International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Natural Language Processing for Cyberbullying Detection

Jerry He[a*], Lisa Chalaguine[b]

[a]Marriotts Ridge High School, 12100 Woodford Drive, Marriottsville 21104, USA

[b]Department of Computer Science, University College London, 66-72 Gower St, London WC1E 6EA, UK

[a]Email: jjhe2019@outlook.com

[b]Email: lisa.chalaguine.16@ucl.ac.uk

Abstract

With the development of digital technologies and the popularity of social media, cyberbullying has become a serious public health concern that can lead to increased risk of mental and behavioral health issues or even suicide. Artificial intelligence like machine learning opens a lot of possibilities to combat cyberbullying, e.g. automatic cyberbullying detection. Most recent research focuses on improving performance by developing complex models that demand more resources and time to run. Those research uses publicly available datasets without carefully evaluating the feasibility and limitations. This study uses natural language processing (NLP) to evaluate the model performance and examine the difference between fine-grained classification and binary classification as well as assess the feasibility and quality of the publicly available dataset. The results show that simple classifier can also achieve similar performance as that of more complex models if appropriate preprocessing is used, and the publicly available dataset may have limitation and quality issues that researchers should consider when using the data.

Keywords: machine learning; natural language processing; cyberbullying.

1. Introduction

Social media has had a profound impact on the way people live and interact with each other. It enables people to connect with friends, family, and strangers worldwide in real-time, regardless of geographic distance. It has also changed the way people consume and share information.

Additionally, social media has significantly impacted people's self-image and self-esteem. The pressure to present a perfect image on social media has increased anxiety and depression among users. Social media naturally became a hotbed for cyberbullying - a form of bullying or harassment in the digital realm, due to its accessibility, ease of communication, anonymity, and impunity.

Received: 9/5/2023
Accepted: 10/10/2023
Published: 10/23/2023

----

* Corresponding author.

According to Stopbullying.gov, "Cyberbullying is bullying over digital devices like cell phones, computers, and tablets. Cyberbullying can occur through SMS, Text, and apps, or online in social media, forums, or gaming where people can view, participate in, or share content".

The consequences of cyberbullying can be devastating for the victims, who often experience anxiety, depression, and other mental health issues. Cyberbullying can also have long-lasting effects on a person's reputation and relationships, and in some extreme cases, it has even led to suicide [1, 5]. Nearly half of U.S. teens ages 13 to 17 (46%) reported ever experiencing at least one of six cyberbullying behaviors and 28% of teens experienced multiple types of cyberbullying according to a survey Pew Research Center conducted in 2022 [6]. Similar results were reported by the European Commission's Joint Research Centre (JRC) [7] based on a survey of more than 6,000 10-18-year-olds in the Summer of 2020, which found that about half (49%) of children had experienced at least one kind of cyberbullying in their lifetime and among the children that have already been the victim of cyberbullying, nearly half (44%) reported an increase of the phenomenon during the Covid-19 spring lockdown.

Social media platforms have taken steps to address cyberbullying, such as setting up policies and enforcing community standards, implementing reporting tools, and using AI technologies to assist in the identification of improper content. For example, Facebook claimed in their Integrity and Transparency Reports for the Third Quarter of 2022 [8] that their proactive rate decreased in Q3 2022 from 76.7% to 67.8% on Facebook, and 87.4% to 84.3% on Instagram for bullying and harassment-related content and attributed the decrease to the bug fix and improved accuracy in their AI technologies. However, more needs to be done to combat cyberbullying, such as educating users on the impact of their online behavior, encouraging bystanders to speak up when they witness cyberbullying, creating safe and supportive online communities, and improving automatic cyberbullying detection.

In this work, the effectiveness and efficiency of applying Natural Language Processing (NLP) to automatically detect cyberbullying is examined. The Twitter cyberbullying dataset created by Wang and his colleagues [9] has been used by other researchers in their published work [10, 12]. First, fine-grained cyberbullying classification is implemented through several classification algorithms, and the outcomes are compared against the results from other publications using the same dataset. The results indicate that simple classical algorithms like logistic regression can also achieve similar accuracy as that of more complex methods when using proper preprocessing. Second, the binary classification (cyberbullying or not cyberbullying as outcome) is implemented on the dataset to compare with the outcome from fine-grained classification. Third, although it has been used by other researchers, the accuracy of the cyberbullying categorization and the limitations of the cyberbullying dataset have not yet been found in any existing research. Human review and computerized review through ChatGPT API are utilized to assess the accuracy of the cyberbullying categorization of the samples extracted from the cyberbullying dataset. The practicality of fine-grained cyberbullying classification is discussed based on social science research on cyberbullying.

The main contributions of this work are: demonstrating that simple classical algorithms like logistic regression can also achieve similar accuracy as those from more complicated approaches that require more computing resources and time to run, which is essential from the perspective of practical application; reassessing the efficacy and limitation of fine-grained cyberbullying classification and pointing out the limitations. Since the goal is to properly detect whether a tweet is cyberbullying or not which is a binary classification, fine-grained classification of cyberbullying based on traits like race, gender, religion, etc. does not necessarily help improve the detection results due to the complexity of cyberbullying categorization in the real world.

The rest of the paper is structured as follows: the related work is reviewed in the next section, followed by the explanation of the dataset and the methodology, and then the results and discussion are presented before the conclusion of this study.

2. Literature review

It's believed that the earliest use of the term cyberbullying was in a 1995 New York Times article on cyberaddiction [13]. The term "cyberbullying" was coined to describe the use of digital technologies to harass, intimidate, or harm others, often through the use of social media, email, or instant messaging. Since then, the concept of cyberbullying has become widely recognized, and there have been numerous efforts to define cyberbullying, identify differences from traditional bullying, set up the measurement, and develop strategies for preventing and responding to cyberbullying behavior. Those efforts set up the foundations (e.g. the traits to determine whether cyberbullying or not) and important features (e.g. cyberbullying types and categorization) that cyberbullying detection should rely on.

Menesini and his colleagues [14] examined the criteria difference for defining traditional bullying and cyberbullying and discussed the measurement challenges given the age, cultural, and linguistic differences. The research on these fundamental concepts and questions of cyberbullying provides the guidance on what traits should be verified for cyberbullying detection. O'Brien and his colleagues [15] presented their findings from a national cyberbullying project which identified some key differences from the previous similar studies which suggested that people exhibited different level of resilience to cyberbullying. The research suggested further studies to be done on the resilience disparities and the gender differences of being cyberbullies. It implies that the impacts may not be balanced across different cyberbullying traits and populations, which should be considered when developing the cyberbullying detection models and datasets for research purpose. Notar and his colleagues [16] did a comprehensive literature review on categorized topics ranging from the definition, reasons, and roles to gender comparisons related to cyberbullying. These research results evidenced the complexities of cyberbullying characterization such as gender differences, which suggested that simply defining several cyberbullying categories like age, gender, etc., and treating them equally may not work well in cyberbullying detection. It is also worth noting that it is more difficult to determine the intention attribute of cyberbullying compared to physical bullying, and the repetition and imbalance of power attributes were rarely included in the existing machine learning approaches for detecting cyberbullying. Cyberbullying is a broader term that encompasses various forms of online harassment and abusive behavior that is closely related to and shares some common traits with several other types of negative or harmful cyber behaviors such as: cyberhate, cyberharassment, cyberstalking, online trolling, flaming, doxxing, swatting, online shaming, and digital abuse. Clearly defining cyberbullying and the fine differences from other similar cyber abusive behaviors is critical to improve the accuracy of cyberbullying detection. Therefore, we not only need to consider the traits targeted in cyberbullying but also should consider the persistence, specificity, intent, target, and the nature of the harmful behavior involved to develop effective cyberbullying detection models.

There also have been numerous research efforts on applying machine learning to help automatically detect cyberbullying. Much of the research aimed to improve detection accuracy by introducing new features. Balakrishnan and his colleagues [17] experimented with improving cyberbullying detection on a manual annotated Twitter data set with 5453 tweets gathered using a specific hashtag by using Twitter users' psychological features including personalities, sentiments, and emotions. They found that cyberbullying detection improved when personalities and sentiments were used whereas a similar effect was not observed for emotions. The research revealed that introducing specific traits like Twitter users' psychological features helped improve cyberbullying detection. But there are some limitations of this research. First, the dataset used in the research was gathered by using the hashtag #Gamergate which heavily focuses on gender. The relatively small data size and the data with the specific hashtag used in the experiment limit the generality and scope of the obtained results. Muneer and his colleagues [18] compiled a dataset of over thirty-seven thousand tweets from two other datasets created by other researchers in their earlier research on hate speech and offensive language detection. Most tweets included in the hate speech and offensive language datasets were racial or homophobic related. The main focus of the research is to evaluate the performance (accuracy, precision, recall, and F1 score) of seven machine learning classifiers: Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), and Support Vector Machine (SVM). Their research showed that LR had the overall best performance with the highest median accuracy and F1 score, but SGD achieved the best precision and SVM achieved the best recall. It is worth noting that hate speech and offensive language have different impact and legal consequences. Offensive, aggressive, and hate speech are broader than cyberbullying. But this research ignored the difference between hate speech/offensive language and cyberbullying in the data. Alam and his colleagues [19] proposed an ensemble based machine learning approach to tackle the cyberbullying detection. The authors randomly sampled nine thousand tweets from the hate speech and offensive language dataset. The dataset only had two columns, one column contained the original tweet text and the other had 0 or 1 to indicate binary labeling of offensive or not offensive. The sampled dataset had about 54% tweets that were labeled offensive. The authors experimented the combination of four machine learning classifiers and three ensemble models with two different feature extraction techniques. They claimed that their ensemble based approach generated higher accuracy than the traditional machine learning classifiers. It is not clear about the composition of the sampled dataset. Although it seems roughly balanced (54% vs 46%) between the binary categorization, it is unclear whether the offensive tweets represented one specific offensive type or various types. Ensemble methods are usually more computationally expensive and time consuming due to the need for training and storing multiple models, and combining their outputs. Additionally, they can be prone to overfitting and underfitting if the base models are too weak or too strong, or if the aggregation method is too simple or too complex. Many machine learning approaches for cyberbullying detection focused on improving the detection by introducing new features. The feature extraction and feature selection became more complicated when the number of the features increased. Deep learning can automatically learn features which eliminates the need for human intervention.

There were recent studies using deep learning for cyberbullying detection. Al-Ajlan and his colleagues [20] experimented cyberbullying detection by proposing a model based on convolutional neural network (CNN) and incorporating semantics through the use of word embedding. The dataset contained thirty-nine thousand tweets extracted using Twitter streaming API with bad words that were likely to return bullying tweets. The experiment outcomes showed that the deep learning model outperformed the traditional classifier SVM. There are some limitations of this research. The construction of the dataset used for the experiment may be flawed since it used certain bad words to extract the tweets. The coverage of the bad words could be limited and a bully tweet didn't necessarily contain a bad word. Wang and his colleagues [9] proposed a Graph Convolutional Network (GCN) classifier and compared its performance using eight embedding methods and six commonly used classifiers. The authors claimed that the GCN model matched or exceeded the performance of the baseline models based on the results from a relatively small sample (4,000 tweets). Although the authors created a balanced fine-grained cyberbullying dataset with "not cyberbullying" category group included, they didn't include the "not cyberbullying" category group in their experiment. There was no comparison of how the models perform on the fine-grained classification vs. binary classification either.

3. Dataset and methodology

The data used in this study is publically available [21]. Wang and his colleagues collected a total of 39224 tweets from six cyberbullying datasets published by other researchers. They then hand-labeled 5,975 tweets from two out of the six published datasets before classifying them into six groups: age, ethnicity, gender, religion, other cyberbullying, and not cyberbullying (Not CB). The first five groups form the aggregated cyberbullying (CB) group. A modified Dynamic Query Expansion (DQE) was used to increase the number of tweets in each of the six groups. The first two rows of Table 1 below are from Wang and his colleagues [9] paper and the last row presents the ratio of DQE expansion to each group. The expansion is unbalanced across different groups. The other cyberbullying group has less than a 5% change and not cyberbullying group has about 11% change. The group of age, ethnicity, religion, and gender were significantly expanded, ranging from about 60% to nearly 6000%.

Table 1: Dataset summary.

Number of tweets	Total	CB	Not CB	Age	Ethnicity	Gender	Religion	Other
Before DQE	39224	17583	21648	165	272	6489	1933	7717
After DQE	69767	50468	19299	10010	12730	10277	9367	8084
Expansion ratio	77.87%	187.03%	-10.85%	5966.67%	4580.15%	58.38%	384.58%	4.76%

Then 8,000 tweets were randomly sampled from each group of the expanded dataset to form a balanced dataset of a total of 48,000 tweets. The balanced dataset consists of six ASCII text files, one for each group and each text file has 8,000 tweets included. The ASCII text data only contains the tweet text but does not have any flag to indicate which tweets were directly retrieved from other researchers' cyberbullying datasets and which tweets were added later by using the modified DQE. There is no comparison or analysis of whether there is any difference between the tweets obtained from existing datasets and the tweets added by using DQE.

Our study aims to evaluate how the simple classifiers perform on cyberbullying detection, how the introduction of the “not cyberbullying” dataset affects the performance of cyberbullying detection, and the limitation of the fine-grained cyberbullying dataset. In our experiments, we divide the train/test data by an 80:20 ratio. The following Python libraries are used: Pandas library is used for cleaning, and manipulating data; Numpy library is used for performing mathematical operations on large, multi-dimensional arrays and matrices generated from the models; Scikit-learn library is used to implement various classification algorithms; Natural Language Toolkit (NLTK) library is used for tokenization, parsing, stemming, and removing the stop words in the tweets.

We tested the following machine learning classifiers in our experiments: Naive Bayes (NB), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and Logistic Regression (LR) on the dataset with only the five cyberbullying groups (age, ethnicity, gender, religion, and other) included. We then repeat the process by implementing some preprocessing like using various stemming techniques (Porter Stemmer, Snowball Stemmer), and removing the stop words from the tweets, and then re-assessing the performance change. To evaluate the impact of including the “not cyberbullying” category, we randomly sampled 1600 tweets from each of the five cyberbullying groups to form a cyberbullying dataset of 8000 tweets. Then this dataset of 8000 cyberbullying tweets is combined with the “not cyberbullying” data of 8000 tweets to form a balanced dataset for binary classification: cyberbullying vs. not cyberbullying. The processes we ran on the fine-grained dataset are repeated except the classification is binary instead of fine-grained. Since the datasets in our experiments are balanced, accuracy is an appropriate measure of evaluating the models. F1 score (harmonic mean of precision and recall) is also calculated for evaluation.

In addition to assessing the model performance, we also evaluate the quality of the fine-grained cyberbullying dataset. We randomly sampled twenty tweets from each of the six groups and then had five volunteers (with various ages, sexes, races/ethnicities, and educational backgrounds) manually review and classify each sampled tweet as cyberbullying or not. The final classification of each tweet was determined by three or more votes from the five human reviewers. The recent development of the Large Language Model (LLM) such as ChatGPT makes it a promising tool for cyberbullying detection. Due to the cost of using the ChatGPT 3.5 API, we randomly sampled 200 tweets from each of the six groups and then ran them through the ChatGPT 3,5 API. The classification results (cyberbullying or not cyberbullying) returned by ChatGPT are then compared against the original classification. The classification differences are then manually reviewed.

4. Results

Our first experiment is to test four different classifiers (NB, SVM, MLP, and LR) on the dataset of the five cyberbullying groups (age, ethnicity, gender, religion, and other) as explained in the methodology section. We first ran the classification without implementing the preprocessing. The scores for each classifier are listed in the row “Prior Preprocessing”. Then we implemented the preprocessing like removing the stopwords before repeating the processes. The updated scores for each classifier are listed in the row “After Preprocessing”. For comparison, the best scores reported by Wang and his colleagues from their experiments of combining different embedding methods and classifiers are also included in the row “Wang and his colleagues in Table 2 (Accuracy) and Table 3 (F1 Score). The results show that the performance measures (accuracy and F1 score) improved greatly after standard preprocessing like removing the stop words. Although the accuracy and F1 scores before preprocessing are lower than the best score reported by Wang and his colleagues the scores after preprocessing match or exceed the reported best scores. Wang and his colleagues ran their experiments through Google Colaboratory on a Tesla P100 GPU, with 25 GB of RAM and 147 GB of disk space. We ran our experiments through Jupyter Notebook on an Intel i7-12700H CPU, 16 GB of RAM and 512 GB of disk space. The hardware used in Wang and his colleagues experiments exceeds the capacity of our hardware configuration. Comparing the results from our experiments and those reported by Wang and his colleagues shows that, through proper preprocessing, traditional machine learning classifiers like LR can yield similar or even better performance outcomes to more computationally intensive and time-consuming complex methods. This is especially significant when considering available computing resources and practicality.

Table 2: Test accuracies – 40,000 tweets.

Accuracy	NB	SVM	MLP	LR
Wang and his colleagues	0.8265	0.9225	0.9154	0.9033
Prior processing	0.8442	0.8637	0.8736	0.8859
After processing	0.8773	0.9277	0.9238	0.9421

Table 3: Test F1 scores – 40,000 tweets.

F1 score	NB	SVM	MLP	LR
Wang and his colleagues	0.8157	0.9272	0.9153	0.9033
Prior processing	0.8428	0.8615	0.8802	0.8865
After processing	0.8745	0.9212	0.9208	0.9388

Our second experiment is to evaluate how the classifiers perform when including the “not cyberbullying” group in the model. The confusion matrix for the fine-grained classification of all six groups by using LR without preprocessing is presented in Figure 1 below. Table 4 below shows the group-specific scores when we ran the Logistic Regression (LR) classifier on the entire dataset with all six groups (age, ethnicity, gender, religion, other cyberbullying, and not cyberbullying) included. The accuracy is 0.8121.

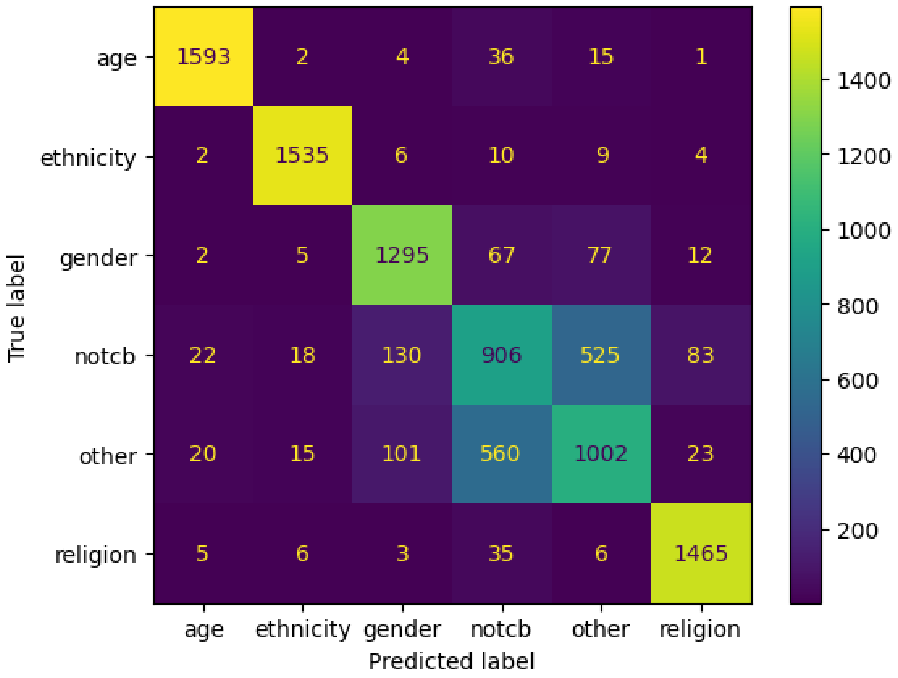

Figure 1: confusion matrix for fine-grained classification using LR without preprocessing.

Table 4: Performance Metrics for Fine-Grained Classification without Preprocessing – 48,000 Tweets.

Group	Precision	Recall	F1 score
Age	0.9649	0.9690	0.9669
Ethnicity	0.9802	0.9709	0.9755
Gender	0.8882	0.8415	0.8642
Religion	0.9638	0.9225	0.9427
Other	0.5822	0.6132	0.5973
Not CB	0.5380	0.5613	0.5494

As In a manner similar to our previous experiment involving the classification of the 5 cyberbullying groups, we conducted preprocessing and then repeated the process to assess any improvements in the results. The confusion matrix can be found in Figure 2, and performance metrics are detailed in Table 5 below. The accuracy saw a slight increase to 0.8283. The results indicate that preprocessing does contribute to performance improvement, although not to the extent observed in experiments that excluded the "not cyberbullying" group. It's worth noting that the "Other Cyberbullying" and "Not Cyberbullying" groups exhibited the lowest performance scores. This suggests that the inclusion of the "Not Cyberbullying" group has had an impact on model performance. Further insights into the potential causes of these effects are elaborated upon in the subsequent Discussion section.

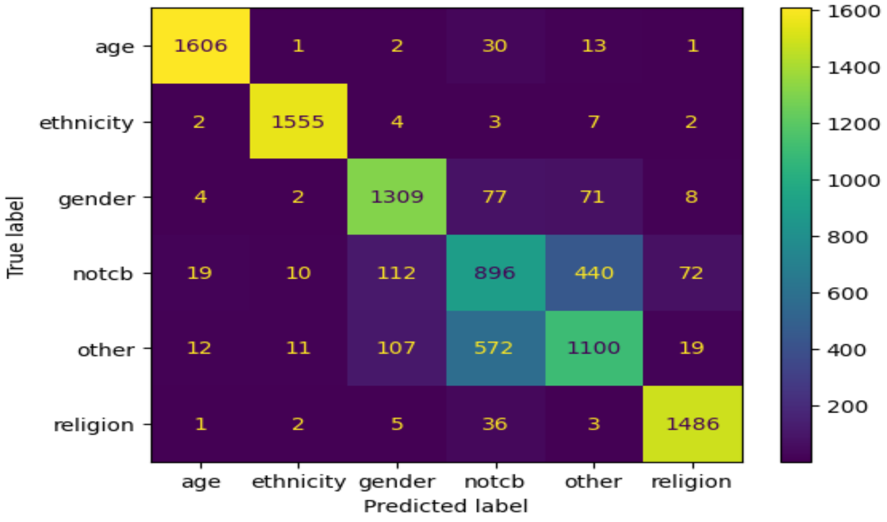

Figure 2: confusion matrix for fine-grained classification using LR with preprocessing.

Table 5: Performance Metrics for Fine-Grained Classification with Preprocessing – 48,000 Tweets.

Group	Precision	Recall	F1 score
Age	0.9716	0.9769	0.9742
Ethnicity	0.9886	0.9836	0.9860
Gender	0.8899	0.8506	0.8698
Religion	0.9693	0.9358	0.9523
Other	0.6041	0.6732	0.6368
Not CB	0.5784	0.5551	0.5666

To gain further insights into how the models perform in binary classification (distinguishing between cyberbullying and not cyberbullying), and to investigate the potential reasons behind the performance drop when including the "not cyberbullying" group, we conducted binary classification on a dataset that incorporated both types of tweets. Given the dataset's composition of 40,000 cyberbullying tweets across five groups and 8,000 tweets from the "not cyberbullying" group, directly merging these two categories would result in an imbalanced dataset, potentially biasing the model. To address this issue, we randomly sampled 1,600 tweets from each of the five cyberbullying groups, creating an 8,000-tweet cyberbullying sub-dataset. This was then combined with the 8,000 "not cyberbullying" tweets. After implementing preprocessing, we performed binary classification using Logistic Regression. The confusion matrix is visualized in Figure 3, and performance metrics are provided in Table 6. The accuracy achieved was 0.8269.

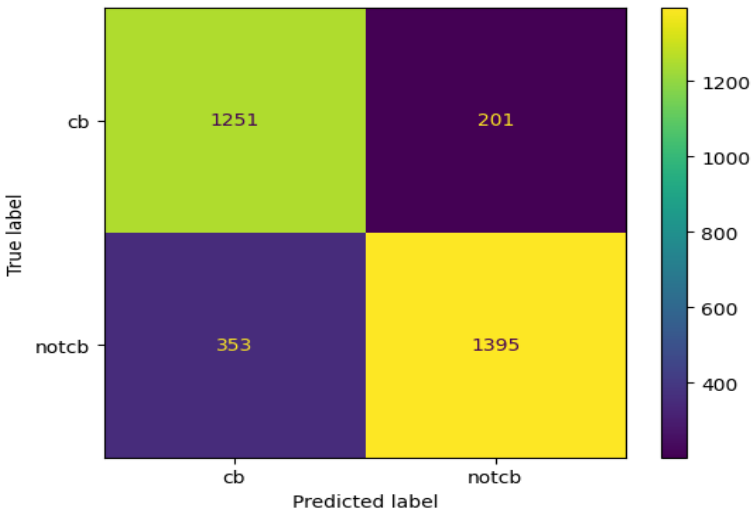

Figure 3: confusion matrix for binary classification using LR with Preprocessing.

Table 6: Performance metrics for binary classification using LR with Preprocessing – 16,000 Tweets

Group	Precision	Recall	F1 score
CB	0.8616	0.7799	0.8187
Not CB	0.7981	0.8741	0.8343

The results of the binary classification on the dataset including the “not cyberbullying” group suggest this group may be related to the performance drop and needs further check. To assess the quality of the fine-grained cyberbullying dataset, we randomly sampled 20 tweets for each of the six groups and had five people manually review each tweet to decide whether it was cyberbullying or not. The choice with three or more votes from the five reviewers will be the human classification. The review results are listed in Table 7 below. The row “Data - CB” represents the original classification as cyberbullying in the dataset. The row “Data - Not CB” means classification as “not cyberbullying” in the dataset. Similarly, the row “Human - CB” stands for classified as cyberbullying by human review, and the row “Human - Not CB” means being classified as “not cyberbullying” by human review. The results show that the groups “Other CB” and “Not CB” have the highest number of discrepancies.

Table 7: Binary classification change with human review – 120 Tweets.

	Age	Ethnicity	Gender	Religion	Other	Not CB
Data - CB	20	20	20	20	20	0
Data - Not CB	0	0	0	0	0	20
Human - CB	18	18	17	19	9	13
Human - Not CB	2	2	3	1	11	7

For the ChatGPT experiment used to evaluate the quality of the fine-grained cyberbullying dataset, the results are included in Table 8 below. The results show that the groups “Other CB” and “Gender” have the highest number of discrepancies. The results from both the human review and ChatGPT classification indicate that the data varieties of the "Other CB" group may introduce additional noise into the data, adversely impacting the model's performance.

Table 8: Binary classification change with ChatGPT – 1200 Tweets.

	Age	Ethnicity	Gender	Religion	Other	Not CB
Data - CB	200	200	200	200	200	0
Data - Not CB	0	0	0	0	0	200
ChatGPT - CB	160	166	127	134	59	56
ChatGPT - Not CB	40	34	73	66	141	144

5. Discussion

Wang and his colleagues [9] described their preprocessing approach, which involved removing links, mentions (@username), the retweet flag "RT," and punctuation before conducting experiments. Notably, they chose not to remove hashtags or stopwords, as they believed these elements might provide valuable context for cyberbullying detection. However, they did not empirically validate this assumption. Our experiment results, as presented in Tables 2 and 3, demonstrate that even basic preprocessing steps, like removing stopwords, can significantly enhance detection performance. While Wang and his colleagues proposed a Graph Convolutional Network (GCN) classifier, they only tested it on a subset of the data (10%, or 4,000 tweets). Complex models often require substantial computational resources and time to execute. In contrast, our results, obtained with the entire dataset (40,000 tweets) following proper preprocessing, reveal that simple classifiers like Logistic Regression (LR) can achieve performance comparable to or even surpass that of more intricate models. This has significant implications for the practicality and efficiency of cyberbullying detection. Wang and his colleagues created a fine-grained cyberbullying dataset comprising five distinct cyberbullying categories and one group labeled as 'not cyberbullying.' They stated their primary interest in cyberbullying tweets and consequently omitted the 'not cyberbullying' group entirely from their experiment. However, for a more comprehensive evaluation, it would have been more compelling if the authors had assessed how well their model could distinguish between cyberbullying and non-cyberbullying, given that each tweet unequivocally falls into one of these two binary categories. Such an evaluation would have revealed the extent to which cyberbullying tweets might have been missed and how many 'not cyberbullying' tweets were included in their fine-grained classification. In our own experiments, when we included the 'not cyberbullying' group, the model's performance decreased compared to when we exclusively employed fine-grained classification. This decrease can partly be attributed to the fact that the 'not cyberbullying' group contains some tweets that indeed qualify as cyberbullying, as confirmed through human review and ChatGPT classification. The human review and ChatGPT classification highlighted certain potential issues with the quality of the fine-grained cyberbullying dataset. For instance, consider the tweet “Last year I invited someone who I would consider one of my "High school bullies" to my wedding. I grew and so did he. People change and grow, it's good. Forgive and love others.” This tweet was categorized as an “age” cyberbullying in the dataset. However, during the human review of the ChatGPT classification output, all five reviewers unanimously agreed that it should not be labeled as cyberbullying. Instead, they found that the tweet conveyed a positive message. Another example is the tweet “Loved every second spent playing games with you. Ur one of the first friends I made on twitch and I am very thankful I met you. I miss you and thanks again for greeting me!” This tweet was tagged as an “ethnicity” cyberbullying in the dataset, but all five human reviewers concluded that it expressed gratitude rather than constituting cyberbullying. During the ChatGPT experiment, a tweet that stated “Racism and rape are not joking matters at all, which are the jokes that he makes. We are not laughing.” was included in the “gender” cyberbullying dataset. However, ChatGPT didn’t classify it as cyberbullying. Human reviewers also agreed that this tweet criticized making jokes about racism and rape and should not be considered as an instance of cyberbullying. It's important to acknowledge that human reviewers did not always reach a consensus on the classification of certain tweets in our experiment. Moreover, there were instances where human reviewers disagreed with ChatGPT's classifications. For instance, there were cases in which ChatGPT did not classify certain tweets containing negative or offensive language as cyberbullying, while the human reviewers held a different perspective. This observation underscores the subjective nature of determining whether a tweet qualifies as cyberbullying. Different individuals, owing to their unique personal experiences, sensitivities, and cultural backgrounds, may interpret online interactions in diverse ways. This subjectivity can lead to differing opinions regarding whether a specific statement or behavior should be considered cyberbullying. Several critical factors come into play when assessing whether an online interaction constitutes cyberbullying, including the tone, intent, frequency, and impact on the recipient. What one person might perceive as harmless banter or a joke could be deeply hurtful and offensive to another. Cultural, social, and individual norms further shape one's perception of what constitutes cyberbullying. Additionally, the context in which a tweet is situated plays a significant role. The same tweet can carry a different meaning in various contexts. Although the dataset created by Wang and his colleagues is a balanced dataset with an equal number of tweets in each cyberbullying classification which helps prevent the model from becoming biased towards one specific class, it should be noted that it is very unlikely that the proposed classification will be represented in equal proportions in the real world. The authors also mentioned that the frequencies of fine-grained classes in the existing dataset before DQE indicated that ageism and racism might not be well represented. Statistics indicate that cyberbullying happens more often among teenagers than adults. Pew Research Center's Report on Teens and Cyberbullying 2022 [6] shows nearly half of U.S. teens ages 13-17 (46%) ever experienced at least one of the six types of cyberbullying behaviors asked about in a survey, compared to 41% of adults in a 2017 study [22]. This Pew Research Center's Report on Teens and Cyberbullying 2022 measures cyberbullying of teens using six distinct behaviors: offensive name-calling, spreading of false rumors about them, receiving explicit images they didn’t ask for, physical threats, constantly being asked where they are, what they’re doing, or who they’re with by someone other than a parent, and having explicit images of them shared without their consent. The percentages of U.S. teens who experienced one of the above six cyberbullying are 32%, 22%, 17%, 15%, 10%, and 7% respectively. There are multiple ways to define cyberbullying types. The types of cyberbullying are not balanced in nature. Also, one single tweet may include multiple cyberbullying classifications instead of just one, e.g. a tweet may include both age and sex or both religion and race/ethnicity. Therefore, it may be problematic to simply classify such tweets in one classification for fine-grained classification. People need to be aware of the limitations when using the fine-grained cyberbullying dataset. This study has some potential limitations. When evaluating the performance of the traditional classifier algorithms before and after preprocessing, we only conducted the experiments with four machine learning algorithms. Although the four algorithms exhibited the similar performance improvement after preprocessing, we wouldn’t conclude that the statement could be generalized to other classifier algorithms without validating through experiments. The cyberbullying detection in this study focuses on textual analysis only. It doesn’t include the cyberbullying detection involving non-textual information such as images, videos, or audios. This will be a potential research topic for our future research. Given the limited time the volunteers could help review the tweets for relabeling, we were only able to relabel 120 tweets through human review. Due to the cost of using the ChatGPT 3.5 API, we were only able to send 1,200 sample tweets to ChatGPT for cyberbullying detection. Although it was ten-fold larger than the human reviewed samples, it was still relative small sample of the original dataset (48,000 tweets). The small sample of the relabeled tweets might be insufficient to draw statistically significant conclusions. The discrepancies in the categorization of cyberbullying, as identified during the relabeling process through human review and ChatGPT, exposed potential labeling issues within the original dataset.

6. Conclusion

This study serves a threefold purpose. Firstly, we aim to reassess the effectiveness of fine-grained cyberbullying classification and the use of a fine-grained cyberbullying dataset. Our experiments reveal that even simpler classification methods, such as Logistic Regression (LR), can achieve comparable accuracy rates to more computationally demanding techniques. This implies that complex methods, which demand additional computational resources and time, may not always be better solutions. Secondly, we delve into a sample of tweets within the fine-grained cyberbullying dataset, subjecting them to manual review and analysis using the ChatGPT API. The disparities between the reclassification and the original classifications underscore the inherently subjective nature of cyberbullying classification. This subjectivity leads to variations in both the dataset and the outcomes of machine learning models. Lastly, we address the real-world imbalance in cyberbullying types based on existing research. We also highlight the limitations of utilizing a fine-grained cyberbullying dataset. Our findings suggest a need for further research in developing improved methods for constructing cyberbullying datasets and refining cyberbullying classification models.

References

[1] M. Mickey. “A 15-year-old boy died by suicide after relentless cyberbullying, and his parents say the Latin School could have done more to stop it.” Internet: https://www.cbsnews.com/chicago/news/15-year-old-boy-cyberbullying-suicide-latin-school-chicago-lawsuit/, Apr. 25, 2022 [Jan. 19, 2023].

[2] A.J. Willingham. “The family of a teen who died by suicide after being outed by cyberbullies is demanding justice.” Internet: https://www.cnn.com/2019/09/30/us/channing-smith-suicide-cyberbullying-tennessee-trnd/index.html, Sep. 30, 2019 [Jan. 19, 2023].

[3] K. Rosenblatt. “Cyberbullying tragedy: New Jersey family to sue after 12-year-old daughter's suicide.” Internet: https://www.nbcnews.com/news/us-news/new-jersey-family-sue-school-district-after-12-year-old-n788506, Aug. 1, 2017 [Jan. 19, 2023].

[4] D. Grau and J. Rybak. “Bullying: Words Can Kill.” Internet: https://www.cbsnews.com/news/bullying-words-can-kill/, Sep. 23, 2013 [Jan. 19, 2023].

[5] Ryan’s Story Presentation LLC. “Ryan’s Story Presentation.” Internet: https://www.ryanpatrickhalligan.org/about, 2022 [Jan. 19, 2023].

[6] E. A. Vogels. “Teens and Cyberbullying 2022.” Internet: https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/, Dec. 15, 2022 [Apr. 6,

2023].

[7] B. Lobe, A. Velicu, E. Staksrud, S. Chaudron, and R. Di Gioia, How children (10-18) experienced online risks during the Covid-19 lockdown - Spring 2020, EUR 30584 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-29763-5, doi:10.2760/562534, JRC124034..

[8] G. Rosen. “Integrity and Transparency Reports, Third Quarter 2022.” Internet: https://about.fb.com/news/2022/11/integrity-and-transparency-reports-q3-2022/, Nov. 22, 2022 [Feb. 23, 2023].

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

[10] M. Mahmud, M. Mamun, and A. Abdelgawad. “A deep analysis of textual features based cyberbullying detection using machine learning,” in *2022 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, 2022, pp. 166-170.

[11] T.H. Aldhyani, M.H. Al-Adhaileh, and S.N. Alsubari. “Cyberbullying identification system based deep learning algorithms.” *Electronics*, vol. 11, pp. 3273, Oct. 2022.

[12] T. Ahmed, S. Ivan, M. Kabir, H. Mahmud, and K. Hasan. “Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying.” *Social Network Analysis and Mining*, vol. 12, pp. 99, Aug. 2022.

[13] S. Bauman. *Cyberbullying: What counselors need to know*. Alexandria, VA: American Counseling Association, 2011, pp. 204.

[14] E. Menesini and A. Nocentini. “Cyberbullying definition and measurement: Some critical considerations.” *Journal of Psychology*, vol. 217, pp. 230-232, Jan. 2009.

[15] N. O’Brien and T. Moules. “Not sticks and stones but tweets and texts: Findings from a national cyberbullying project.” *Pastoral Care in Education*, vol. 217, pp. 53-65, Mar. 2013.

[16] C.E. Notar, S. Padgett, and J. Roden. “Cyberbullying: A review of the literature.” *Universal Journal of Educational Research*, vol. 1, pp. 1-9, Jun. 2013.

[17] V. Balakrishnan, S. Khan, and H.R. Arabnia. “Improving cyberbullying detection using Twitter users’ psychological features and machine learning.” *Computers & Security*, vol. 90, pp. 101710, Mar. 2020.

[18] A. Muneer and S.M. Fati. “A comparative analysis of machine learning techniques for cyberbullying detection on twitter.” *Future Internet*, vol. 12, pp. 187, Oct. 2020.

[19] K.S. Alam, S. Bhowmik, and P.R.K. Prosun. “Cyberbullying detection: an ensemble based machine learning approach,” in *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 2021, pp. 710-715.

[20] M.A. Al-Ajlan and Y. Mourad. “Deep learning algorithm for cyberbullying detection.” *International Journal of Advanced Computer Science and Applications*, vol. 9, pp. 199-205, Sep. 2018.

[21] J. Wang, K. Fu, and C. Lu. “IEEE Big Data 2020 Cyberbullying Dataset.” Internet: https://drive.google.com/drive/folders/1oB2fan6GVGG83Eog66Ad4wK2ZoOjwu3F?usp=sharing, Aug. 23, 2020 [Oct. 18, 2022].

[22] M. Duggan. “Online Harassment 2017.” Internet: https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/, Jul. 11, 2017 [Apr. 6, 2023].