Cyberbullying Detection using Machine Learning

***Harsh Kumar, Harshit Pandey, Harsh Kumar Singh***

*Department of CSIT, KIET GROUP OF INSTITUTIONS*

# ABSTRACT: This research paper presents that with the rise of internet and social media, cyber bulling or online harassment is becoming more and more widespread. Many people use social media platform to spread harass others and spread hate online that has resulted in disastrous consequences that why there is a pressing need for its detection and prevention. In this project, we aim to build a system that tackles the Cyber bullying by identifying the text and placing it into appropriate category. The target of this system is to deal with Cyber bullying and also create awareness about this issue between youngsters as well as elders.

**Keywords: Cyber bullying detection, social media, Twitter, Online harassment, Hate speech, NLP (Natural Language Processing), Machine learning.**

# INTRODUCTION

With the evolution of technology, the ability to bully is also evolved. Instead of restricting to school or colleges bullying has now entered into the online world. Today, kids or teenagers are more connected with friends over the internet than in real life interactions and almost every teenager wants to get popular on social media sites or platform which sometimes leads them to fall in unfortunate conditions that not only affect their mental health but also sometimes force them to take wrong steps. Cyberbullying can happen anywhere there is online social interaction for example on social media like meta, Instagram, in video games, etc. Students sometimes use their private information to access different websites and unknowing fall in trap. Thus, there is a severe need to make youngsters as well as parents aware about such things.

Suicide is the act of taking one’s own life. Suicide is the second leading cause of death in individual people 10 to 34 years of age, occurring at a rate of 10.2 per 100,000 individuals, and its incidence in this age group has risen 56 percent between 2006 and 2016. According to 2019 CDC Youth Risk Behavior Surveillance (YRBS) data, 18.8 percent of high school students seriously considered attempting suicide, 15.7 percent of high school students made a suicide plan, 8.9 percent of high school students attempted suicide, and 2.5 percent of high school students were injured in a suicide attempt. Suicide not only affect the descendent but also the family members and has a ripple effect on the community. But the important thing is that suicides is often a preventable cause of mortality.

The primary objective of this research project is to tackle the issue of cyberbullying through the utilization of Natural Language Processing (NLP) and Machine Learning techniques. Our aim is to develop a system capable of detecting instances of cyberbullying in text and appropriately categorizing them. By doing so, our intention is not only to combat cyberbullying but also to foster awareness about this problem among both the younger generation and adults.

This research paper aims to explore the intersection of machine learning and cyberbullying detection, search into the potential of advanced algorithms to sift through the digital noise and identify instances of harmful behavior. By taking advantages of the power of artificial intelligence, we seek to develop a proactive and efficient approach to detect, categorize, and respond to cyberbullying incidents.

# II RELATED WORK

In an effort to model the cyberbullying, Kelly Reynolds and April Kontosthatis, 2011 used machine learning to train the data collected from FromSpring.me which is a social networking site, the data which is used was labeled using Turk an Amazon Web Service. The number of bad words were used in this model as a feature to train it.

In a study by Dinakar, states that the use of individual topic-sensitive classifiers proves more effective in identifying instances of cyberbullying. They experimented it on a large corpus of comments collected from Youtube.com.

Dadvar, analyzed the gender perspective within the cyberbullying detection, employing it to the social networking platform Myspace, a platform featuring interactive, user-submitted communities of friends, personal profiles, blogs, groups, and more. The study focused on the content of user-generated posts while disregarding user profile information. Employing a Support Vector Machine (SVM) model, a dedicated gender text classifier was trained. The dataset consists of about 381.000 posts. The results obtained by the gender-based approach improved the baseline by 39% in precision, 6% in recall, and 15% in F-measure.

At MIT, Dinakar tested various computer systems on a bunch of labeled You Tube comments they had tagged by hand. Their method got it right about 66.7% of the time, and they also made use of a special kind of learning tool called an SVM.

R. R. Dalvi suggests a way to find and stop online abuse on Twitter using advanced computer learning tools. In this study, they gather tweets in real-time using the Twitter API to create datasets. They test two different machine learning methods, Support Vector Machine and Naive Bayes, on these datasets. They use something called the TFIDF vectorizer to pick out important features. The results reveal that the cyberbullying model based on the Support Vector Machine is around 71.25% accurate, which is better than the Naive Bayes method, which is only about 52.75% accurate.

Xu suggested that using various language tools to find signs of bullying in online content. They outlined how a bullying incident is structured and identified possible roles. The authors used Sentiment Analysis to figure out roles and Latent Dirichlet Analysis to find topics. They turned the cyberbullying detection into a kind of yes/no problem, training a linear SVM with a dataset labeled by hand. The outcome showed an 89% accuracy during cross-validation, proving that even simple features and common tools can be effective in spotting cyberbullying signals in text.

J. Yadav introduced a fresh way to spot cyberbullying on social media. They used the BERT model with a single neural network layer on top as a kind of detective. They trained and tested this detective on data from the Formspring forum and Wikipedia. The new model did really well, with a 98% accuracy for Formspring and 96% for Wikipedia, which is better than older models. Interestingly, it worked especially well for Wikipedia because of its big size, so they didn't need to do anything special to the data, while for Formspring, they had to do some extra steps to make it work well.

Altaf Mahmud attempted to tell the difference between statements that state facts and those that insult by examining comments with some language rules. However, they didn't really focus on comments aimed at both the people involved and those who are not. Another study by Rajavi used a fixed list of words and a three-tier classification method, using features based on bags of words. This approach involved using a list of words that is not easy to find or access.

S. M. Kargutkar put forward a system to give a more detailed understanding of cyberbullying. They used Convolutional Neural Network (CNN) and Keras to analyze content since the existing methods back then gave a vague view with less accuracy. This study looked at data from both Twitter and YouTube, and the CNN accuracy turned out to be 87%. Deep learning models, like the one used here, are becoming popular for recognizing instances of online harassment. They can overcome the limitations of traditional models and enhance their effectiveness.

Notice that many of these research studies use a method where a supervisor guides the computer to learn how to spot cyberbullying. They often use pre-trained helpers, like SVM, to do the heavy lifting. They also use CNN, Naïve bayes and BERT Model. People manually label the data, but it's a slow process and usually covers only a small part of the available information. These studies often use Natural Language Processing (NLP), which is like teaching the computer to understand and analyse text, to help with cyberbullying detection. Most of the time, these language tasks are done at the beginning, before the computer really gets to work.

Many studies have looked into categorizing text or detecting cyberbullying in the English language. They focused on things like YouTube comments, where each comment was labeled by hand, and different ways of sorting them were tried out. Out of all these methods, SVM stood out for being really good at sorting out different types of text. Some recent research even found that the NB Classifier works well for sorting Indian text. Interestingly, when it came to Urdu language, SVM did a better job than the NB method. So, this research aims to dig into different machine learning methods to find out what works best.Top of Form

**III PROPOSED METHODOLOGY**

As we know, Twitter dataset may easier to extracted compared to other mediums such as Facebook, Instagram, and YouTube. Even though the cyber bullying occurred most in Facebook but only data from public profiles could be extracted easily such as Twitter that the data is publicly available. The main job was to gather information from social media using the platform's tools. After that, we had to clean up and get the data ready. Since the data we got had different languages and lots of emojis, we had to clean it up so it would be more accurate. We tried out different machine learning methods to see which one worked the best.

Frequent use of SVM by researchers shows that SVM is popular among other classifiers in supervised learning approaches. SVM works well for sorting through lots of text, especially when it's imbalanced, like when we're trying to find cyberbullying using the content of messages. It handles different situations, like when there's missing data or when we're dealing with different types of features, and it still does a good job compared to other sorting methods.

We have divided all the features into these five categories:

● Sentimental Features

● Sarcastic Features

● Syntactic Features

● Semantic Features

● Social Features

In SVM based researches these features are generally obtained by statistical analysis of documents (tweets or sentences):

***Bad words:***

In many studies, it's clear and makes sense that if a text has certain "bad" words, it's likely to be considered a potential cyberbullying sentence. Following the approach used in other research, we've also compiled a list of insults and swear words, totaling 550 terms. We gathered these terms from various sources available online.

***Bad words density:***

In our study, we also examine how many "bad" words appear in a sentence, and we treat this as a single feature. This feature represents the proportion of bad words in a sentence for each severity level, divided by the total number of words in that sentence.

We also introduce another feature to measure the overall "badness" of a text. This feature is calculated by finding the weighted average of the "bad" words, where each word is given a weight based on its severity level.

***Density of upper-case letters:***

We got the idea for this feature from Dadvar's findings. We're looking at whether there are capital letters in a text message, thinking of it as a sign of possibly "shouting" at someone, which is often how people interpret it in social networks. This feature is calculated by finding the ratio of the number of uppercase letters to the total length (number of characters) of the entire sentence.

***Exclamations and questions marks:***

The preprocessing step is done in the following:

- Tokenization: Here, we break down the text into sentences or the entire paragraph, and then we present the inputted text as individual words in a list.

- Lowering text: This takes the list of words that got the out of tokenization and then lower all the letters Like: “HEY I AM HARSH” is going to be ‘hey i am harsh’.

- Removing Stop Words and Encoding Cleaning: This is an important step in getting the text ready, where we get rid of words like "the" and special characters like '\*', which don't give useful information to the classifiers.

Now, let’s talk about the features we have extracted. We sorted all these features based on what we found in existing studies, and each feature has its own way of pinpointing the text. Picking features that give useful, detailed, and independent information is a really important step. It makes a big difference in how well the computer programs can recognize patterns and classify things correctly.

In the sentimental features, we aim to figure out if a text is positive or negative. Studies reveal that when humans analyze these texts, they tend to agree about 80-85% of the time. We used this agreement as a starting point when teaching our system to score sentiments.

In sarcastic features, we pay attention to context incongruity, where nonverbal behavior clashes with spoken words. For example, a text might have some things in a matching context, and others in a conflicting one. This matters in cyberbullying detection since the true meaning of a sarcastic comment might be missed in sentiment analysis due to this context clash. We also look at practical features like emojis and mentions to spot the sarcastic tone in the source material.

In syntactic features, we not only find insults but also keep track of how many of these bad words or insults are in a single sentence. Then, we measure the density of these bad words in the sentence. We've checked the overall badness of the sentence based on factors like density range. We also pay attention to the use of uppercase letters when someone is making mean statements because it can be like shouting on social media. Similarly, we consider special characters or patterns made by them when figuring out syntactic features.

Semantic Features help us understand the connection between words in a language. They represent the meaning of words. In our case, we've focused on finding groups of three words and pairs of two words used in text references. We pay attention to sentence negation and consider how different pronouns, whether implicit or explicit, might be used to refer to someone while engaging in online harassment.

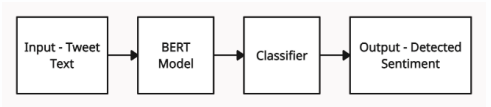
Social features involve looking at the social behavior of both the victim and the bully. Just analyzing the post, itself isn't enough to understand the text's nature. We've looked at patterns in bullies' behaviors and identified a few features. This includes directly tagging the victim when using hate speech. We also try to understand the post's context by looking at previous interactions between the bully and the victim. Additionally, we can profile the author to find out about their past interactions and involvement in similar harmful activities on social media.

We proposed a cyberbullying detection model based on transformers. A recent improvement on the natural language task introduced the BERT. The BERT is a recent paper published by researchers at Google AI Language. BERT stands for Bidirectional Encoder from Transformers. It's a model that learns from both the left and right directions using unlabeled texts, making it grasp the meaning from both contexts. BERT is powerful in Natural Language Processing (NLP) tasks because of its semi-supervised learning approach. We can use this model to build a cutting-edge machine learning model for a specific task by adding a task-specific layer on top of BERT's architecture. BERT is bidirectional, meaning it tries to understand the word's meaning from both the left and right contexts to get a richer understanding during training.

We saw a **bat.**

This **bat** was given to me by my brother.

In the first sentence, if we look at the context around the underlined word "bat" from the left, it's talking about the nocturnal animal. On the other hand, in the second sentence, if we focus on the context from the right, it's referring to the cricket bat. This creates a challenge for machines to predict the exact meaning of the word without considering both contexts. BERT solves this problem because it's a bidirectional model, understanding the meaning by looking at the context from both the left and right sides.



**IV PROBLEM STATEMENT**

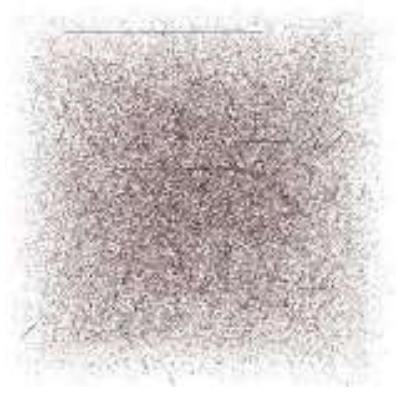
We want to create a software app that can find instances of bullying on social media platform but because the data is not easily available for all other social media platforms, we are creating it for Twitter. Our focus is on detecting bullying for social reasons, ranging from "reducing the number of suicides caused by bullying" to "making microblogging a no bullying zone."

**V RESULT ANALYSIS**

We took a bullying-related tweet from Twitter, applied it to our model, and in the classification report, labels 0 and 1 represent Bullying and Non-Bullying, respectively. There is a confusion matrix based on our testing results. The accuracy of SVM and Naive Bayes, which are 71.25% and 52.70%, respectively, when used on the same dataset from [3]. The accuracy of the BERT model on the same dataset, showing improved accuracy in sentiment analysis on the Twitter dataset. Our proposed model achieved an accuracy of 91.90% when applied to the Twitter dataset for sentimental analysis, which is notably higher compared to traditional machine learning models on similar datasets.

To assess our results, we also looked into using Amazon's Mechanical Turk (Crowd sourcing) to categorize unlabeled data and confirm newly labeled data. Combining Crowd sourcing with machine learning algorithms, we built an infrastructure capable of detecting instances of bullying in Twitter's public timeline.

Even with a lot of data collected, Twitter's system would pinpoint and remove potential bullies (spammers) faster than we could thoroughly examine their social connections. This led to missing chunks of information, creating a disorderly bullying graph that made further analysis impossible. Interestingly, this observation could be seen as a sign of Twitter's success in effectively using powerful spam filter technology.



**Bullying Graph**

**VII CONCLUSION AND FUTURE SCOPE**

This study aims to tackle cyberbullying on the Twitter platform using Machine Learning. We experimented with both supervised and unsupervised machine learning techniques. We found that selecting the correct keywords is crucial for improving results in sentiment analysis. The outcomes show that our model performs reasonably well and could be effectively used to develop practical monitoring applications, helping to address the serious social issue of cyberbullying.

The experiment shows that the SVM-based method gives the highest accuracy, and its performance gets even better when user-specific data is included. Because of the high-dimensional input space, a few irrelevant features, and the linearly separable nature of the text dataset, SVM outperforms other classification algorithms in text classification. In the future, we can study the significance of individual features to further improve the method.

But on contrast, we also suggested a semi-supervised method for spotting cyberbullying using five features to define a cyberbullying post or message with the BERT model. Focusing on just one feature, sentimental features, the BERT model achieved 91.90% accuracy when trained over dual cycles, outperforming traditional machine learning models. The BERT model could give even more accurate results with a larger dataset. To improve cyberbullying detection, considering all proposed features in this research could be beneficial. Creating an application based on these features could detect bullying traces and aid in reporting such posts. Additionally, combining other models with the BERT model in the future could create a cutting-edge model for specific Natural Language Processing tasks in detecting cyberbullying.

**VIII REFERENCES**

[1] J. Xu, K. Jun, X. Zhu, and A. Bellmore, “Learning from Bulling Traces in Social Media,"Proc.2012 Conf.North Am. Chapter Assoc. Comput. Linguist. Hun. Lang. Technol, pp. 656-666,2012

[2] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. “Modeling the detection of Textual Cyberbullying.” The Social Mobile Web 11.02(2011):11-17

[3] K. Dinkar, R. Reichart and H. Lieberman, “Modeling the Detection of Textual Cyberbullying,” MIT. International

[4] M. Dadvar and F.de Jong. 2012.” Cyberbullying detection: step toward a safer internet yard”. In Proceedings of the 21st International Conference on World Wide Web (WWW ’12 Companion). ACM, New York, NY, USA, 121-126

[5] R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893. (2020)

[6] J. Yadav, D. Kumar and D. Chauhan, Cyberbullying Detection using Pre-Trained BERT Model, ICESC, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700. (2020)

[7] S. M. Kargutkar and V. Chitre, A Study of Cyberbullying Detection Using Machine Learning Techniques, ICCMC, pp. 734-739, doi: 10.1109/ICCMC48092.2020.ICCMC-00013 7. (2020)