Article

Not peer-reviewed version

# Use of Large Language Model for Cyberbullying Detection

Bayode Ogunleye * and Babitha Dharmaraj

Posted Date: 15 June 2023

doi: 10.20944/preprints202306.1075.v1

Keywords: BERT; Cyberbullying; RoBERTa; Language Model; Machine learning; Online abuse; Natural language processing; NLP

*Article*

# Use of Large Language Model for Cyberbullying Detection

**Bayode Ogunleye [1,*] and Babitha Dharmaraj [2]**

[1] Department of Computing & Mathematics, University of Brighton, Brighton, BN2 4GJ, United Kingdom
[2] Department of Digital, Data and Technology, Ofgem, London, E14 4PU, United Kingdom
* Correspondence: B.Ogunleye@brighton.ac.uk

**Abstract:** The dominance of social media has added to the channels of bullying to perpetrators. Unfortunately, cyberbullying (CB) is the most prevalent phenomenon in today's cyber world and is a severe threat to the mental and physical health of citizens. This opens the need to develop a robust system to prevent bullying content from online forums, blogs, and social media platforms to manage the impact in our society. Several machine learning (ML) algorithms have been proposed for this purpose. However, their performances are not consistent due to high class imbalance issue and generalisation. In recent years, large language models (LLM) like BERT and RoBERTa have achieved state of the art (SOTA) results in several natural language processing (NLP) tasks. Unfortunately, the LLMs have not been applied extensively. In our paper, we explored the use of these models for cyberbullying (CB) detection. We have prepared a new dataset (D2) from existing studies (Formspring and Twitter). Our experimental results for dataset D1 and D2 showed RoBERTa outperformed other models.

**Keywords:** Large Language model; Cyberbullying; Machine Learning; Online abuse; Natural language processing; RoBERTa; social media analytics; BERT

## 1. Introduction

The emergence of social technologies like Facebook, Twitter, TikTok, WhatsApp and Instagram have improved communication amongst the people and businesses across the globe. However, despite the huge advantage of these platforms, they have also added channels of bullying to perpetrators. Cyberbullying (CB), often referred to as online bullying is becoming an important issue that requires urgent attention. For illustration, in the USA, Pew research center (2018) reported that around two-third of US adolescents have been subjected to cyberbullying. Statista (2021) reported in their survey that 41% of adults in the USA had experienced cyberbullying. Pew research center (2022) reported 46% of teens in the USA aged 13 to 17 have been cyberbullied. Office for National Statistics (2020) reported that 19% of children aged 10 to 15 (this equates to 764,000 children) have experienced cyberbullying in England and Wales. Patchin and Hinduja (2020) found out that 90% of teens (9 to 12 years old) utilise social media or gaming apps, and 20% of tweens are involved in CB either as a victim, an offender, or a bystander.

This problem of cyberbullying (CB) is a relatively new trend that has recently gained more popularity as a subject. Cyberbullying is the repetitive, aggressive, targeted, and intentional behaviour aimed to hurt an individual's or a group's feelings through an electronic medium (Emmery et al. 2021; Dinakar et al. 2011). CB takes many forms, including flaming, harassment, denigration, impersonation, exclusion, cyberstalking, grooming, outing, and trickery (Emmery et al. 2021; Reynolds et al. 2011). Cyberbullies are more likely to be technologically astute than physically stronger, making them better able to access victims online, conceal their digital footprints, and involve in posting rumours, insults, sexual comments, threats, a victim's private information, or derogatory labels (Aboujaoude et al. 2015). The fundamental causes of any bullying incident are the imbalance of power and the victim's perceived differences in race, sexual orientation, gender, socioeconomic level, physical appearance, and mannerism. Xu et al. (2012) stated that CB participants can play the role of either a bully, victim, bystander, bully assistant, reinforcer, reporter, or accuser.

Prior studies found out that CB impacts anxiety (Huang et al. 2021; Hellfeldt et al. 2020; Nixon, 2014), depression (Jin et al. 2023; Karki et al. 2022), social isolation (Piccoli et al. 2022), suicidal thoughts (Peng et al. 2019; Kim et al. 2019; Zaborskis et al. 2018), and self-harm (Islam et al. 2022; Eyuboglu et al. 2021). Messias et al. (2011) stated victims of cyberbullying have higher rates of depressive illnesses and suicidality than victims of traditional bullying. Patchin and Hinduja (2020) stated CB victims admit that they frequently feel awkward or afraid to attend school and it impacts their academic performance. In addition, they found out that nearly 70% of teens who reported being victims of cyberbullying stated it had a negative impact on their self-esteem and nearly one-third claimed it had an impact on their friendships. Despite the impact and increasing rate of CB, unfortunately, there is limited attention paid to developing sophisticated approaches for automatic CB detection.

CB studies are yet to extensively explore the use of large language models (LLMs) for CB detection (Elsafoury et al. 2021). CB is commonly misinterpreted, leading to flawed systems with little practical use. Additionally, several studies only evaluated using swear words to filter CB, which is only one aspect of this topic, and swear words may not always indicate bullying on platforms with a high concentration of youngsters (Emmery et al. 2021; Rosa et al. 2018). Thus, it is practically useful for developers and media handlers to have a robust system that understand context better to enhance CB detection. In our study, we aim to evaluate the performance of large language models for CB detection. Unfortunately, there are some obstacles to CB detection. One is the issue of unavailable balanced and enriched benchmark datasets (Emmery et al. 2021; Elsafoury, et al. 2021; Rosa et al. 2019). The issue of class imbalance has been a popular problem in machine learning (ML) applications as the ML algorithms tend to be biased towards the majority class (Ogunleye, 2021). Past studies emphasised on the class imbalance problem in the CB context (Yi & Zubiaga, 2022). In most studies, the proportion of bullying post is in the range of 4 - 20% of the entire dataset compared to non-bullying post (Emmery et al. 2021; Agrawal and Awekar 2018; Di-Capua et al. 2016; Kontostathis et al. 2013). This opens the need to create a new, enriched dataset with balanced classes for effective CB detection and make it publicly available. To this end, we propose the use of Robustly optimized BERT approach (RoBERTa), a pre-trained large language model for cyberbullying detection. Thus, our contributions can be summarised as follows. We prepared a new dataset (D2) from existing studies for the development of algorithms on CB detection. We conducted an experimental comparison of sophisticated machine learning algorithms with two datasets (D1 & D2). We ascertained RoBERTa as the state-of-the-art (SOTA) method for automated cyberbullying detection. The rest of the paper is organised as follows. Section 2 will review the literature to provide background knowledge to this study. Section 3 will present the methodology. Section 4 will present and discuss the results and section 5 will provide conclusions and recommendations.

## 2. Related Work

Cyberbullying (CB) is the most prevalent phenomenon in today's digital world and is a severe threat to the mental and physical health of cybercitizens (Jin et al. 2023; Centers for Disease Control and Prevention, 2014; Xu et al. 2012). Several studies have proposed various techniques for automated CB detection. For example, Xu et al. (2012) crawled 1762 tweets from Twitter using keywords such as "bully, bullied, bullying". The data was labelled by five human annotators such that 684 were labelled as bullying and 1078 as non-bullying. They compared four traditional machine learning models namely, Naïve Bayes (NB), Support Vector Machines (linear SVM), Support Vector Machines (RBF - SVM), and Logistic regression (LR). Their result showed linear SVM achieved the best performance with 77% F1 score. Agrawal and Awekar (2018) compared machine learning (ML) models namely, naïve bayes (NB), support vector machines (SVM), random forest (RF), convolutional neural network (CNN), long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), and BiLSTM with attention mechanism. They used dataset from three different social media platforms namely, Formspring (a total of 12773, split into 776 bully and 11997 non bullying text) which was collected by authors of Reynolds et al. (2011), Twitter (16090, split into 1937 bullying through racism, 3117 bullying through sexism and 11036 non-bullying text) collected by authors of Waseem & Hovy (2016), and Wikipedia (115864, split into13590 attack and 102274 not

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1075.v1



attack text) collected by authors of Wulczyn et al. (2017). They oversampled the minority class using the SMOTE (Synthetic Minority Oversampling technique). Their BiLSTM with attention implementation achieved at least 87% F1 score for the bullying across the three different social media platforms. Similarly, Huang & Qi (2023) reproduced the experiment in Agrawal and Awekar (2018) with Formspring dataset (Reynolds et al. 2011) only. Their result showed SVM performed better than logistic regression (LR), decision tree (DT), and random forest (RF) with 98% accuracy and F1 score of 93% (86% f1 score for bullying class). Alduailaj & Belghith, (2023) compared SVM and NB for cyberbullying detection in Arabic language context. They collected 30,000 Arabic comments on 5 February 2021 from Twitter and Youtube and the comments were labelled manually as bullying and non-bullying using most common and frequent Arabic bullying keywords detected from the comments. Their result showed term frequency inverse document frequency (TF-IDF) vectors deployed to SVM achieved accuracy of 95% and F1 score of 88%. Dewani et al. (2023) showed that SVM and embedded hybrid N-gram approach performed best in detecting cyberbullying in the Roma Urdu language context with an accuracy of 83%. Suhas-Bharadwaj et al. (2023) applied extreme learning machine to classify cyberbullying messages and achieved accuracy of 99% and F1 score of 91%. Woo et al. (2023) conducted a systematic review of CB literature. Their literature review findings suggest SVM and Naïve Bayes (NB) are the best performing models for CB detection.

Recently, there have been development of large language models (LLM) which has taken the world by surprise. The LLMs have been applied to several NLP tasks like topic modelling (Ogunleye, 2023), sentiment analysis (Zhao & Yu, 2021), recommendation system (Yang et al. 2022), and harmful news detection (Lin et al. 2022). In the context of CB detection, Paul & Saha, (2022) compared bidirectional encoder representations from transformers (BERT) to BiLSTM, SVM, LR, CNN, and hybrid of RNN and LSTM, using three real-life CB datasets. The datasets are from Formspring (collected by authors of Reynolds et al. 2011), Twitter (collected by authors of Waseem & Hovy, 2016), and Wikipedia (collected by authors of Wulczyn et al. 2017). They used SMOTE to rebalance the dataset and thus, showed BERT outperformed other models across the datasets with at least 91% F1 score. Similarly, Yadav et al. (2020) applied BERT to the Formspring dataset (Reynolds et al. 2011), and Wikipedia dataset (Wulczyn et al. 2017). Their approach achieved F1 score of 81% for the Wikipedia dataset. They rebalanced the Formspring dataset thrice and achieved F1 score of 59% with first oversampling rate, F1 score of 86% with second oversampling dataset and 94% F1 score in the third oversampling dataset. However, it is worth mentioning that they have tested their model on the oversampled dataset and thus might not be reliable in terms of generalisation. Yi & Zubiaga (2022) used the same dataset from Formspring (Reynolds et al. 2011), Wikipedia (Wulczyn et al. 2017) and Twitter (Waseem & Hovy, 2016). They proposed XP-CB, a novel cross-platform adversarial framework based on transformers and adversarial learning models for cross platform CB detection. They showed XP-CB can enhance a transformer leveraging unlabelled data from the source and target platforms to come up with a common representation while preventing platform-specific training. They showed XP-CB achieved an average macro F1 score of 69%. In summary, popular data source for CB detection are Wikipedia, Formspring, and Twitter (Paul & Saha, 2022; Yi & Zubiaga 2022; Reynolds et al. 2011; Yadav et al. 2020). Our literature review findings suggest that very few studies have used transformers models (pre-trained large language models) for the CB detection. CB literature survey conducted by Woo et al. (2023) found that most studies have used traditional machine learning models for CB detection. Thus, our paper looks to compare the performance of state-of-the-art (SOTA) language models for CB detection.

## 3. Methodology

We propose the use of fine-tuned Robustly optimized BERT approach (RoBERTa) for automatic cyberbullying (CB) detection. We conducted an experimental comparison of large language models to traditional machine learning models such as support vector machine (SVM) and random forest (RF).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1075.v1



*3.1. Bidirectional Encoder Representations from Transformers (BERT)*

BERT (Devlin et al. 2018) is a self-supervised autoencoder (AE) language model developed by Google in 2018 for training NLP systems. BERT is a bidirectional transformer-based model pre-trained on a large-scale Wikipedia corpus using the Masked Language Model (MLM) and next-sentence prediction tasks. BERT (base) is implemented based on transformer and attention mechanism which uses encoder to read in input and decoder to output. The BERT base model consists of 12 layers of transformer blocks, 768 hidden layer size, and 12 self-attention heads. The model comprises of two stages which are pre-training stage and fine-tuning stage. Without making significant task-specific architecture alterations, the pre-trained BERT model can be improved with just one extra output layer to produce cutting-edge models for a variety of tasks, including question answering and language inference. The objective of the masked language model is to predict the actual vocabulary id of a masked word only based on its context after randomly masking some of the tokens from the input. The MLM's intent permits the representation to combine the left and the right context, in contrast to the left-to-right language model pre-training, which allows us to pre-train a deep bidirectional Transformer. To fine-tune the pre-trained BERT model, the model is first instantiated with default parameters (used when pre-trained) and then the parameters are fine-tuned using labelled data from downstream tasks (text classification in our case). Every sequence will start with a particular classification token as the first token ([CLS]). For classification tasks, the last hidden state matching to this token is used as the aggregate sequence representation (Figure 1 below). The sum of the token embeddings, segmentation embeddings, and position embeddings constitutes the input embeddings as shown in Figure 2 below.

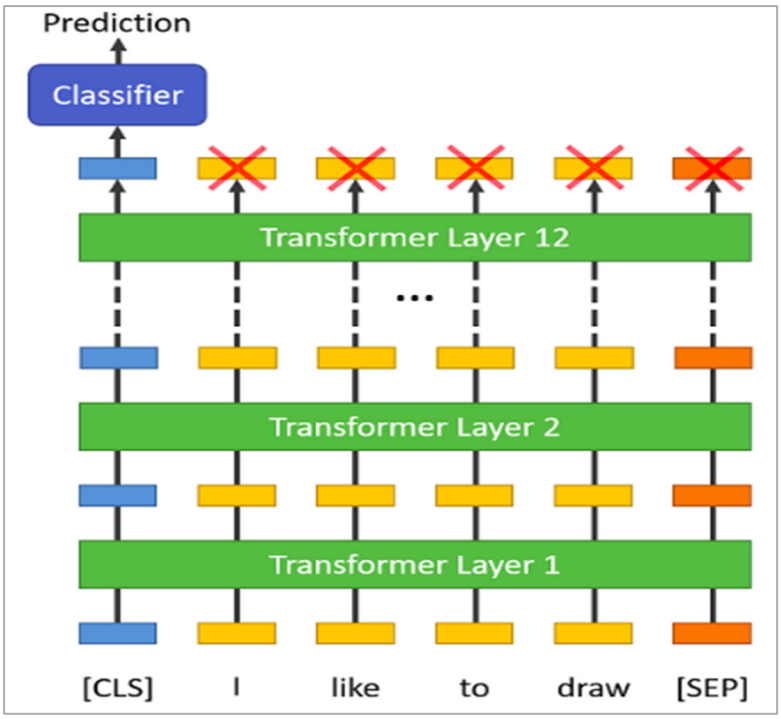


**Figure 1.** Illustration of BERT model architecture.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1075.v1



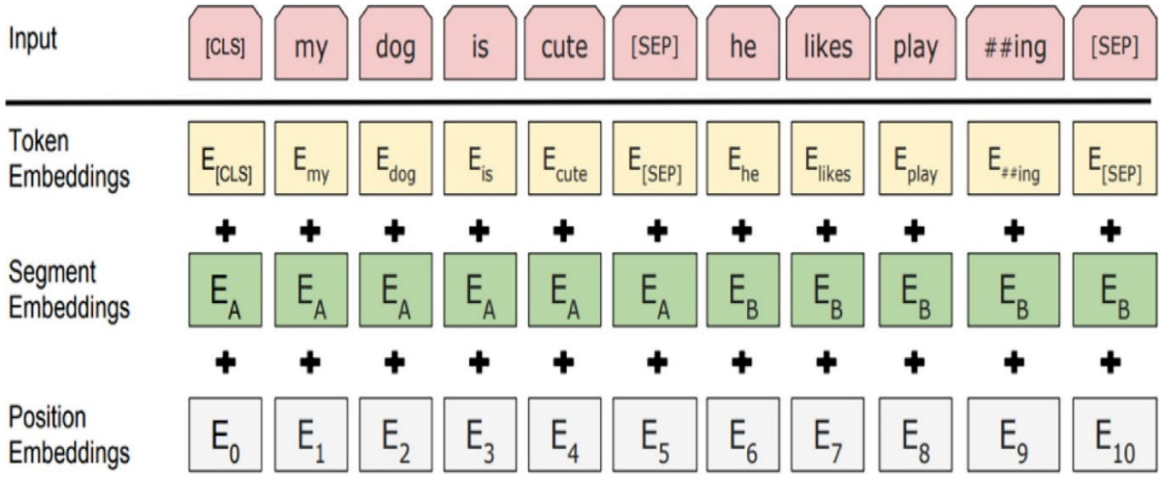


**Figure 2.** BERT input representation (Devlin et al. 2018).

### *3.2. RoBERTa*

Facebook AI Research (FAIR) identified the limitations of Google's BERT and proposed Robustly optimized BERT approach (RoBERTa) in 2019. Liu et al. (2019) stated that BERT was undertrained, and they modified the training method by (i) using dynamic masking pattern instead of static, (ii)training with more data with large batches, (iii) removing next sentence prediction, and (iv) training on longer sentences and proposed RoBERTa. As a result, RoBERTa outperforms BERT in terms of the masked language modelling objective and performs better on downstream tasks. For training the model, the researchers employed pre-existing unannotated NLP datasets as well as the unique collection CC-News, which was compiled from publicly available news stories. RoBERTa is a component of the effort by Facebook to improve the state-of-the-art in self-supervised models so that they may be created with less dependency on time and data-labelling.

### *3.3. XLNET*

XLNet (Yang et al. 2019) is a permutation based autoregressive transformer that combines the finest aspects of autoencoding and autoregressive language modelling while seeking to get around their drawbacks. BERT (autoencoder language model) ignores the dependency between the masked positions and leads to pretrain-finetune discrepancy because it relies on masking the input to corrupt it. On the other hand, conventional autoregressive (AR) language models predict the next word based on the word's context either in forward or in backward direction but not in both. XLNet training objective calculates the likelihood of a word based on all possible word permutations in a sentence, rather than only those to the left or right of the target token. Integrating Transformer-XL and a carefully thought-out two-stream attention mechanism are just two of the ways that the XLNet neural architecture is designed to perform in perfect harmony with the AR (autoregressive) mission. It is anticipated that to capture bidirectional context, each position would learn to make use of contextual data from all positions.

### *3.4. XLM-RoBERTa*

The multilingual version of RoBERTa is called XLM-RoBERTa (Conneau, 2019) was released by Facebook as an update to their XLM-100 model. It was trained on 100 languages from 2.5TB of filtered common crawl data. The "RoBERTa" part in XLM-RoBERTa originates from the fact that it uses the identical training procedures as the monolingual RoBERTa model, with the Masked Language Model training objective. There is no ALBERT-style sentence order prediction or BERT-style Next Sentence prediction in XLM-RoBERTa.

 doi:10.20944/preprints202306.1075.v1



*3.5. Evaluation Metrics*

This section discusses the evaluation metrics of the models to understand how well they have performed in this context and help decide on the best model. In a classification task, the common evaluation metrics are accuracy, precision, recall, and f1-measure. The proportion of correctly and wrongly labelled classes will be described in the form of true positive, true negative, false positive and false negative.

Where:

True Positive (TP) i.e. Positive class classified correctly

False Positive (FP) i.e. Negative class wrongly predicted as positive (Type I error)

False Negative (FN) i.e. Positive class wrongly predicted as negative (Type II error)

True Negative (TN) i.e. Negative class classified correctly

Accuracy: is the ratio of number of samples predicted correctly to the total number of samples.

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision: denoted as P is the percentage % of selected items that is correct.

$$\frac{TP}{TP + FP} \quad (2)$$

Recall: denoted as R is the percentage % of correct items that are selected.

$$\frac{TP}{TP + FN} \quad (3)$$

F1 measure provides the balance between the precision and recall and can be denoted as

$$2PR / P + R. \quad (4)$$

*3.6. Dataset*

In this study, we have used data from existing cyberbullying (CB) studies. The datasets were collected from Formspring.me and Twitter. We have named the datasets D1, and D2, for easy identification. In Dataset (D1), we used the dataset of Agrawal and Awekar (2018). They collected the data from Formspring and employed three human annotators from Amazon Mechanical Turk service to label the data as bullying or non-bullying. The data is publicly available via this link (GitHub - sweta20/Detecting-Cyberbullying-Across-SMPs). Table 1 below shows the class distribution of the dataset (D1).

**Table 1.** D1 – Formspring (CB) data distribution (Agrawal and Awekar, 2018).

| Label | Count |
|---|---|
| Not Cyberbullying (CB) | 11997 |
| Cyberbullying | 776 |

Wang et al. (2020) prepared new dataset from six existing studies (Agrawal and Awekar, 2018; Bretschneider et al. 2014; Chatzakou et al. 2019; Davidson et al. 2017; Waseem and Hovy, 2016; Xu et al. 2012). Their datasets were manually annotated into fine-grained CB target classes namely, victim's

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1075.v1



age, ethnicity, gender, religion, other quality of CB and not cyberbullying (notcb). The datasets were imbalanced; hence, they applied a modified Dynamic Query Expansion (DQE) to augment the dataset in a semi-supervised manner. They then randomly sampled approximately 8,000 tweets from each class to have a balanced dataset of approximately 48000 tweets. The dataset is publicly available and can be accessed via this link below. Table 2 below provides the class distribution of their dataset. https://drive.google.com/drive/folders/1oB2fan6GVGG83Eog66Ad4wK2ZoOjwu3F.

**Table 2.** Twitter (CB) data distribution (Wang et al. 2020).

| Label | Count |
|---|---|
| Age | 7992 |
| Ethnicity | 7961 |
| Gender | 7973 |
| Religion | 7998 |
| Other CB | 7823 |
| Not CB | 7945 |

The issue of data scarcity and class imbalance is a popular problem in CB detection domain. To resolve this, we took a different approach to Wang et al. (2020). This is because we aim to prepare a dataset which is comparable to prior and future work and to enhance the development of CB detection algorithms. Thus, we prepared our binary classification **dataset (D2)** from existing CB studies including Wang et al. (2020). In dataset **(D2),** we converted the multi-class dataset of Wang et al. (2020) into binary classes by labelling the 'age', 'ethnicity', 'gender', 'religion' and 'other cyberbullying' classes as "bullying" class. Agrawal and Awekar (2018) have 11,997 CB non- bullying instances and only 776 bullying instances from FormSpring. Thus, we concatenated the two imbalanced binary class datasets to create ours. The instances that had less than 3 words or more than 100 words have been considered outliers and removed to obtain a more representational dataset. Table 3. below presents the distribution of our dataset.

**Table 3.** D2 - distribution (Our study).

| Label | Count | Source | Annotation |
|---|---|---|---|
| Bullying | 19553 | FormSpring + Twitter | Manual |
| Non-Bullying | 19526 | FormSpring + Twitter | Manual |

*3.7. Experimental Set Up*

The TF-IDF weighted feature vectors and SBERT embeddings are used as input to the RF and SVM models. Support Vector Machine was employed using 'radial basis function' (rbf) as kernel function with C=5 and gamma=0.5 for non-linear classification. The Random Forest classifier was employed with the number of estimators as 100 and the bootstrap parameter set to True (to allow taking subsamples for each decision tree). Additionally, we utilized HuggingFace transformers to implement pre-trained language models. The number of training epochs used is 4. For training and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1075.v1



testing, we divided the data into stratified samples of 90% and 10%. With all the above descriptions for the models, the study was conducted in two cases. In case 1, we used the dataset D1 which is highly imbalanced with 10702 negative class instances and 703 positive class instances for training. This imbalanced dataset was given as input to pre-trained language models namely, BERT, RoBERTa, XLNET, and XLM-RoBERTa. Similarly, in case 2, the dataset D2 was used as input to the ML algorithms, and we present the result in subsequent section.

## 4. Experimental Result

This section presents the result of our experimental comparison of the classification algorithms. Our experiment was in two folds. In the first experiment, we implemented the algorithms with the imbalanced dataset (D1) and named this case 1. In the second experiment, we implemented the algorithms with the balanced dataset (D2) prepared, and this is named case 2. Table 4 below presents the evaluation report of our case 1.

**Table 4.** Evaluation report of classification algorithms (case 1).

| Algorithms | Class | Training Size | Accuracy (%) | Precision (%) | Recall (%) | Macro F1-Score (%) |
|---|---|---|---|---|---|---|
| BERT | 0 | 10792 | 0.96 | 0.98 | 0.98 | 0.98 |
| | 1 | 703 | | 0.66 | 0.63 | 0.64 |
| XLNet | 0 | 10792 | 0.95 | 0.97 | 0.99 | 0.98 |
| | 1 | 703 | | 0.70 | 0.41 | 0.52 |
| RoBERTa | 0 | 10792 | 0.95 | 0.98 | 0.98 | 0.98 |
| | 1 | 703 | | 0.63 | 0.68 | **0.66** |
| XLM - | 0 | 10792 | 0.94 | 0.98 | 0.96 | 0.97 |
| RoBERTa | 1 | 703 | | 0.53 | 0.68 | 0.60 |

In our experiment, the positive class instances ('bully') is denoted '1' and this class is of interest. Firstly, we implemented the traditional machine learning classifiers namely, support vector machine (SVM), random forest (RF), and logistic regression. However, all these classifiers produced poor result. Thus, they were excluded in case 1. The result in Table 1 above shows that RoBERTa achieved the best performance with F1 score of 0.66. In general, the algorithms struggled with the positive class instances ('bully'), when compared to negative instances ('non bully'), especially the XLNet model. This is unsurprising as its due to the class imbalance of the dataset (D1). However, the results are superior to the result of Agrawal and Awekar (2018). Agrawal and Awekar (2018) is the creator of dataset D1, and they applied bidirectional long short-term memory (BiLSTM) with attention mechanism for CB detection. In their study, their BiLSTM implementation achieved a F1 score of 0.51 for the positive class instance ('bully'). In comparison to our study, all our large language models applied to D1 showed better performances.

Furthermore, in the experiments of Agrawal and Awekar (2018), they improved on D1 by oversampling the minority class using the SMOTE (Synthetic Minority Oversampling technique). Their BiLSTM implementation achieved F1 score of 0.91 for the positive class instance ('bully'). However, Emmery et al. (2021) criticised their implementation. Emmery et al. (2021) reproduced their experiment and discovered the overlap between train-test data due to the oversampling method. They showed the results of Agrawal and Awekar (2018) for oversampled case is not reliable. Thus, Emmery et al. (2021) modified the implementation (BiLSTM with attention mechanism) by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1075.v1



oversampling only the training data and achieved F1 score of 0.33 for 'bully' class. To conclude for case 1 (D1), all our models namely, BERT, RoBERTa, XLNet and XLM-RoBERTa achieved better performance than that of Agrawal and Awekar (2018) and Emmery et al. (2021) as reported in Table 4 above. The performance of the large language models can be attributed to their power to understand context better and effective understanding of long sequence text. Thus, it is not surprising to see that the large language models have performed better than the traditional machine learning models and the hybrid algorithms.

**Table 5.** Evaluation report of classification algorithms (case 2).

| Algorithms | Class | Training Size | Accuracy (%) | Precision (%) | Recall (%) | Macro F1-Score (%) |
|---|---|---|---|---|---|---|
| BERT | 0 | 17573 | 0.85 | 0.85 | 0.86 | 0.86 |
| | 1 | 17598 | | 0.86 | 0.85 | 0.86 |
| XLNet | 0 | 17573 | 0.86 | **0.88** | 0.84 | 0.86 |
| | 1 | 17598 | | 0.84 | 0.88 | 0.86 |
| RoBERTa | 0 | 17573 | **0.87** | 0.87 | 0.86 | 0.86 |
| | 1 | 17598 | | 0.86 | 0.87 | **0.87** |
| XLM-RoBERTa | 0 | 17573 | 0.86 | 0.86 | 0.86 | 0.86 |
| | 1 | 17598 | | 0.86 | 0.86 | 0.86 |
| SBERT + SVM | 0 | 17573 | 0.85 | 0.84 | 0.87 | 0.86 |
| | 1 | 17598 | | 0.87 | 0.83 | 0.85 |
| SBERT +RF | 0 | 17573 | 0.81 | 0.79 | 0.86 | 0.82 |
| | 1 | 17598 | | 0.84 | 0.77 | 0.81 |
| TF-IDF + SVM | 0 | 17573 | 0.84 | 0.84 | 0.86 | 0.85 |
| | 1 | 17598 | | 0.86 | 0.83 | 0.85 |
| TF-IDF + RF | 0 | 17573 | 0.84 | 0.80 | **0.90** | 0.85 |
| | 1 | 17598 | | **0.88** | 0.78 | 0.83 |

Table 5 above presents the case 2 of our experiment. Using our dataset (D2), we ascertain RoBERTa as the state-of-the-art model as the algorithm achieved the best performance overall with F1 score of 0.87 for positive class ('bully'). Also, the results of all four language models prove that they are all comparable in performance with not much variance. This agrees with the result of Paul and Saha (2022) that showed Bidirectional Encoder Representation from Transformer (BERT) performs than deep learning algorithms like (BiLSTM) for automated cyberbullying detection. Our experimental findings showed that pre-trained language models are powerful and competitive with other models in the detection of cyberbullying on social media sites. Furthermore, it is worth noting that because we have used a balanced training dataset (D2), the traditional machine learning models also showed good performance, most notably, the support vector machine (SVM) that achieved F1 score of 0.85 for the positive class ('bully'). This is consistent with the result of Ogunleye (2021) that

 doi:10.20944/preprints202306.1075.v1



showed SVM is a robust classification algorithm when fed with a balanced training dataset. In summary, we propose the use of RoBERTa for CB detection.

## 5. Conclusions

In this study, we aimed to ascertain state-of-the-art (SOTA) language model for automated cyberbullying (CB) detection. We prepared a new dataset (D2) from existing CB studies. The datasets were originated from FormSpring and Twitter and were manually annotated. We used the dataset in our implementation and our results showed RoBERTa performed well in both experiments, case 1 & 2. For a classification task, we argue that large language models (LLMs) can predict the minority class better than the approach of using traditional machine learning approach and/or oversampling technique (case 1). This is due to the ability of language model to understand context of long and short text. In addition, we showed that RoBERTa perform better than deep learning approach like BiLSTM with attention mechanism. We also evidenced that when dataset is balanced, traditional machine learning approach produces good performance, however, RoBERTa yielded a state-of-the-art (SOTA) performance. To conclude, our contributions can be summarised as follows. We prepared a new dataset (D2) for the development of algorithms in the field of CB detection. The dataset (D2) has been made publicly available for research access and use. We demonstrated how large language models can be used for automated CB detection with two datasets (D1 & D2). We presented SOTA results for CB detection by fine tuning RoBERTa.

In theory, the use of machine learning algorithm yields poor performance when fed with imbalance dataset compared to large language models. Similarly, language models yield better result with balanced dataset. This implies that the performance of RoBERTa is consistent across different categorises (balanced or not) of cyberbullying dataset. In practice, our application is useful for social network owners, the government, and developers to implement cyberbullying detection algorithm to prevent and reduce the act. For future work, we consider implementing a multimodal approach to develop and enhance algorithms for CB detection. In addition, the language models can be tuned with external corpus to improve on SOTA models. This implementation can be extended to detect other forms of online abuse including hate speech, and cyber molestation.

**Author Contributions:** "Conceptualization, B.O. and B.D.; methodology, B.O. and B.D.; software, B.D.; validation, B.O and B.D.; formal analysis, B.D.; investigation, B.O. and B.D.; resources B.D.; data curation, B.O. and B.D.; writing—original draft preparation, B.O and B.D.; writing—review and editing, B.O.; visualization, B.O. and B.D.; supervision, B.O.; project administration, B.O.; All authors have read and agreed to the published version of the manuscript.".

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data availability Statement:** The dataset and code used in this work is available on Github repository (please see link below).

GitHub - Babitha23/Cyberbullying-detection

Case 1:

https://github.com/Babitha23/Cyberbullying-detection/tree/main/Case1

Case 2:

https://github.com/Babitha23/Cyberbullying-detection/tree/main/Case2

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Agrawal, S., & Awekar, A. (2018, March). Deep learning for detecting cyberbullying across multiple social media platforms. In Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings (pp. 141-153). Cham: Springer International Publishing.
2. Alduailaj, A. M., & Belghith, A. (2023). Detecting Arabic Cyberbullying Tweets Using Machine Learning. Machine Learning and Knowledge Extraction, 5(1), 29-42.
3. Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., & Goujon, A. (2021, April). A comparison of pre-trained language models for multi-class text classification in the financial domain. In Companion Proceedings of the Web Conference 2021 (pp. 260-268).
4. Behzadi, M., Harris, I. G., & Derakhshan, A. (2021, January). Rapid Cyber-bullying detection method using Compact BERT Models. In 2021 IEEE 15th International Conference on Semantic Computing (ICSC) (pp. 199-202). IEEE.
5. Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
6. Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. D., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting cyberbullying and cyberaggression in social media. ACM Transactions on the Web (TWEB), 13(3), 1-51.
7. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
8. Dewani, A., Memon, M. A., Bhatti, S., Sulaiman, A., Hamdi, M., Alshahrani, H., ... & Shaikh, A. (2023). Detection of Cyberbullying Patterns in Low Resource Colloquial Roman Urdu Microtext using Natural Language Processing, Machine Learning, and Ensemble Techniques. Applied Sciences, 13(4), 2062.
9. Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the timeline meets the pipeline: A survey on automated cyberbullying detection. IEEE access, 9, 103541-103563.
10. Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., ... & Daelemans, W. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Language Resources and Evaluation, 55, 597-633.
11. Eyuboglu, M., Eyuboglu, D., Pala, S. C., Oktar, D., Demirtas, Z., Arslantas, D., & Unsal, A. (2021). Traditional school bullying and cyberbullying: Prevalence, the effect on mental health problems and self-harm behavior. Psychiatry research, 297, 113730.
12. Hellfeldt, K., López-Romero, L., & Andershed, H. (2020). Cyberbullying and psychological well-being in young adolescence: the potential protective mediation effects of social support from family, friends, and teachers. International journal of environmental research and public health, 17(1), 45.
13. Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. In Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7 (pp. 49-66). Springer International Publishing.
14. Huang, H., & QI, D. (2023). Cyberbullying detection on social media. Higher Education and Oriental Studies, 3(1).
15. Huang, J., Zhong, Z., Zhang, H., & Li, L. (2021). Cyberbullying in social media and online games among Chinese college students and its associated factors. International journal of environmental research and public health, 18(9), 4819.
16. Islam, M. I., Yunus, F. M., Kabir, E., & Khanam, R. (2022). Evaluating risk and protective factors for suicidality and self-harm in Australian adolescents with traditional bullying and cyberbullying victimizations. American journal of health promotion, 36(1), 73-83.
17. Jin, X., Zhang, K., Twayigira, M., Gao, X., Xu, H., Huang, C., ... & Shen, Y. (2023). Cyberbullying among college students in a Chinese population: Prevalence and associated clinical correlates. Frontiers in Public Health, 11.
18. Karki, A., Thapa, B., Pradhan, P. M. S., & Basel, P. (2022). Depression, anxiety, and stress among high school students: A cross-sectional study in an urban municipality of Kathmandu, Nepal. PLoS Global Public Health, 2(5), e0000516.
19. Kim, S., Kimber, M., Boyle, M. H., & Georgiades, K. (2019). Sex differences in the association between cyberbullying victimization and mental health, substance use, and suicidal ideation in adolescents. The Canadian Journal of Psychiatry, 64(2), 126-135.

20. Kim, S., Razi, A., Stringhini, G., Wisniewski, P. J., & De Choudhury, M. (2021). A human-centered systematic literature review of cyberbullying detection algorithms. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-34.
21. Lin, S. Y., Kung, Y. C., & Leu, F. Y. (2022). Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis. Information Processing & Management, 59(2), 102872.
22. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
24. Nixon, C. L. (2014). Current perspectives: the impact of cyberbullying on adolescent health. Adolescent health, medicine and therapeutics, 143-158.
25. Office for National statistics (2020) Online Bullying in England and Wales; year ending March 2020 https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/onlinebullyingineng landandwales/yearendingmarch2020 Accessed on the 16th March, 2023.
26. Ogunleye, B. O. (2021). Statistical learning approaches to sentiment analysis in the Nigerian banking context (Doctoral dissertation, Sheffield Hallam University).
27. Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. Applied Sciences, 13(2), 797.
28. Paul, S., & Saha, S. (2022). CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. Multimedia Systems, 28(6), 1897-1904.
29. Paul, S., & Saha, S. (2022). CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. Multimedia Systems, 28(6), 1897-1904.
30. Peng, Z., Klomek, A. B., Li, L., Su, X., Sillanmäki, L., Chudal, R., & Sourander, A. (2019). Associations between Chinese adolescents subjected to traditional and cyber bullying and suicidal ideation, self-harm and suicide attempts. BMC psychiatry, 19(1), 1-8.
31. Pew Research Center (2018) A Majority of Teens Have Experienced Some Form of Cyberbullying. https://www. pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/ Accessed on the 16th March, 2023.
32. Pew Research Center (2022) Teens and Cyberbullying. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/ Accessed on the 16th March, 2023.
33. Piccoli, V., Carnaghi, A., Bianchi, M., & Grassi, M. (2022). Perceived-Social Isolation and Cyberbullying Involvement: The Role of Online Social Interaction. International Journal of Cyber Behavior, Psychology and Learning (IJCBPL), 12(1), 1-14.
34. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
35. Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In 2011 10th International Conference on Machine learning and applications and workshops (Vol. 2, pp. 241-244). IEEE.
36. Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. P. (2018, July). A "deeper" look at detecting cyberbullying in social networks. In 2018 international joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.
37. Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., ... & Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review. Computers in Human Behavior, 93, 333-345.
38. Statista (2021) Share of adult internet users in the United States who have personally experienced online harassment as of January 2021. https://www.statista.com/statistics/333942/us-internet-online-harassment-severity/ Accessed on the 16th March, 2023.
39. Suhas-Bharadwaj, R., Kuzhalvaimozhi, S., & Vedavathi, N. (2023). A Novel Multimodal Hybrid Classifier Based Cyberbullying Detection for Social Media Platform. In Data Science and Algorithms in Systems: Proceedings of 6th Computational Methods in Systems and Software 2022, Vol. 2 (pp. 689-699). Cham: Springer International Publishing.
40. Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2022). Pre-Trained Language Models and Their Applications. Engineering.
41. Wang, J., Fu, K., & Lu, C. T. (2020, December). Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 1699-1708). IEEE.
42. Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).

43. Woo, W. H., Chua, H. N., & Gan, M. F. (2023). Cyberbullying Conceptualization, Characterization and Detection in social media–A Systematic Literature Review. International Journal on Perceptive and Cognitive Computing, 9(1), 101-121.
44. Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012, June). Learning from bullying traces in social media. In Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies (pp. 656-666).
45. Yadav, J., Kumar, D., & Chauhan, D. (2020, July). Cyberbullying detection using pre-trained bert model. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1096-1100). IEEE.
46. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.
47. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.
48. Yi, P., & Zubiaga, A. (2022). Session-based Cyberbullying Detection in Social Media: A Survey. arXiv preprint arXiv:2207.10639.
49. Yi, P., & Zubiaga, A. (2022, May). Cyberbullying detection across social media platforms via platform-aware adversarial encoding. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 1430-1434).
50. Zaborskis, A., Ilionsky, G., Tesler, R., & Heinz, A. (2018). The association between cyberbullying, school bullying, and suicidality among adolescents. Crisis.