cyberbullying, gender-related cyberbullying, religion-related cyberbullying, other types of cyberbullying, and tweets that did not contain cyberbullying. Types of these cyberbullying tweets are listed in Table I. The dataset was later split into 8:1:1 for training data, validation data, and testing data using stratified shuffle split techniques.

1. Dataset Class

| **Cyberbullying type** | **Tweet text** |
| --- | --- |
| Age | 7992 |
| Ethnicity | 7961 |
| Gender | 7973 |
| Not\_cyberbullying | 7945 |
| Other\_cyberbullying | 7823 |
| Religion | 7998 |

## Preprocessing

In this research paper, the label from the dataset is a categorical feature which is still a string type. The labels need to be converted to numerical values by using an ordinal encoder. An ordinal encoder is a technique which can be used to convert categorical variables with an inherent order or ranking into numerical values. Padding the list of tweets text is also needed since the length of the text differs one from another before the text is tokenised.

## Tokenization

Tokenization is the process of breaking down a text or sequence into smaller units called tokens. Tokenization plays a crucial role in the text classification task. Tokenization can be done in various ways. This research used one of the TensorFlow libraries to do the tokenization task on the tweets dataset. This is done by simply giving a number to each word within the text. At this preprocessing step, there is still no semantic information engrained on the token itself.

## Embedding

Embedding is a technique used to extract the semantic information of the dataset which is an essential task for the model to have better accuracy in classifying sentiment analysis such as this research topic. This embedding converts a tokenized word to a vector representation, by bringing the numerical value to higher dimensions, it enables the model to learn a similar word or even words that may or may not have a relation to each other. In this research, we set the embedding as a learnable vector representation rather than using pre-trained embeddings.

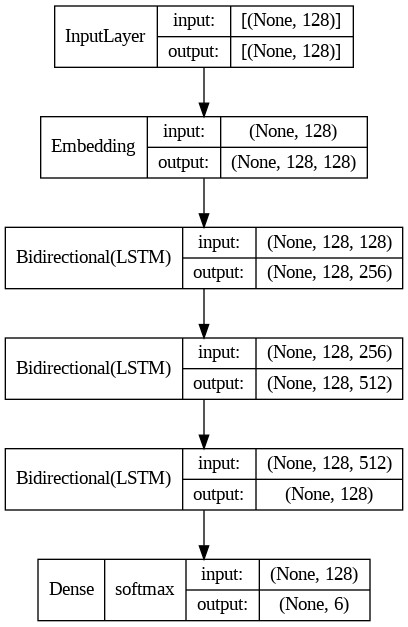
## BERT

BERT is a linguistic model, this model is presented by scientists from the Google AI Language laboratory in 2018. The deep preliminary learning of bidirectional text representation is the purpose of this model. This architecture is based on a multilayer bidirectional transformer, it uses text embeddings to represent an input sequence. The BERT model, which enables further training of the fundamental model for a particular task with just one extra layer of neurons, is also utilized for text categorization tasks. After more training using a specific approach and BERT, creating models with the current top performance in text classification challenges is feasible.

## LSTM and Bi-LSTM

Long Short-Term Memory (LSTM) were first proposed by Hochreiter and Schmidhuber in 1997. LSTM is a sequential neural network. LSTM offers a solution for the inability of basic RNNs to process long sequences without losing much of the information from the early words in this case. LSTM is a module that incorporates gating and splitting information and feeds a part of them to a fully connected layer. Unlike Basic RNN which only outputs one value, LSTM yields 2 outputs which are the long-term information and the short-term information.

Bidirectional Long Short-Term Memory (Bi-LSTM) is a deep learning model that is commonly used for sentiment analysis. It has two inputs consisting of forward and backward input, as a result, each input sequence may be used by Bi-LSTM to learn past and future information. This model is a derivative variant of the LSTM model. To process a sentence it is clear that it is better to evaluate a sentence forward or backwards to have a better understanding of the overall meaning of the text. This way of evaluating a text is also performed by humans without realizing it, we as a human often re-read previous words to understand the text, this behaviour corresponds to the backward evaluation of Bi-LSTM

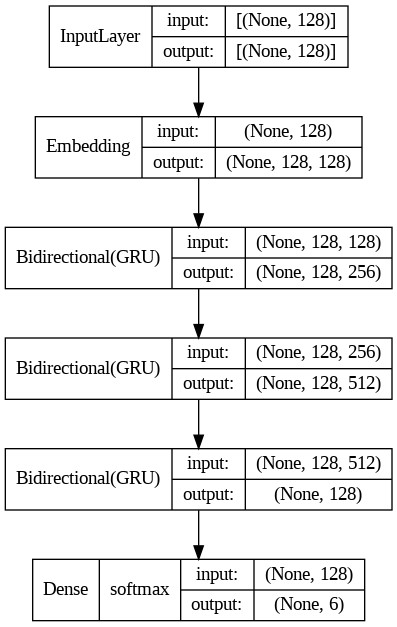


1. Bi-LSTM model architecture

Fig. 2 shows the architecture for the Bi-LSTM model. The model consists of an Embedding layer, 3 layers of Bi-LSTM, and a dense layer with softmax activation as the classifier head. From this diagram, it is clear to see that this model is a sequence-to-vector model.

## Bidirectional-Gated Recurrent Unit

Bidirectional-GRU (Bi-GRU) is a decoder that models the information flow in one direction. Its structure is a combination of two separate directional GRUs. Fig. 3 shows the overall architecture for the Bi-GRU model. To yield an objective comparison and minimize the bias, the authors build the Bi-GRU model as similar as possible to the Bi-LSTM, it also consists of Embedding layer, 3 layers of Bi-Gru and a classifier head with softmax activation.



1. Bi-GRU model architecture

## Metrics

Accuracy, precision, recall, and F1-Score are performance metrics for evaluating the classification models.

Accuracy is arguably the most well-known classification metric, it measures how often the classifier model makes correct predictions

|  |  |
| --- | --- |
|  | (1) |

Precision measures how much confidence the model offers given the model outputs a true label. It is known to be the ratio of an expected positive result to any anticipated positive data.

|  |  |
| --- | --- |
|  | (2) |

Recall is a metric that measures the ability of the model to detect a true date and assign it with a true label. It represents the ratio of all correct positive prediction data with all available positive data.

|  |  |
| --- | --- |
|  | (3) |

F1-Score is a harmonic mean between precision and recall hence it provides a fair assesment of precision and recall to find the sweet spot with the best combination of precision and recall.

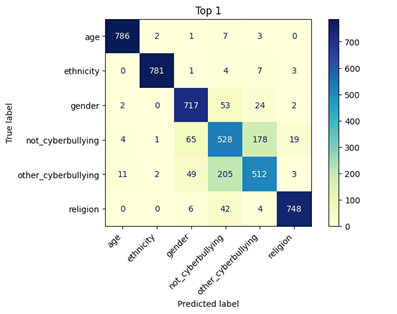
|  |  |
| --- | --- |
|  | (4) |

# Result and Discussion

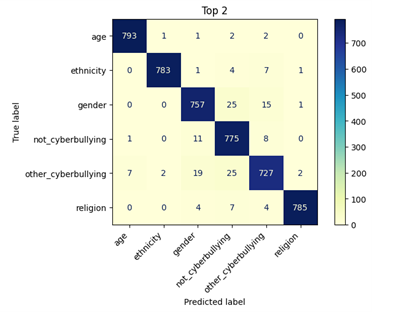
Result of this research on cyberbullying classification topic using models like BERT, Bi-LSTM, Bi-GRU.

1. Model Evaluation

| Model | Metric | | | |
| --- | --- | --- | --- | --- |
| Accuracy | Precision | Recall | F1 Score |
| BERT (Top 1) | 0.85367 | 0.8541 | 0.8529 | 0.8531 |
| BERT (Top 2) | 0.9686 | 0.9689 | 0.9684 | 0.9685 |
| Bi-LSTM (Top 1) | 0.8252 | 0.8304 | 0.8248 | 0.8171 |
| Bi-LSTM (Top 2) | 0.9524 | 0.9527 | 0.9523 | 0.9524 |
| Bi-GRU (Top 1) | 0.8111 | 0.8198 | 0.8099 | 0.8114 |
| Bi-GRU (Top 2) | 0.9465 | 0.9491 | 0.9462 | 0.9464 |



1. Confusion Matrix for BERT (Top 1) model



1. Confusion Matrix for BERT (Top 2) model.

As shown in Table II, the BERT model showed the best results with F1-Score with 0.9685 as the best model and overall model score from the model top 2 which takes account precision as well as recall. The table also shows the results from the other models, which results are worse when compared to the BERT model. However, the disparities were relatively minor.

Confusion metrics from the models also shown on the figure Fig. 4 to Fig. 9. From the figures, concluded that the prediction results are sufficiently positive as seen from the confusion metrics, the most common predicted values align with the diagonal of the metrics, which means correct classes prediction is made.

# Conclusion

In conclusion, the use of natural processing models such as BERT, Bi-LSTM, and Bi-GRU shows great performance in classifying cyberbullying tweets. When these results were compared to each other, the BERT architecture model showed the best performance from the model evaluation score such as the Accuracy score and F1-Score. The accuracy score of these models showed good results in classifying cyberbullying types from tweets dataset. This research also concluded that using the BERT model gives the best results in this particular tweets cyberbullying dataset. The BERT architecture indicates a better balance between precision and recall for the BERT model.

The results of this research can be considered as a promising result since it hasn't gone through a parameters-tuning process. This shows that classification in this topic still can be improved using various ways, one of which is parameters tuning these models to further improve the performance of the proposed models. Future works should concentrate on improving these models, and expanding the datasets to include wider and more diverse linguistic contexts. Additionally, optimizing more advanced models by fine-tuning to boost the performance in classifying cyberbullying tweets

##### References

[1] “Teens and Cyberbullying 2022”.

[2] S. Arnon et al., “Association of Cyberbullying Experiences and Perpetration with Suicidality in Early Adolescence,” JAMA Netw Open, vol. 5, no. 6, Jun. 2022, doi: 10.1001/jamanetworkopen.2022.18746.

[3] “Rising Levels of Hate Speech & Online Toxicity During This Time of Crisis.”

[4] “BULLYING IN INDONESIA: Key Facts, Solutions, and Recommendations.” [Online]. Available:https://www.liputan6.com/news/read/2361551/mensos-bunuh-diri-anak-indonesia-40-persen-karena-bullying

[5] 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, Dec. 2020, pp. 1699–1708. doi: 10.1109/BigData50022.2020.9378065.

[6] C. Raj, A. Agarwal, G. Bharathy, B. Narayan, and M. Prasad, “Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques,” Electronics, vol. 10, no. 22, p. 2810, Nov. 2021, doi: 10.3390/electronics10222810.

[7] T. H. H. Aldhyani, M. H. Al-Adhaileh, and S. N. Alsubari, “Cyberbullying Identification System Based Deep Learning Algorithms,” Electronics (Switzerland), vol. 11, no. 20, Oct. 2022, doi: 10.3390/electronics11203273.

[8] M. Alotaibi, B. Alotaibi, and A. Razaque, “A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media,” Electronics, vol. 10, no. 21, p. 2664, Oct. 2021, doi: 10.3390/electronics10212664.

[9] 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, Oct. 2020, doi: 10.3390/fi12110187.

[10] Islam, Manowarul & Uddin, Md Ashraf & Islam, Linta & Akhter, Arnisha & Sharmin, Selina & Acharjee, Uzzal. (2021). Cyberbullying Detection on Social Networks Using Machine Learning Approaches. 10.1109/CSDE50874.2020.9411601.

[11] M. I. Mahmud, M. Mamun, and A. Abdelgawad, “A Deep Analysis of Textual Features Based Cyberbullying Detection Using Machine Learning,” in 2022 IEEE Global Conference on Artificial Intelligence and Internet of Things, GCAIoT 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 166–170. doi: 10.1109/GCAIoT57150.2022.10019058.

[12] E. Cihan Ates, E. Bostanci, and S. Güzel, “Comparative Performance of Machine Learning Algorithms in Cyberbullying Detection: Using Turkish Language Preprocessing Techniques.”

[13] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, “Cyberbullying Detection on Social Networks Using Machine Learning Approaches,” in 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/CSDE50874.2020.9411601.

[14] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, “Social media cyberbullying detection using machine learning,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 703–707, 2019, doi: 10.14569/ijacsa.2019.0100587.

[15] K. Shah, N. Mehendale, E. Engineering, and K. Rajpara, “Cyber bullying detection for Hindi-English language using machine learning Chaitanya Phadtare.” [Online]. Available: <https://ssrn.com/abstract=4116143>

[16] A. Perera and P. Fernando, Accurate Cyberbullying Detection and Prevention on Social Media. 2021.

[17] Kumar, A., Sachdeva, N. A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. World Wide Web 25, 1537–1550 (2022).

[18] T. Yazdizadeh and W. Shi, “Performance Evaluation for the use of ELMo Word Embedding in Cyberbullying Detection,” in Data Science and Machine Learning, Academy and Industry Research Collaboration Center (AIRCC), Sep. 2022, pp. 131–144. doi: 10.5121/csit.2022.121511.

[19] Nabi Rezvani and Amin Beheshti, “Towards attention-based context-boosted cyberbullying detection in ...” [Online]. Available: https://rintonpress.com/xjdi2/xjdi2-4/418-433.pdf. [Accessed: 19-Mar-2023].

[20] S. Paul and S. Saha, “CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification,” in Multimedia Systems, Springer Science and Business Media Deutschland GmbH, Dec. 2022, pp. 1897–1904. doi: 10.1007/s00530-020-00710-4.

[21] H. Saleh, A. Alhothali, and K. Moria, “Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model,” Applied Artificial Intelligence, vol. 37, no. 1, Dec. 2023, doi: 10.1080/08839514.2023.2166719.

[22] Stephen Afrifa and Vijayakumar Varadarajan, “Cyberbullying Detection on Twitter Using Natural Language Processing and Machine Learning Techniques”, ijitis, vol. 5, no. 4, pp. 1069–1080, Dec. 2022.

[23] S. Khan and A. Qureshi, “Cyberbullying Detection in Urdu Language Using Machine Learning,” in 2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/ETECTE55893.2022.10007379.

[24] H. R. Kirk, Bertie Vidgen, and S. A. Hale, “Is More Data Better? Re-thinking the Importance of Efficiency in Abusive.” [Online]. Available: https://arxiv.org/pdf/2209.10193v1.pdf. [Accessed: 19-Mar-2023].

[25] H. H.-P. Vo, H. T. Tran, and S. T. Luu, “Automatically Detecting Cyberbullying Comments on Online Game Forums,” Jun. 2021, doi: 10.1109/RIVF51545.2021.9642116.

[26] S. Das, P. Mandal, and S. Chatterji, “Probabilistic Impact Score Generation using Ktrain-BERT to Identify Hate Words from Twitter Discussions,” Nov. 2021.

[27] M. Yao, C. Chelmis, and D.-S. Zois, “Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media,” in The World Wide Web Conference, New York, NY, USA: ACM, May 2019, pp. 3427–3433. doi: 10.1145/3308558.3313462.

[28] M. F. Ahmed, Z. Mahmud, Z. T. Biash, A. A. N. Ryen, A. Hossain, and F. Bin Ashraf, “Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language,” Jun. 2021.

[29] Ghosh, S., Chaki, A., & Kudeshia, A. (2021). Cyberbully Detection Using 1D-CNN and LSTM. Lecture Notes in Electrical Engineering, 728 LNEE. [https://doi.org/10.1007/978-981-33-4866-0\_3](https://doi.org/10.1007/978-981-33-4866-0_37)7

[30] F. Elsafoury, S. Katsigiannis, S. R. Wilson, and N. Ramzan, “Does Bert Pay attention to cyberbullying?,” ACM Conferences, 01-Jul-2021. [Online]. Available: <https://dl.acm.org/doi/10.1145/3404835.3463029>.

[31] A. R. Vanigotha, M. R. N. Kumar, S. Hiremath, S. S. Adityan, and Dr. M. J. Basha, “Effective Cyberbullying Detection with SparkNLP,” Int J Res Appl Sci Eng Technol, vol. 11, no. 3, pp. 101–106, Mar. 2023, doi: 10.22214/ijraset.2023.49369.

[32] R. Nayak and R. Joshi, “Contextual Hate Speech Detection in Code Mixed Text using Transformer Based Approaches,” Oct. 2021.

[33] A. F. M. de Paula, I. Bensalem, P. Rosso, and W. Zaghouani, “Transformers and Ensemble methods: A solution for Hate Speech Detection in Arabic languages,” Mar. 2023.

[34] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, “Cyberbullying Detection: An Ensemble Based Machine Learning Approach,” in 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), IEEE, Feb. 2021, pp. 710–715. doi: 10.1109/ICICV50876.2021.9388499.

[35] L. Cheng, K. Shu, S. Wu, Y. N. Silva, D. L. Hall, and H. Liu, “Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model,” in Proceedings of the 29th ACM International Conference on Information & Knowledge Management, New York, NY, USA: ACM, Oct. 2020, pp. 185–194. doi: 10.1145/3340531.3411934.

[36] P. K. Roy and F. U. Mali, “Cyberbullying detection using deep transfer learning,” Complex & Intelligent Systems, vol. 8, no. 6, pp. 5449–5467, Dec. 2022, doi: 10.1007/s40747-022-00772-z.