Self-Training for Cyberbully Detection: Achieving High Accuracy with a Balanced Multi-Class Dataset

Mohamad Ahmadinejad
Department of Computer Science
University of Regina
Regina, SK, Canada
mha808@uregina.ca

Nashid Shahriar
Department of Computer Science
University of Regina
Regina, SK, Canada
nashid.shahriar@uregina.ca

Lisa Fan
Department of Computer Science
University of Regina
Regina, SK, Canada
lisa.fan@uregina.ca

***Abstract*—Cyberbullying is a growing concern in the digital age, causing significant harm to its victims. The ability to automatically detect cyberbullying in social media is crucial for protecting vulnerable individuals. In this study, we propose a Machine Learning/Deep Learning-based approach for cyberbully detection in social media. The approach involves curating a balanced dataset for training the model and implementing a semi-supervised self-training algorithm for increasing the size of the labeled dataset. The model is trained and tested on real-world social media data, and its performance is evaluated using various metrics such as precision, recall, and F1-score. We present our annotated dataset of 99,991 tweets, which we make publicly available for future scientific investigation. The results show that the proposed approach outperforms state-of-the-art methods for cyberbully detection, demonstrating the effectiveness of Machine Learning/Deep Learning techniques for this problem. The findings from this study provide insights for future research in this area and have practical implications for developing automated systems for detecting cyberbullying in social media.**

***Index Terms*—Cyberbully detection, Self-training, Machine Learning, Deep Learning, Text Classification**

I. Introduction

While social media provide excellent communication capabilities, they also make people more susceptible to cyberbullying. Recent research indicates that cyberbullying is becoming an increasing concern among young people that can negatively impact their mental health [1]–[3]. With so much information flowing in online social media platforms, it is near to impossible for humans to monitor any platform for cyberbullying. Artificial intelligence (AI), in particular, Machine Learning (ML) approaches have shown promises for automated detection of cyberbullying on social media platforms [4]. Cyberbully detection is a binary classification problem that entails marking a content (e.g., status) as either cyberbullying or non-cyberbullying, resulting in two possible classes or labels. The multi-classification problem in the context of cyberbullying, on the other hand, entails classifying a content into multiple predefined categories or labels (e.g., race, gender, religion) identifying the context of the cyberbully.

One of the main challenges in developing ML models for cyberbullying detection is the lack of availability of labeled data. Labeled data is necessary for training ML models, but is often difficult and time-consuming to obtain in large quantities let alone with proper balance among classes. Moreover, the cost of manual labeling by human annotators is high, which limits the size of labeled datasets. This can result in a lack of diversity in the data making the models prone to overfitting [5]. Another challenge in developing ML-based cyberbully detection is that existing cyberbully datasets are highly imbalanced which can lead to poor class-based performance. To address these challenges, in this paper, we first propose a self-training based cyberbully dataset generation approach that utilizes a smaller labeled dataset from existing literature to guide the labeling of additional unlabeled contents gathered from a popular social media platform such as Twitter [7]. Using the proposed approach, we generated a large and balanced dataset for the task of cyberbully detection in social media and make the dataset publicly accessible. We also present an effective multi-class classification strategy to enhance the performance of cyberbully detection and evaluate it with several ML and deep learning models on our dataset. Specifically, we make the following contributions in this paper:

- **Devising a self-training based mechanism to generate a labeled and balanced dataset**: We propose a self-training based data annotation approach that can annotate unlabeled contents with high prediction confidence. The approach can be used to augment a labeled dataset with additional labeled data eliminating the imbalance and introducing diversity into the dataset. This can help ML models better generalize to new and unseen data and reduce the effects of bias and overfitting that can occur when training on imbalanced data.
- **Improving self-training reliability**: A problem with self-training is that the predictions made by a self-trained model on the unlabeled sample may not be accurate. To address this issue, we used multiple ML/Deep Learning models with different architectures and hyperparameters to make predictions on unlabeled samples. We leveraged ideas from co-training [8] and tri-training [9] to make the newly labeled dataset more reliable, but instead of two or three classifiers, we used six different models, and only the samples on which all six classifiers agreed are added to the labeled set. The samples of newly labeled data were then verified by social media experts.

- **Publishing the balanced and labeled dataset**: Since most publicly accessible labeled datasets for cyberbully detection have a small number of records (the largest dataset available has around 62,000 records [10]), we have provided a balanced dataset with a reasonably large (e.g., 99,991) number of records for future research [1].
- **Proposing a two-phase multi-class classification**: When the classes in the multi-class classification are not of the same level, i.e., there is a non-cyberbully class as well as different types of cyberbully in the target labels, the classification performance can be improved by splitting the problem into binary-target classification and multi-class classification among cyberbully types.
- **Cyberbully detection with high accuracy**: Finally, several supervised learning algorithms were applied on the curated dataset, yielding high accuracy that could be employed in real-world applications.

The remainder of the paper is structured as follows: Section 2 reviews related works. The self-training-based dataset generation approach and multi-class classification method are discussed in Section 3. Comparing different models and scenarios, Section 4 investigates evaluation settings and outcomes. In the final section of the paper, section 5, future research directions are discussed.

II. Related Work

Previous work, such as [11], has explored the underpinnings of ML and AI in depth, thus laying a critical foundation for our exploration of various models for cyberbully detection.

There have been numerous studies on the use of ML models for cyberbully detection, and while these models have shown promise, they also have several limitations and shortcomings.

The majority of ML approaches rely on supervised [12] [13] [14] or semi-supervised learning [15]. The former includes building a classifier from labeled training data, whereas semi-supervised techniques rely on classifiers generated from a training corpus that contains a small number of labeled and a large number of unlabeled examples. To identify disturbed youths, [12] employed three labels: sexuality, race and culture, and intelligence for binary classification and a combination of these three labels for multi-class classification. Their label-specific binary classifiers outperformed multi-class classifiers by using domain-specific content features learned from training classifiers on a set of messages clustered on sensitive topics such as race, culture, sexuality, and intelligence to detect bullying messages within each cluster. Their discovery demonstrates that bullying, including intentional abuse and vulgarity, was significantly easier to identify than bullying, incorporating sarcasm and euphemism.

For each labeled message, [13] employed unigrams, bigrams, and trigrams for feature engineering. They also used counts from the Linguistic Inquiry and Word Count (LIWC) lexicon to assess the text's linguistic and psychological processes. They used several cyberbully criteria, such as aggression, repetition, harmful intent, peer visibility, and target power. Authors in [14] presented a system for extracting semantic characteristics from texts and using ML classification models (Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Random Forest) to predict if a text is low-level cyberbullied, non-cyberbullied, medium-level cyberbullied, or high-level cyberbullied. They employed the SMOTE oversampling approach to address the data imbalance problem. Authors in [6] used balanced data and graph-based classifiers to detect multi-class cyberbully. Similarly, [10] introduced a new English Twitter-based dataset for online abuse and cyberbully identification using query phrases targeting various forms of bullying and offensive content. In contrast, [16] used a chi-square test to show how social media variables might indicate cyberbullying.

Dataset quality and composition constrain ML/Deep Learning models for cyberbullying detection. Previous studies suffered from a shortage of labeled data and employed numerous small and diverse datasets, making the evaluation cumbersome. The context-dependency of cyberbullying makes data labeling problematic. The majority of approaches rely on manually crafted features, which can result in inaccurate and biased results [17]. Previous research on cyberbullying detection also lacks empirical evidence. Many studies employ artificially generated or annotated datasets, which may not accurately reflect the dynamics of online abuse in the real world [18], [19]. When models perform well on training data but unfavorably on new data, this is an example of overfitting. Additionally, prior research on cyberbullying detection has been hindered by imbalanced datasets in which one class is highly skewed [14] [20]. The trained model on such datasets may be biased toward the dominant class and fail to identify instances of minority classes, leading to inaccurate results. Existing methods for detecting the severity of cyberbully cannot distinguish between various severity levels [20]. More research is needed to collect and annotate vast and diverse datasets that authentically depict the complexities of cyberbully detection.

In conclusion, the lack of labeled, representative, and balanced datasets to be used to train ML/Deep Learning models for cyberbully detection are critical limitations that can have a significant impact on the performance of these models. In this paper, we address these challenges by utilizing self-training techniques that allow us to generate a balanced labeled dataset for training robust multi-class classification models.

III. Methodology

To detect cyberbullying tweets, a sufficiently big labeled dataset with classes covering the various aspects of cyberbully is required. In this section, we describe our methodology to create a balanced multi-class cyberbully dataset using self-training followed by an effective multi-class classification strategy to be applied to the dataset. Fig. 1 depicts the progression of our methodology.

A. Curating a Balanced Dataset using Self-training

1) Problem Statement: Curating a balanced dataset for cyberbullying detection in social media is a significant problem

[1] Dataset Link

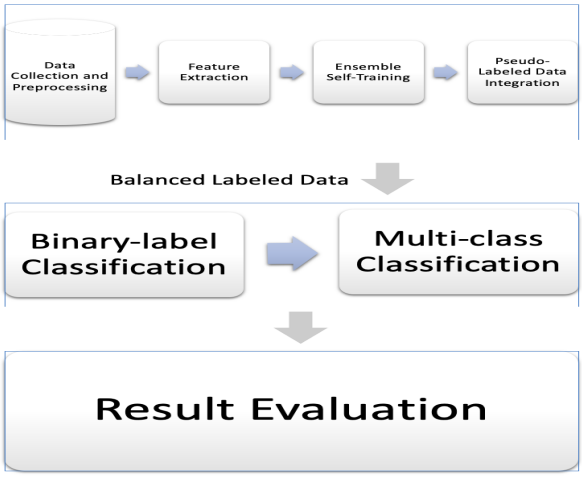


Fig. 1: Methodology Flow Diagram

in developing successful ML models for identifying and combating cyberbullying. Challenges associated with the generation of a cyberbullying dataset include data imbalance, correct annotation, and an insufficient number of labeled datasets. The varying degrees of negativity and positivity in social media content, the time-consuming and error-prone process of labeling and annotation, the need for data filtering and preprocessing for quality assurance, and the lack of publicly accessible multi-class labeled datasets with sufficient records despite existing initiatives all contribute to these challenges. To address these challenges, it is essential to curate a balanced dataset that is representative of the target population and includes a diverse range of positive and negative examples of social media posts and comments. This requires a thorough and systematic approach to data collection, preprocessing, annotation, and balancing that we discuss in the following.

2) Data Collection and Pre-Processing: The data collection for this study was done in two phases. The first phase involved collecting labeled data from previous studies on cyberbullying detection in social media [6] [10] [16]. This data was used to train an ML model, which was then applied to new unlabeled data collected in the second phase. In the second phase, new unlabeled data was collected from Twitter using the Twitter API. We collected roughly 4,000,0000 tweets covering a diverse range of topics over the course of two months.

The tweet data collected from previous work and the Twitter API can contain irrelevant information, typos, and misspelled words, which need to be cleaned and processed before being used for data annotation. This includes removing irrelevant information such as URL links, punctuation marks, special characters, emojis, hashtags, and extra white spaces and converting all text to lowercase to reduce the impact of capitalization on our analysis. Then the stop words, such as 'the', 'is', 'an' were removed as these words do not add any meaningful information. In the next step, words were reduced to their base or root form (Stemming/ Lemmatization [21]). This helps in reducing the size of the vocabulary and also helps in reducing the sparsity of the data. Finally, the text data was transformed into a numerical representation using the Feature Extraction process described below.

3) Feature Extraction: Feature extraction techniques transform the raw text data into a numerical representation that captures the data's meaningful information and reduces the feature space's dimensionality. This helps to improve the performance of Natural Language Processing (NLP) models by reducing the noise in the data and increasing the signal-to-noise ratio [22]. **Bag-of-Words (BoW)** and **TF-IDF** are two of the most commonly used text representation techniques that were used in this work.

Embedding methods, on the other hand, map text data into a continuous high-dimensional vector representation that captures the semantic meaning and relationships between words. This allows ML models to capture the meaning of the text and make more informed predictions [23]. **Global Vectors for Word Representation (GloVe)** [24], **Keras Embedding (KerasE)** [25], and **BERT (Bidirectional Encoder Representations from Transformers)** [26] were used as embedding methods. There are several pre-trained language models in the BERT family models that we investigated, and "bert-base-cased" [26] produced the best results. The term "cased" in the model name indicates that the model is case-sensitive, meaning that it distinguishes between upper and lower case letters in words, while the word "base" refers to the model architecture's size, which is either base or large.

4) Ensemble Self-training: Ensemble Self-training step involves selecting a set of best-performing ML/Deep Learning models and annotating unlabeled tweets collected using the Twitter API. In addition, there are two important parameters in model selection and in our self-training process:

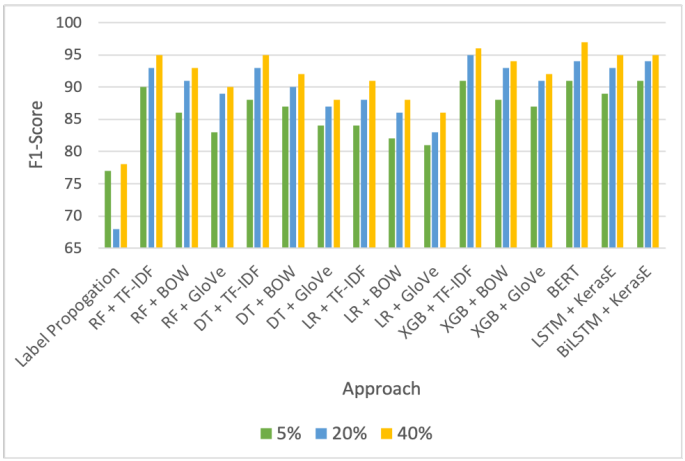
- **Pre-Train Size**: it refers to the percentage of labeled examples used to initially train the model before the self-training process begins. A smaller pre-train size typically results in more unlabeled data being labeled during the self-training process, as the model will have less initial data to deal with. However, this can also make the model more susceptible to overfitting, especially if the unlabeled data is noisy or unrepresentative of the labeled data.
- **Prediction Confidence Threshold (PCT)**: it refers to the level of confidence required for a predicted label to be added to the labeled data. This threshold is typically a value between 0 and 1, and it determines how confident the model needs to be in its prediction before adding it to the labeled data. A higher prediction confidence threshold will result in fewer examples being added to the labeled dataset, resulting in a smaller percentage of unlabeled data being labeled. This is because the model will only add examples with high confidence and can miss some useful examples. On the other hand, a lower prediction confidence threshold means that more predicted labels will be added to the labeled data, resulting in a larger percentage of unlabeled data being labeled. However, this may also lead to the addition of less reliable examples to

the labeled data, which can lead to overfitting and poor model performance.

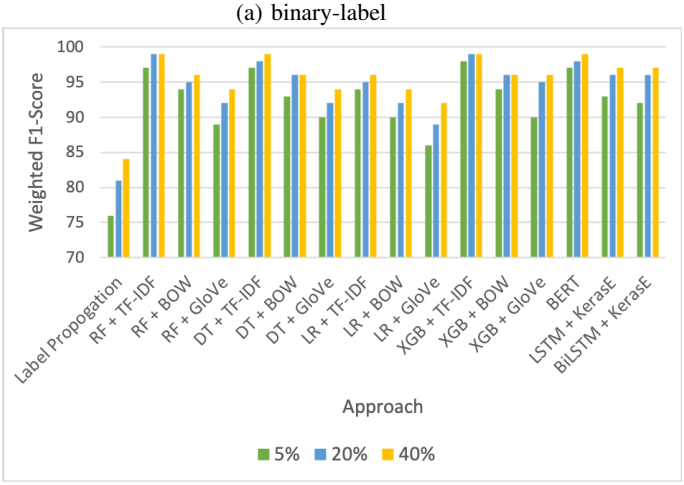
As ML/Deep learning model and these two parameters can impact the percentage of unlabeled data that our self-training models can label, they need to be tuned carefully to achieve optimal performance. In the following, we describe our analysis that helps select ML/Deep learning models as well as parameters.

a) Model Selection for Self-training: In this step, we performed self-training on a balanced and labeled dataset in order to select the best ML/Deep Learning models along with feature extraction methods to be used in our self-training task with unlabeled tweets. Since existing datasets are highly imbalanced among cyberbully and non-cyberbully classes, we have combined tweets from different sources [6], [27], and [28] to create a balanced and labeled dataset. The dataset has around 34,000 records, out of which 17,000, 5,500, 5,500, and 5,500 records correspond to non-cyberbullying, religion, ethnicity/race, and gender/sexual class, respectively, making them balanced among cyberbully and non-cyberbully classes. Afterward, we train different ML/Deep Learning models and feature extraction methods with labeled tweets of several Pre-Train Sizes from the dataset and allow the models to label the remaining tweets in the dataset in a self-training manner. For instance, for 20% Pre-train size, each model was trained with around 7,000 labeled tweets and asked to label the remaining 27,000 tweets. Then we evaluated the models' performances in terms of their F1-Score of labeling and selected the best-performing models and feature extraction methods. AutoML (Automated Machine Learning) was employed for this task which automates the selection of the optimal ML/Deep Learning model for a given dataset. We used Auto-Sklearn [29] to discover the top ML models while Auto-Keras [30] was used to determine the appropriate model architecture for Deep Learning models.

Among the classifiers we evaluated in this step, **Random Forest (RF)**, Decision Tree (DT) and **Logistic Regression (LR)** were implemented using sci-kit learn [31]; **XGBOOST (XGB)** was deployed using its open source library [32]; **Long Short-Term Memory (LSTM)** and **Bidirectional Long Short-Term Memory (BiLSTM) models** were built using Keras, and **BERT** was used as a classifier by fine-tuning its pre-trained models on the labeled dataset. This fine-tuning process involved training the last layer of the BERT model with a text classification task. We combined these ML/Deep Learning models with feature extraction methods discussed earlier. To establish a baseline approach, we also tested a semi-supervised Label Propagation approach. All the models were additionally tuned using grid search to ensure they have the optimal hyperparameters. Fig. 2 presents the F1-Score of labeling achieved by different approaches for both binary and multi-classification tasks with a fixed PCT of 0.7. We have also measured other metrics such as accuracy and the trend is similar. Fig. 2 shows that out of all approaches, RF with TF-IDF, DT with TF-IDF, XGB with TF-IDF, BERT using its own tokenizer, LSTM with KerasE, and BiLSTM with KerasE have



(a) binary-label



(b) multi-label

Fig. 2: F1-Score of labeling obtained by various methods with varying pre-train sizes (5%, 20%, and 40%) and PCT = 0.7

superior performances. Based on this evaluation, we selected RF, DT, and XGB combined with TF-IDF; BERT with its own tokenizer; and LSTM and BiLSTM with KerasE for our self-training task with unlabeled tweets. Since the performance is worst at 5% Pre-train size, we selected 20% and 40% Pre-train size for further analysis.

b) Data Annotation with Self-training: This is the most critical step of our dataset generation as it involves annotating unlabeled tweets with accurate cyberbully classes using a self-training approach. The self-training algorithm was performed with the top six models (RF with TF-IDF, DT with TF-IDF, XGB with TF-IDF, BERT using its own tokenizer, LSTM with KerasE, and BiLSTM with KerasE) we selected in the previous step. For each of the six models, training was carried out using the balanced dataset of 34,000 tweets. Then, in an iterative process, each model was tasked with providing pseudo labels to the 4,000,000 unlabeled tweets. Afterward, we adopted a majority voting strategy that considers the pseudo labels given to each tweet by the six models. Our voting strategy would label a tweet with a pseudo label only when all six models assigned the same pseudo label to the tweet. Otherwise, the

tweet is ignored from adding to our target dataset. Such a strong voting strategy was chosen to ensure that the labeled data in our target dataset is consistent across all the models and to reduce the risk of having incorrect or inconsistent labels in the dataset.

TABLE I and II show the F1-Score and percentage of unlabeled data that could be labeled with different models at different confidence levels with a fixed 20% Pre-train size. TABLEs III and IV demonstrate the same assessment for a Pre-train size of 40%. After careful analysis of these results, we chose a 20% pre-train size with a PCT of 0.7. As a result, our 34,000 labeled records were used as pre-training and a set of 136,000 unlabeled records was fed for pseudo-labeling in each self-training stage. This generated a dataset of 2,400,000 labeled tweets. Intuitively, in the generated dataset the majority of general tweets were of the non-cyberbully class and the number of tweets pertaining to cyberbully types was not equal creating an imbalanced issue. Hence, we randomly chose around 99,991 tweets to create a balanced dataset that not only has an equal number of tweets from both non-cyberbully and cyberbully classes but also has an equal number (approximately 17,000) of tweets from cyberbully forms (e.g., Gender/Sexual, Religion, and Ethnicity/Race). Although we created a dataset of 99,991 tweets, one can easily expand the dataset using the proposed self-training approach avoiding the costly human annotation process.

c) Dataset Verification: To assure the accuracy of the newly labeled data, a total of five batches, each containing 1,000 tweets, were randomly selected. These samples were then distributed to three social media specialists with experience in detecting cyberbully. The experts were instructed to confirm the labels' veracity and provide feedback. After analyzing the data, the specialists determined that the accuracy of the classifications exceeded 90%. This procedure assisted us in ensuring the accuracy of the labeled data and minimizing any possible biases in the data. The feedback was also used to examine the contents of the tweets with false labels and to clean the dataset.

Fig. 3 presents the word clouds generated from the labeled tweets of our dataset for the three cyberbully forms. Word clouds visually represent the frequency of words within the text, with larger font sizes indicating higher frequencies. The word clouds provide a quick overview of the most common words used in each class and can help identify the underlying themes and patterns. The distinctive word clouds of each class demonstrate the differences in languages and contents used in different cyberbully types.

TABLE I: F1-Score of labeling obtained by various methods with varying PCTs (with 20% pre-train)

Model/ Threshold	0.5	0.6	0.7	0.8	0.9
RF + TF-IDF	93	93	95	95	97
DT + TF-IDF	93	94	95	95	97
XGB + TF-IDF	94	94	95	96	96
BERT	93	93	94	95	97
LSTM + KerasE	90	91	93	93	93
BiLSTM + KerasE	92	92	94	95	95

TABLE II: Percentage of unlabeled data that various approaches with varying PCTs could label (with 20% pre-train)

Model/ Threshold	0.5	0.6	0.7	0.8	0.9
RF + TF-IDF	100	99	98	93	86
DT + TF-IDF	100	99	98	94	88
XGB + TF-IDF	100	99	99	98	96
BERT	99	99	98	95	90
LSTM + KerasE	100	100	99	99	98
BiLSTM + KerasE	100	99	99	99	97

TABLE III: F1-Score of labeling obtained by various methods with varying PCTs. (with 40% pre-train)

Model/ Threshold	0.5	0.6	0.7	0.8	0.9
RF + TF-IDF	92	94	96	97	98
DT + TF-IDF	94	95	96	97	98
XGB + TF-IDF	92	94	96	96	97
BERT	94	95	97	98	99
LSTM + KerasE	92	93	94	95	96
BiLSTM + KerasE	91	92	95	96	97

TABLE IV: Percentage of unlabeled data that various approaches with varying PCTs could label (with 40% pre-train)

Model/ Threshold	0.5	0.6	0.7	0.8	0.9
RF + TF-IDF	100	97	98	90	79
DT + TF-IDF	100	99	99	95	89
XGB + TF-IDF	100	98	99	92	80
BERT	100	99	98	95	87
LSTM + KerasE	100	99	99	96	82
BiLSTM + KerasE	100	99	99	96	82

B. Multi-Class Classification

1) Problem Statement: Detecting cyberbully in tweets can be formulated as a multi-class classification problem, where tweets are classified into different categories such as non-cyberbully and specific cyberbully based on religion, gender, and race/ethnicity. This is a supervised learning task as we assume that the input dataset to this problem is already labeled and balanced. Otherwise, the dataset can be curated with a process described in the previous sub-section. In this sub-section, we present a multi-class classification approach that can accurately classify tweets in any labeled cyberbully dataset.

2) Classification Approach: It is difficult to find a balanced dataset for the cyberbullying multi-classification problem where the four classes (non-cyberbully, religion, gender, and race/ethnicity) have an abundance of records. This is due to the fact that cyberbullying is not a usual phenomenon and each cyberbully type has different probability to occur. We observed similar behavior in our generated dataset that is skewed towards non-cyberbully class. To address such an imbalance issue, we could downsample the non-cyberbully class reducing the diversity of data. Alternatively, we could upsample cyberbully classes using synthetically generated data. Since both downsampling and upsampling with synthetic data have their disadvantages, we adopt a two-phase approach that does not rely on upsampling or downsampling. In the first phase, we train a binary classifier to distinguish between

(a) Religion (b) Ethnicity/Race (c) Gender/Sexual

Fig. 3: Word clouds of newly labeled records in different classes

cyberbully and non-cyberbully classes. Since this is a generic task, we can combine data from different cyberbully types into one class making balanced datasets. In the second phase, we train a multi-class classifier only on cyberbully classes which are more likely to be balanced among themselves. Similar two steps would be needed to classify an unlabeled tweet as shown in Fig. 4. Despite the fact that dividing the problem into these two steps would need more time, our evaluation demonstrated that the outcome would be superior.

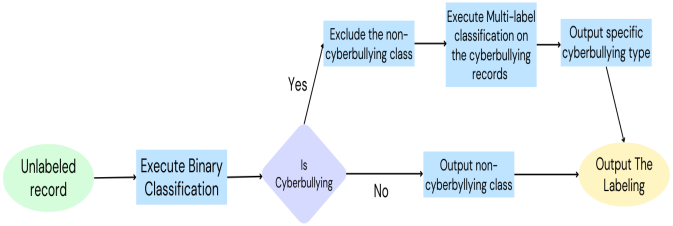


Fig. 4: Two-phase Multi-label Classification

IV. Evaluation Results

Our evaluation first focuses on benchmarking our multi-classification approach against existing strategies and datasets. Afterward, we evaluate different ML and Deep Learning models using our multi-classification approach on the generated dataset. We used several evaluation metrics, including accuracy, precision, recall, and F1-score, to evaluate the performance of the approaches. The experiments were carried out on a machine equipped with an Intel Core i9-12900 CPU, 64GB of memory, and an Nvidia RTX A5000, 3660T GPU.

A. Comparison with Existing Strategies and Dataset

1) Evaluation Settings: For the multi-class classification, [6], [27], and [28] used similar datasets to classify among non-cyberbully, age, gender, ethnicity, and religion classes. Among them, XGB with TF-IDF and BoW as the text representation methods of [6] achieved the best results. Hence, we use XGB with TF-IDF and BoW from [6] as our baseline approach to compare with. The authors in [6] used a balanced dataset with only 8,000 instances per class totaling around 40,000 instances and considered a 75:25 train/test data split. We used the same dataset to reproduce their result as well as to evaluate our multi-classification approach using a 70:10:20 ratio for training, validation, and test sets.

2) Results and Analysis: TABLE V compares reported results in [6] with our implementation of XGB with TF-IDF, BoW, and GloVe as well as BERT. Due to the fact that the authors of [6] only presented F1-score and accuracy as performance metrics, we could only include those metrics in the table. TABLE V shows that our implementation of XGB with TF-IDF and BoW with hyperparameter tuning produces a superior outcome. Additionally, BERT could surpass that result. TABLE VI shows class-wise performances for XGB + TF-IDF revealing that when all the classes are included in the multi-class classification, the non-cyberbully class score is lower than the others. The findings in this table are for XGB + TF-IDF, however, the results for other techniques are similar. This is potentially due to the lack of diversity in the dataset. TABLE VI also shows using our two-phase multi-classification approach, the scores of the non-cyberbully class, denoted as Non-Cyberbully*, could be improved by up to 10% compared with the single-phase multi-classification approach.

TABLE V: Comparison of classification performance using different embedding methods and classifiers

Model/Metric	Weighted Precision	Weighted Recall	Weighted F1-Score	Accuracy
XGB + TFIDF [6]	-	-	93.8	93.7
XGB + BoW [6]	-	-	94.4	94.3
XGB + TFIDF	95.1	94.4	94.7	94.6
XGB + BoW	94.8	93.9	94.5	94.3
XGB + GloVe	92.2	90.4	91.4	91.3
BERT	96.1	95.4	95.7	95.6

TABLE VI: Class Score Using XGB + TF-IDF

Class/ Metric	Precision	Recall	F1-Score
Religion	98	94	96
Age	99	98	99
Ethnicity	99	99	99
Gender	95	87	91
Non-Cyberbully	82	93	87
Non-Cyberbully*	92	97	95

B. Evaluation on Our Generated Balanced Dataset

1) Evaluation Settings: Auto-Sklearn was used to discover the best ML models for our generated balanced dataset with 99,991 records. Auto-Keras was used to determine the appropriate model architecture for Deep Learning models. All the

top-performing models using AutoML and Auto-Keras were evaluated, and the superior models were selected. Moreover, some recent language models such as **RoBERTa** (Robustly Optimized BERT Pretraining Approach) [33], and **GPT-3** (Generative Pre-trained Transformer 3) [34] were examined. The dataset was split randomly into training, validation, and test sets in a 70:10:20 ratio. Additionally, 5-fold cross-validation was performed on the training set to evaluate the models' performance. The cross-validation folds were created randomly, with each fold containing an equal number of samples from each class. The models were trained on the training set and evaluated on both the validation and test sets, with the final reported performance being the average over the 5 folds. Finally, a grid search was carried out for hyperparameter tuning to ensure that the models utilized the most suitable hyperparameters.

2) Results and Analysis: TABLE VII and TABLE VIII show the performance metrics for binary and multi-class classification using 5-fold cross-validation, with the superior results being boldfaced. These two tables show that much better classification performances can be achieved using our balanced dataset compared to the same achieved using existing datasets (see TABLE V). The tables also show that the balanced dataset enables chosen ML and Deep Learning models to perform near perfection with XGB, BERT, and RoBERTa being the best performers. Overall, the results suggest that BERT is a more suitable model for cyberbully detection compared to XGB and RF. However, the good results obtained by XGB highlight the potential of ML models for this task.

We also analyzed the performance of BERT using confusion matrices in Fig. 5 and Fig. 6. These figures verify that BERT has high precision and recall for detecting cyberbullying contents in our balanced dataset.

TABLE VII: Binary Classification using Cross Validation

Model/ Metric	Precision	Recall	F1-score	Accuracy
RF + TF-IDF	99.40	99.40	99.40	99.40
DT + TF-IDF	99.50	99.40	99.50	99.50
XGB + TF-IDF	99.60	99.60	99.60	99.60
LSTM + KerasE	99.50	99.50	99.50	99.50
BiLSTM + KerasE	99.40	99.50	99.50	99.50
BERT	99.70	**99.70**	**99.70**	**99.70**
RoBERTa	**99.80**	99.60	**99.70**	**99.70**
GPT3	99.50	99.60	99.50	99.50

TABLE VIII: Multi-label Classification with only Cyberbullying classes using Cross Validation

Model/ Metric	Weighted Precision	Weighted Recall	Weighted F1-score	Accuracy
RF + TF-IDF	99.50	**99.80**	99.70	99.70
DT + TF-IDF	99.60	**99.80**	99.70	99.70
XGB + TF-IDF	**99.80**	**99.80**	**99.80**	**99.80**
LSTM + KerasE	99.70	99.60	99.60	99.60
BiLSTM + KerasE	99.50	99.60	99.60	99.60
BERT	**99.80**	**99.80**	**99.80**	**99.80**
RoBERTa	**99.80**	99.70	99.80	**99.80**
GPT3	99.70	99.70	99.70	99.70

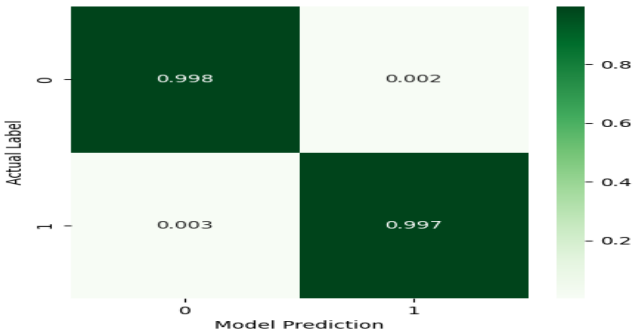


Fig. 5: BERT Binary-Label Confusion Matrix

Fig. 6: BERT Multi-Label Confusion Matrix

Finally, in the real world, cyberbullying occurs less frequently than non-cyberbullying, resulting in imbalanced datasets. As a result, a test set representative of real-world settings would yield more credible results. TABLE IX demonstrates that even when the test set has only 10% records of cyberbullying class and 90% records of the non-cyberbullying class, the models, particularly BERT and XGB + TF-IDF, perform well. Similar results were obtained when additional cyberbullying to non-cyberbullying ratios, such as 5 to 95, 20 to 80, and 30 to 70, were also examined.

TABLE IX: Model Performance When the Test Set Is Imbalanced

Model/ Metric	Precision	Recall	F1-Score	Accuracy
RF + TF-IDF	98.50%	99.50%	99.00%	99.60%
DT + TF-IDF	99.10%	99.40%	99.30%	99.60%
XGB + TF-IDF	**99.80%**	**99.70%**	**99.70%**	**99.80%**
BERT	99.50%	99.60%	99.50%	**99.80%**
Roberta	99.40%	99.50%	99.40%	99.70%
GPT3	99.50%	99.50%	99.50%	99.60%

V. ACKNOWLEDGEMENTS

We extend our gratitude to AI4PH[2] for their supportive funding and valuable insights throughout the course of this research. We appreciate their commitment to advancing cyberbully detection and promoting a safer online environment.

[2]https://ai4ph-hrtp.ca/

VI. Conclusion

Cyberbullying is a critical problem in today's society, and detecting it automatically requires developing effective methods. This paper examines the use of ML and deep learning models for automated cyberbully detection. The paper highlights the challenges faced in this task, such as limited labeled data and imbalanced datasets. To address these issues, we propose a self-training-based approach to annotate unlabeled data collected from a popular social media platform and generate a large balanced dataset published on the Internet. Furthermore, we present an effective multi-class classification strategy to enhance the performance of cyberbully detection and evaluate it with several ML and deep learning models on our dataset. Our evaluation results suggest that with sufficient data and appropriate preprocessing, ML and Deep Learning approaches can effectively detect cyberbullying on social media platforms. We believe our proposed approach and the dataset will ignite further research to address the critical problem of cyberbully detection. The proposed model can be deployed as a web form or a browser extension, allowing users to input text and receive predictions regarding whether the text contains cyberbullying or not and, if so, which category of cyberbullying it falls under. In the future, we plan to examine the robustness and generalizability of our method. This could entail evaluating its efficacy in a variety of contexts using diverse datasets from a variety of sources or populations.

References

[1] Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, De Pauw G, Daelemans W, Hoste V. Automatic detection of cyberbullying in social media text. PloS one. 2018 Oct 8;13(10):e0203794.

[2] Nixon CL. Current perspectives: the impact of cyberbullying on adolescent health. Adolescent health, medicine and therapeutics. 2014 Aug 1:143-58.

[3] Nazir S. The rise of bullying as a public health issue. 2018

[4] Bharti S, Yadav AK, Kumar M, Yadav D. Cyberbullying detection from tweets using deep learning. Kybernetes. 2021 Jul 13;51(9):2695-711.

[5] He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009 Jun 26;21(9):1263-84.

[6] Wang J, Fu K, Lu CT. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In2020 IEEE International Conference on Big Data (Big Data) 2020 Dec 10 (pp. 1699-1708). IEEE.

[7] Amini MR, Feofanov V, Pauletto L, Devijver E, Maximov Y. Self-training: A survey. arXiv preprint arXiv:2202.12040. 2022 Feb 24.

[8] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory 1998 Jul 24 (pp. 92-100).

[9] Zhou ZH, Li M. Tri-training: Exploiting unlabeled data using three classifiers. IEEE Transactions on knowledge and Data Engineering. 2005 Sep 26;17(11):1529-41.

[10] Salawu S, Lumsden J, He Y. A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. Association for Computational Linguistics.

[11] Wang Y, Kwong S, Leung H, Lu J, Smith MH, Trajkovic L, Tunstel E, Plataniotis KN, Yen GG, Kinsner W. Brain-inspired systems: A transdisciplinary exploration on cognitive cybernetics, humanity, and systems science toward autonomous artificial intelligence. IEEE Systems, Man, and Cybernetics Magazine. 2020 Jan 15;6(1):6-13.

[12] Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying. InProceedings of the International AAAI Conference on Web and Social Media 2011 (Vol. 5, No. 3, pp. 11-17).

[13] Ziems C, Vigfusson Y, Morstatter F. Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. InProceedings of the International AAAI Conference on Web and Social Media 2020 May 26 (Vol. 14, pp. 808-819).

[14] Talpur BA, O'Sullivan D. Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in twitter. InInformatics 2020 Nov 15 (Vol. 7, No. 4, p. 52). MDPI.

[15] Nahar V, Al-Maskari S, Li X, Pang C. Semi-supervised learning for cyberbullying detection in social networks. InDatabases Theory and Applications: 25th Australasian Database Conference, ADC 2014, Brisbane, QLD, Australia, July 14-16, 2014. Proceedings 25 2014 (pp. 160-171). Springer International Publishing.

[16] Elsafoury F. Cyberbullying datasets. Mendeley. com.[Online]. Available: https://data.mendeley.com/datasets/jf4pzyvnpj/1. 2020 Jun.

[17] Emmery C, Verhoeven B, De Pauw G, Jacobs G, Van Hee C, Lefever E, Desmet B, Hoste V, Daelemans W. Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Language Resources and Evaluation. 2021 Sep;55:597-633.

[18] Elsafoury F, Katsigiannis S, Pervez Z, Ramzan N. When the timeline meets the pipeline: A survey on automated cyberbullying detection. IEEE access. 2021 Jul 21;9:103541-63.

[19] Muneer A, Fati SM. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet. 2020 Oct 29;12(11):187.

[20] Talpur BA, O'Sullivan D. Cyberbullying severity detection: A machine learning approach. PloS one. 2020 Oct 27;15(10):e0240924.

[21] Khyani D, Siddhartha BS, Niveditha NM, Divya BM. An interpretation of lemmatization and stemming in natural language processing. Journal of University of Shanghai for Science and Technology. 2021;22(10):350-7.

[22] Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. InProceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies 2011 Jun (pp. 142-150).

[23] Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, Okruszek L. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Research. 2021 Oct 1;304:114135.

[24] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. InProceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 Oct (pp. 1532-1543).

[25] Chollet F. Deep learning with Python. 2017. Shelter Island, NY: Manning. 2018.

[26] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

[27] Ahmed T, Kabir M, Ivan S, Mahmud H, Hasan K. Am I being bullied on social media? An ensemble approach to categorize cyberbullying. In2021 IEEE international conference on big data (Big data) 2021 Dec 15 (pp. 2442-2453). IEEE.

[28] Mahmud MI, Mamun M, Abdelgawad A. A Deep Analysis of Textual Features Based Cyberbullying Detection Using Machine Learning. In2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT) 2022 Dec 18 (pp. 166-170). IEEE.

[29] Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. Advances in neural information processing systems. 2015;28.

[30] Jin H, Chollet F, Song Q, Hu X. AutoKeras: An AutoML Library for Deep Learning. Journal of Machine Learning Research. 2023;24(6):1-6.

[31] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

[32] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

[33] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. 2019 Jul 26.

[34] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.