2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom) | 979-8-3503-0948-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/MeditCom61057.2024.10621093

Cyberbullying Detection Using Bidirectional Encoder Representations from Transformers (BERT)

Razan Sujud[1], Walid Fahs [1], Rida Khatoun[2] and Fadlallah Chbib[2]
[1] Computer and Communication, IUL, LTCI, Wardanieh, Lebanon
[2] Polytechnic Institute of Paris, Telecom Paris, LTCI, France
razanhusseinsujud@gmail.com[1], walid.fahs@iul.edu.lb[1], rida.khatoun@telecom-paris.fr[2],
fadlallah.chbib@telecom-paris.fr[2]

***Abstract*—Bullying is described as an undesirable behavior by others that harms an individual physically, mentally, or socially. Cyberbullying is a virtual form (e.g., textual or image) of bullying. Thus, detecting cyberbullying behaviors early can prevent long-term psychological harm and promote safer digital spaces. Researchers have increasingly turned to machine learning techniques for effective cyberbullying detection in response to this pressing concern. However, the detection methods that rely on machine learning struggle with contextual understanding. Recently, there has been a shift toward using deep learning models, which have produced novel outcomes. Bidirectional Encoder Representations from Transformers (BERT) specifically utilizes a deep learning approach to learn contextualized representations of words or tokens in a given text corpus. It has been widely used for various Natural Language Processing (NLP) tasks, including text classification. In this paper, we propose a novel approach to building a robust system leveraging BERT, designed to effectively detect and categorize instances of cyberbullying across various online platforms. By employing sophisticated NLP techniques, the objective is to develop a model that can analyze and understand complex contextual details and identify cyberbullying behavior effectively. The system is trained on a diverse collected dataset from different platforms such as YouTube, LinkedIn, and Twitter. Our experimental results demonstrated the ability of our detection model to discriminate between potentially hazardous information and benign interactions according to different performance metrics such as Recall, Precision, and F1-score.**

***Index Terms*—Cyberbullying, Deep Learning (DL), Natural Language Processing (NLP), Bidirectional Encoder Representations from Transformers (BERT).**

I. Introduction

Cyberbullying has emerged as a prevalent and distressing issue in today's digital landscape. The anonymity and widespread accessibility of online platforms have intensified the frequency and severity of cyberbullying incidents. It presents in various ways, including transmitting hurtful messages regarding sensitive subjects such as racist and sexual, spreading rumors or lies, sharing embarrassing photos or videos without permission, or isolating individuals from online communities. Predominantly impacting adolescents, this type of aggression predominantly occurs on social media such as Facebook, Twitter, Instagram, Snapchat, and TikTok, presenting substantial hurdles in identification and mitigation. Unlike conventional bullying, which typically occurs in person, cyberbullying transcends geographical boundaries, rendering its detection and resolution more challenging [1], [2]. Research shows that Instagram is the most popular site for cyberbullying, with the greatest victimization rates, followed by Facebook and Snapchat. A considerable proportion of teenagers, ranging in age from 10 to 17, have reported experiencing harassment or bullying on social media and internet platforms. In particular, 42% of teenagers reported experiencing cyberbullying on Instagram, closely followed by 37% on Facebook and 31% on Snapchat. Twitter and YouTube both had smaller shares, at 9% Figure 1.

Text-based cyberbullying detection methods are classified into four main categories: supervised learning, lexicon-based, rule-based, and mixed-initiative approaches [3]. Lexicon-based approaches face a challenge due to their reliance on predefined wordlists, which may not capture emerging terms on online platforms. Supervised methods employ various algorithms such as Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR). However, traditional machine learning approaches struggle with contextual understanding, leading to difficulties in interpreting internet slang and sarcasm, potentially resulting in misclassification or false positives, as noted in [4]. To address this issue, deep learning techniques like Long-Short Term Memory (LSTM), MultiLayer Perceptron (MLP), and Convolutional Neural Network (CNN) are increasingly utilized [5]. In general, besides those challenges, there is a lack of labeled datasets and skewed distributions of available data, where instances of cyberbullying are in the minority.

Bidirectional Encoder Representations from Transformers (BERT) is a more advanced deep learning model that utilizes self-attention mechanisms and transformer encoders to capture contextual relationships among words and sentences in a given text [6]. This enables the generation of contextual word embeddings, where the embedding of a word can vary depending on its context within a sentence. BERT has gained extensive adoption in the field of Natural Language Processing (NLP), playing a crucial role in various applications including machine translation, sentiment analysis, computer vision and audio processing. In our work, We propose a novel model using BERT to create word embeddings which are then fed into a single neural network layer used for classification of text

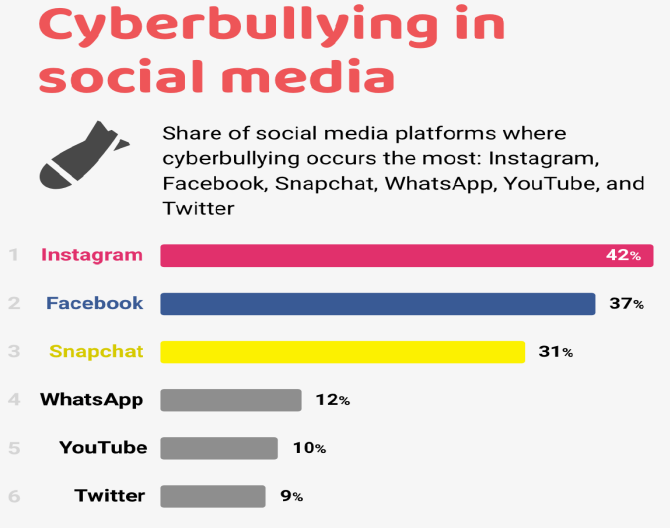

Fig. 1. Cyberbullying statistics in social media (2022) [7]

as cyberbullying or not. Our corpora consist of three sources of joined datasets from YouTube, WhatsApp, and Twitter to overcome the limitations of scarce cyberbullying data and imbalanced datasets.

The main contributions of this research encompass the following:

- Analysis of text-based, image-based, and multimodal approaches used by the research community for detecting cyberbullying.
- Compilation of a comprehensive dataset derived from diverse sources, including YouTube, Twitter, and WhatsApp.
- Introduction of a novel classification algorithm utilizing the BERT (Bidirectional Encoder Representations from Transformers) classification model.

We organize the rest of the paper as follows. Section II presents the literature review. Section III explains the data set, the proposed methodology, implementation, and experimental results. Finally, Section V concludes the paper.

II. Related Works

Over the previous decade, researchers have tried to develop a variety of effective methods for detecting cyberbullying. These techniques have included text-based, image-based, and multimodal approaches that combine various types of inputs, as demonstrated in the following literature reviews:

In the realm of text analysis, machine-learning techniques have traditionally held precedence. Various feature extraction methods have been employed to facilitate the analysis process, including Bag of Words (BoW), n-grams, Term Frequency-Inverse Document Frequency (TF-IDF), hashtags, emoticons for sentiment analysis, and pronouns. A notable study by [8] integrated TF-IDF, profanity, pronouns, sentiment analysis, and BoW as features for classification using Support Vector Machines (SVM). Through experimentation, they assessed each feature individually and explored optimal feature combinations. Linear regression identified sentiment analysis paired with TF-IDF as the most effective combination, resulting in a notable 75% accuracy, precision, and recall. Further investigations by [9] and [10] delved into the efficacy of BoW models with character n-grams, surpassing results obtained from word n-grams BoW models. Extending beyond textual content, [11] incorporated additional features such as post length, hashtag count, uppercase word count, and URL count to enhance analysis capabilities. These multifaceted approaches have led to significant advancements in accuracy and precision, providing deeper insights into text-based data analysis.

The effectiveness and performance of deep learning algorithms in detecting insults within Social Commentary were analyzed empirically by the authors of [12]. They utilized four deep learning models: Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Recurrent Neural Network (RNN) specifically for this purpose. Results indicate that the BLSTM model achieved notably high accuracy and F1-measure scores compared to RNN, LSTM, and GRU. Furthermore, the utilization of Convolutional Neural Networks (CNNs) [11] and [13] has contributed to enhancing cyberbullying detection. The identification of cyberbullying was further explored using the Long Short Term Memory (LSTM) model, a type of Recurrent Neural Network (RNN) model. The authors of [14] developed a representation learning framework specifically designed to identify cyberbullying. By applying different weights to a list of pre-established disparaging phrases, they enhanced it using word embeddings to produce bullying traits. These properties are combined with latent semantic features and Bag-of-Words to create the final representation, fed into a linear SVM classifier. They conducted an empirical study on a Twitter dataset, contrasting the proposed approach with many baseline text representation learning models and cyberbullying detection methods. The authors of [15] introduced an innovative application of BERT for identifying cyberbullying. Employing a simple classification model with BERT yielded state-of-the-art outcomes across three real-world datasets: Formspring, Twitter, and Wikipedia. Their comparison involved pitting their fine-tuned BERT against CNN, LSTM with an attention layer, and two commonly used traditional machine learning-based text classification models. BERT consistently outperformed other methods across all datasets. The authors of [16] presented a novel approach to cyberbullying detection on social media platforms, utilizing the novel pre-trained BERT model in combination with a single linear neural network layer as a classifier, yielding improvements over previous results. Two social media datasets were used for training and testing this model: the bigger Wikipedia dataset and the relatively smaller Form Spring dataset. Using different oversampling rates, the model's performance was evaluated on the Form Spring dataset. The results showed that the model performs better when the prevalence of bully posts increases. On the other hand, the Wikipedia dataset was not oversampled because the BERT model works well with big enough datasets.

Results from the suggested approach are noticeably better and more reliable, especially when used with bigger datasets like Wikipedia.

III. Proposed Approach

Recurrent neural networks (RNNs) are highly effective in modeling sequential data for prediction purposes. However, RNNs suffer from a limitation known as short-term memory. In 2017, Google introduced the Attention algorithm, which addresses this issue by calculating the significance of each word and enabling the network to focus its ”attention” on the most important words. In natural language processing (NLP), every model works with word embeddings, which are vector representations of words. In attention mechanism for each word in the initial sequence, a word embedding is created and multiplied by three distinct matrices, resulting in three different vectors: Query, Key, and Value. These vectors represent different perspectives of the original word embedding. This process, known as multi-head attention, is repeated eight times with randomly initialized matrices. BERT is composed of multiple encoders. An encoder is composed of Multi-Head attention and Feed-forward network sublayers. After every sublayer a residual connection and layer normalization are used to accelerate the training process. The residual connection is adding the initial vector to the result after each sublayer.

Considering the exceptional performance of BERT and the demand for contextual and semantic comprehension of text, we opted to employ BERT-based systems for cyberbullying detection. In this section, we outline our proposed approach, detailing our preprocessing techniques, methodology, and findings.

A. Data set

We integrated numerous publicly accessible datasets and had experts annotate the combined dataset based on specific criteria. During annotation, certain posts received conflicting annotations from different annotators, leading to the assignment of the most frequently occurring annotation to those posts. The resultant joined dataset comprised 370,000 instances categorized into two classes: cyberbully and non-cyberbully. However, only 93,000 samples were labeled as cyberbullying, resulting in a heavily skewed dataset. To address this imbalance and enhance results, we randomly selected 93,000 instances from the non-cyberbully class and augmented it to match the size of the cyberbully class. Our final balanced dataset consisted of 186K instances of YouTube, Twitter, and WhatsApp content as seen in Table I.

B. Preprocessing

Besides text, social media instances usually include tags, hashtags, links, and emojis. Those are valuable sources of information in classifying each instance, so as done in [26], we decided not to remove them but instead to preprocess them. We started by removing non-English instances found within our English dataset. We expanded contractions and removed numbers, stop words, punctuations, URLs, extra white spaces, and posts of less than three words. We replaced elongated words with their base form and filtered special characters such as & and $ present in some words, cleaned hashtags at the end of the sentence, and kept those in the middle by removing just the hashtag symbol. For Twitter datasets, we also removed the retweet abbreviation “RT” like [17] and [18]. We used an open-source Python library available on GitHub called Emoji to convert each emoji instance to the meaning behind that emoji. For example, an emoji showing a happy face will be converted to a”: happy face:”. We then removed the “: from each instance. After finding out that 5% of the instances’ length lies between 290 and 1883 words, we decided to only keep those tweets whose length is less than or equal to 290 words [14]. Finally, we converted every uppercase letter into lowercase as in [18] and [17]. This conversion is only necessary because the BERT model we fine-tuned is only trained using uncased text.

C. Methodology

1) BERT classification model: BERT [6] is a pretrained language model comprised of a series of transformer blocks. Pretraining for BERT involved utilizing two extensive corpora: The Books Corpus (800M words) and Wikipedia (2500M words). BERT underwent pretraining via two tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM involves training the model to predict masked words within a sentence, constructed by randomly masking 15% of tokens and replacing 80% with the [MASK] token, 10% with a random token, and 10% with the original token. NSP focuses on teaching the model to understand relationships between two sentences by determining if one sentence logically follows the other. The BERT-base model encompasses 12 transformer block layers with a hidden size of 768 and 110M parameters. In contrast, the BERT-large model contains 24 layers with a hidden size of 1024 and 340M parameters. For our purposes, we utilize the 110M version of the model and append a dense layer with softmax activation atop the BERT model for classification tasks.

2) Experiment settings: In training our model, we opt to use binary cross-entropy as our loss function. This decision is based on its effectiveness in binary classification tasks,

TABLE I
Sample of Dataset

Text	Preprocessed Text	Annotation
RT user Jonathan stop doing such stupid thinks like you idiot!!!!!!!!!stupidgirl	Jonathan stop doing such stupid thinks like you idiot stupidgirl	1
Andrea was here but actually, she is a fuckinnn retard, she keeps on forgetting what happens.I REALLY HATE HERRRRRRR	andrea was here but actually she is a fuckin retard she keeps on forget what happens i really hate her	1
What do you wantt ?? I will be running in the marathon next week bro come over later R	what do you want i will be run in the marathon next week bro come over later happy face	0

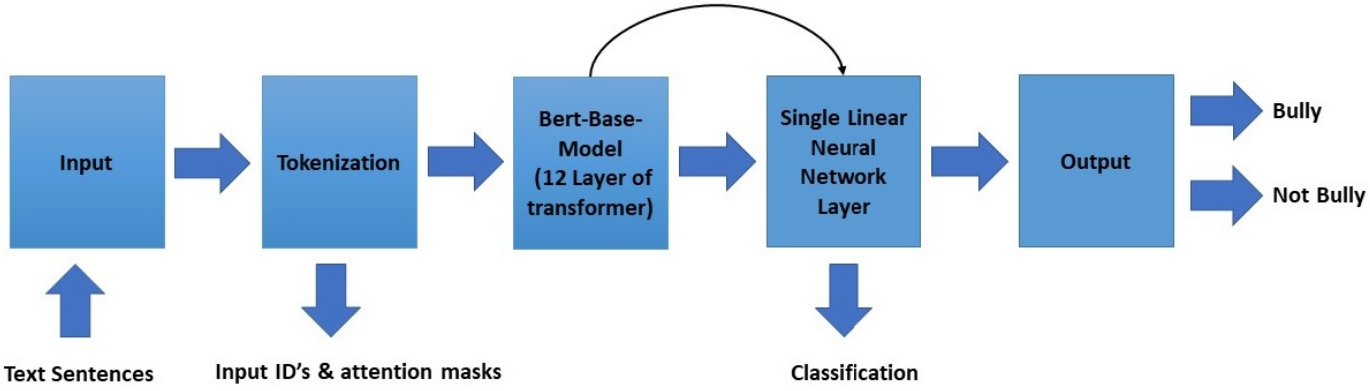

Fig. 2. Training pipeline of BERT-based classification.

including determining if a text contains bullying incidents or not Figure 2. The difference between the actual labels and the expected probabilities is quantified by binary cross-entropy, which gives us a precise indication of how well our model fits the intended results. Our model continuously modifies its parameters to better capture the underlying patterns in the data by minimizing this loss function throughout the training phase, which improves the model's capacity to categorize bullying incidents properly Equation 1.

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (1)$$

$L(y, \hat{y})$ represents the binary cross-entropy loss between the true labels y and the predicted probabilities $\hat{y}$,N is the number of samples, y_i and $\hat{y}_i$ are the true label and predicted probability for the $i-th$ sample, respectively.

We implement Adam's optimizer to efficiently update weights during training. The result would be a probability showing whether the input is classified as a cyberbully or not. Before feeding the Bert model with the dataset, it was divided into $70/30$ training and validation sets accordingly. Then, each set was further preprocessed by Bert's preprocessor which converts the input text into a form to be accepted by the transformer encoder as needed by Bert. To develop this project, we use Keras and TensorFlow hub in the Google Collaboratory environment. We were able to use the GPU engines. For fine-tuning, the BERT model was trained for 60 epochs with a batch size of 64 and a learning rate of $2e5$. We stopped on Epoch 28 because the results weren't getting better. In addition, we use for the dense layer a dropout probability of 30%. Figure 3 shows the block diagram of our proposed model.

D. Experimental Results

In this subsection, the experimental results of Bert-based cyberbullying detection are presented. The study aimed to detect cyberbullying by employing Bert for creating context-dependent word embeddings with careful consideration of

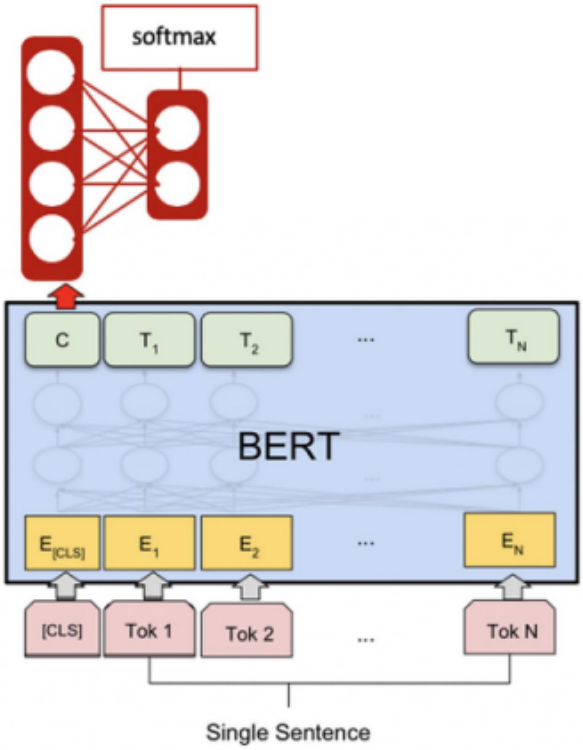

Fig. 3. BERT Classification Model

our balanced dataset, hyperparameters like number of epochs, learning rate, maximum input sequence length, and batch size. To evaluate the performance, we used 27K tweets, and various evaluation metrics like accuracy, precision, recall, and F1 as seen below in Table II A high F1 score indicates the strong overall performance of a binary classification model. It signifies that the model can effectively identify positive cases while minimizing false positives and false negatives. The performance evaluation metrics can be computed as follows:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

TABLE II
RESULTS ON THE VALIDATION AND TEST SETS.

	Validation (%)	Test
Accuracy	83	80
Precision	82	70
Recall	70	70
F1	76	70

where precision and recall are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall \ = \frac{TP}{FN + TP} \tag{4}$$

$$F1 \ score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{5}$$

As seen in Figure 4, we obtained a test accuracy of 80% with a higher validation accuracy of 83% as seen in Figure I Accuracy kept on increasing until reaching a plateau with small changes from epoch 20 and above until epoch 28. After hyper-tuning, this is the best result we could get and the highest accuracy and F1 score that our dataset could achieve.

Our model yielded a validation accuracy of 83%. Our dataset is balanced so we can rely on accuracy as an evaluation metric. These results offer critical insights into cyberbullying detection methods, potentially advancing the domain of text classification and sentiment analysis.

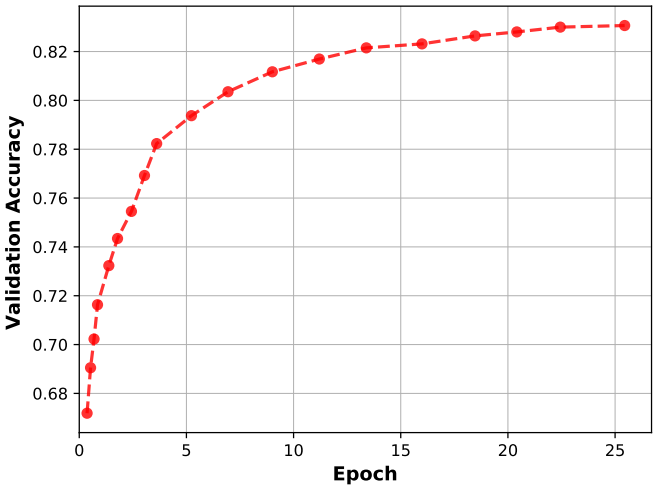

Fig. 4. Validation Accuracy

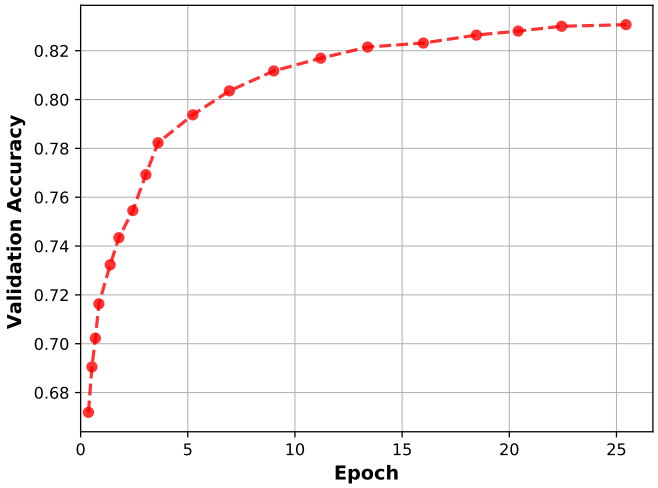

Fig. 5. Test accuracy

IV. CONCLUSION

This work focused on natural language processing within the realm of identifying cyberbullying in social media text, employing BERT, a cutting-edge advancement in word embedding technology. This context-aware transformer encoder addresses the challenge of slang and humor, which might otherwise lead to misclassification as cyberbullying or benign content. By utilizing context-dependent word embeddings, BERT captures semantic relationships and understands the context behind each word. Thus, it produces a different vector for the same word in different sentences with different contexts. Those vectors are fed into a single neural network layer for classification that was able to classify and predict the cyberbullying instances with an accuracy of up to 83%. As part of our future work, we can use some Generative models or techniques, which help in creating cyberbullying instances that can be concatenated with the available corpus and fed into a classification model. In addition, instead of the utilization of a single neural network layer, a CNN, LSTM, or other deep neural network layers can be studied to be added on top of BERT for better results.

REFERENCES

[1] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 3–24, 2017.

[2] E. V. Altay and B. Alatas, "Detection of cyberbullying in social networks using machine learning methods," in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, pp. 87–91, IEEE, 2018.

[3] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, vol. 11, no. 1, pp. 3–24, 2017.

[4] X. Guo, U. Anjum, and J. Zhan, "Cyberbully detection using bert with augmented texts," in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1246–1253, IEEE, 2022.

[5] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," *arXiv preprint arXiv:2003.01200*, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[7] R. Kumar and A. Bhat, "A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media," *International Journal of Information Security*, vol. 21, no. 6, pp. 1409–1431, 2022.
[8] A. Perera and P. Fernando, "Accurate cyberbullying detection and prevention on social media," *Procedia Computer Science*, vol. 181, pp. 605–611, 2021.
[9] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399, 2017.
[10] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in *European conference on information retrieval*, pp. 141–153, Springer, 2018.
[11] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pp. 745–760, Springer, 2018.
[12] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Systems*, vol. 29, no. 3, pp. 1839–1852, 2023.
[13] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, and J. P. Carvalho, "A "deeper" look at detecting cyberbullying in social networks," in *2018 international joint conference on neural networks (IJCNN)*, pp. 1–8, IEEE, 2018.
[14] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th international conference on distributed computing and networking*, pp. 1–6, 2016.
[15] S. Paul and S. Saha, "Cyberbert: Bert for cyberbullying identification: Bert for cyberbullying identification," *Multimedia Systems*, vol. 28, no. 6, pp. 1897–1904, 2022.
[16] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained bert model," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1096–1100, IEEE, 2020.
[17] D. Sultan, B. Omarov, Z. Kozhamkulova, G. Kazbekova, L. Alimzhanova, A. Dautbayeva, Y. Zholdassov, and R. Abdrakhmanov, "A review of machine learning techniques in cyberbullying detection.," *Computers, Materials & Continua*, vol. 74, no. 3, 2023.
[18] J. Qiu, M. Moh, and T.-S. Moh, "Multi-modal detection of cyberbullying on twitter," in *Proceedings of the 2022 ACM Southeast Conference*, pp. 9–16, 2022.