# Cyber Bullying Detection on Social Media using Machine Learning

*Aditya* Desai[1], *Shashank* Kalaskar[2], *Omkar* Kumbhar[3], and *Rashmi* Dhumal[4]

[1,2,3]Student, Department of Computer Engineering, Ramrao Adik Institute of Technology, Nerul
[4]Assistant Professor, Department of Computer Engineering, Ramrao Adik Institute of Technology, Nerul

**Abstract.** Usage of internet and social media backgrounds tends in the use of sending, receiving and posting of negative, harmful, false or mean content about another individual which thus means Cyberbullying. Bullying over social media also works the same as threatening, calumny, and chastising the individual. Cyberbullying has led to a severe increase in mental health problems, especially among the young generation. It has resulted in lower self-esteem, increased suicidal ideation. Unless some measure against cyberbullying is taken, self-esteem and mental health issues will affect an entire generation of young adults. Many of the traditional machine learning models have been implemented in the past for the automatic detection of cyberbullying on social media. But these models have not considered all the necessary features that can be used to identify or classify a statement or post as bullying. In this paper, we proposed a model based on various features that should be considered while detecting cyberbullying and implement a few features with the help of a bidirectional deep learning model called BERT.

**Keywords**: Cyberbullying, Social Media, BERT, NLP, Semi-supervised learning , Twitter API.

## 1 Introduction

Millions of young people spend their time on social networking, and the sharing of information is online. Social networks have the ability to communicate and to share information with anyone, at any time, and in the number of people at the same time. There are over 3 billion social media users around the world. According to the National Crime Security Council (NCPC), cyberbullying is available online where mobile phones, video game apps, or any other way to send or send text, photos, or videos deliberately injure or embarrass another person. Cyberbullying can happen at any time all day, week and you can reach anyone anywhere via the internet. Text, photos, or videos of cyberbullying may be posted in an undisclosed manner. It can be difficult, and sometimes impossible, to track down the source of this post. It was also impossible to get rid of these messages later. Several social media platforms such as Twitter, Instagram, Facebook, YouTube, Snapchat, Skype, and Wikipedia are the most common bullying sites on the internet. Some of the social networking sites, such as Facebook, and the provision of guidance on the prevention of bullying. It has a special section that explains how to report cyber-bullying and to prevent any blocking of the user. On Instagram, when someone shares photos and videos made by the user to be uncomfortable, so the user can monitor or block them. Users can also report a violation of our Community and make Recommendations to the app.

As the social lifestyle exceeds the physical barrier of human interaction and contains unregulated contact with strangers, it is necessary to analyze and study the context of cyberbullying. Cyberbullying makes the victim feel that he is being attacked everywhere as the internet is just a click away. It can have mental, physical, and emotional effects on the victim. Cyberbullying mainly takes place in the form of text or images on social media. If bullying text can be distinguished from non-bullying text, then a system can act accordingly. An efficient cyberbullying detection system can be useful for social media websites and other messaging applications to counter such attacks and reduce the number of cyberbullying cases. The objective of the cyberbullying detection system is to identify the cyberbullying text and also take its meaning into consideration. One first analyzes the various aspects of a particular text and then applies the previous information or visuals to find the context of the text. There is a need to create a personalized system that can access such a text effectively and efficiently.

## 2 Literature Survey

M. Di Capua, et al. [1] proposes an unsupervised approach to develop a cyberbullying model based on an amalgam of features, based on traditional textual features as well as some “social features”. The features were separated into 4 categories as Syntactic features, Semantic features, Sentiment features, and Social features. The author used a Growing Hierarchical Self Organizing Map (GHSOM) network, with a grid of 50 x 50 neurons and 20 features as the input layer. M. Di Capua, et al have applied the clustering algorithm k-means to classify the input dataset along with GHSOM on the Formspring dataset. The results of this hybrid unsupervised methodology surpassed the previous

* Corresponding author: adityadesai1703@gmail.com

results. The author then tested the youtube dataset with 3 different Machine Learning Models: a Naive Bayes Classifier, Decision Tree Classifier(C4.5), and a Support Vector Machine(SVM) with a Linear Kernel. It was observed that clustering results for the hate posts turned out to have a lower precision in the youtube dataset when compared to the FormSpring tests, as textual analysis and syntactical features perform differently on both sides. When this hybrid approach was applied to the Twitter dataset, it resulted in a weak recall and F1 Score. The model proposed by the authors can be improved and used in building constructive applications to mitigate cyberbullying issues.

J. Yadav, et al.[2] proposes a new approach to cyberbullying detection in social media platforms by using the BERT model with a single linear neural network layer on top as a classifier. The model is trained and evaluated on the Formspring forum and Wikipedia dataset. The proposed model gave a performance accuracy of 98% for the Form spring dataset and of 96% for the Wikipedia dataset which is relatively high from the previously used models. The proposed model gave better results for the Wikipedia dataset due to its large size g without the need for oversampling whereas the Form spring dataset needed oversampling.

R. R. Dalvi, et al.[3] suggests a method to detect and prevent Internet exploitation on Twitter using Supervised classification Machine Learning algorithms. In this research, the live Twitter API is used to collect tweets and form datasets. The proposed model tests both Support Vector Machine and Naive Bayes on the collected datasets. To extract the feature, they have used the TFIDF vectorizer. The results show that the accuracy of the cyberbullying model based on the Support Vector Machine is almost 71.25% that is better than the Naive Bayes which was almost 52.75%.

Trana R.E., et al. [4] goal was to design a machine learning model to minimize special events involving text extracted from image memes. The author has compiled a database containing approximately 19,000 text views published on YouTube. This study discusses the effectiveness of the three machine learning machines, the Uninformed Bayes, the Support Vector Machine, and the convolutional neural network used on the YouTube database, and compares the results with the existing Form databases. The authors further investigated algorithms for Internet cyberbullying in sub-categories within the YouTube database. Naive Bayes surpassed SVM and CNN in the following four categories: race, ethnicity, politics, and generalism. SVM has passed well with the inexperienced Naïve Bayes and CNN in the same gender group, and all three algorithms have shown equal performance with central body group accuracy. The results of this study provided data that can be used to distinguish between incidents of abuse and non-violence. Future work could focus on the creation of a two-part segregation scheme used to test the text extracted from images to see if the YouTube database provides a better context for aggression-related clusters.

N. Tsapatsoulis, et al. [5] a detailed review of cyberbullying on Twitter is presented. The importance of identifying different abusers on Twitter is given. In the paper, various practical steps required for the development of an effective and efficient application for cyberbullying detection are described thoroughly. The trends involved in the categorization and labeling of data platforms, machine learning models and feature types, and case studies that made use of such tools are explained. This paper will serve as an initial step for the project in Cyberbullying Detection using Machine learning.

G. A. León-Paredes et al.[6] have explained the development of a cyberbullying detection model using Natural Language Processing (NLP) and Machine Learning (ML). A Spanish cyberbullying Prevention System (SPC) was developed by applying machine learning techniques Naïve Bayes, Support Vector Machine, and Logistic Regression. The dataset used for this research was extracted from Twitter. The maximum accuracy of 93% was achieved with the help of three techniques used. The cases of cyberbullying detected with the help of this system presented an accuracy of 80% to 91% on average. Stemming and lemmatization techniques in NLP can be implemented to further increase the accuracy of the system. Such a model can also be implemented for detection in English and local languages if possible.

P. K. Roy, et al. [7] detail about creating a request for the discovery of hate speech on Twitter with the help of a deep convolutional neural network. Machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-nearby Neighbors (KNN) has been used to identify tweets related to hate speech on Twitter and features have been removed using the tf-idf process. The best ML model was SVM but it managed to predict 53% hate speech tweets in a 3: 1 dataset to test the train. The reason behind the low prediction scale was the unequal data. The model is based on the prediction of hate speech tweets. Advanced forms of learning based on the Convolutional Neural Network (CNN), Long-Term Memory (LSTM), and their Contextual LSTM (CLSTM) combinations have the same effects as a separate distributed database. 10-fold cross-validation was used along with the proposed DCNN model and obtained a very good recall rate. It was 0.88 for hate speech and 0.99 for non-hate speech. Test results confirmed that the k-fold cross-validation process is a better decision with unequal data. In the future, the current database can be expanded to achieve better accuracy.

S. M. Kargutkar, et al. [8] had proposed a system to give a double characterization for cyberbullying. The system uses Convolutional Neural Network (CNN) and Keras for content examination as the relevant strategies at that time provided a guideless view with less precision. This research involved data from Twitter and YouTube. CNN accuracy was 87%. In-depth learning-based models have found their way to identify digital harassment episodes, they can overcome the imprisonment of traditional models, and improve adoption.

Jamil H. et al. [9] have described the implementation of a new social network model and its query language called GreenShip. They showed that with the support devices, GreenShip users can be more effective in the fight against online bullying, and loss of traditional, online, and social networks. The reputation of a management model that has been introduced to restrict access to harmful information, which focuses on the denial of the criminal code, the means for the dissemination of information to the users associated with the target. GreenShip has a reputation as a model that provides safe, “green” friends, due to the recognition of the different types of friendships on Facebook. The damage is as a result of bad friends and that was very limited, and the more complex, that there are many forms of friendship, and the communication lines are put aside for the sake of the benefits of privacy and control.

Rasel, Risul Islam, et al. [10] focuses on the removal of the comments made on social networks, and the analysis of the question as to whether these observations provide an offensive meaning. The reactions can be divided into three categories: offensive, hate speech, and neither of the two. The proposed model classifies the notes on the species), with an accuracy of more than 93%. Latent Semantic Analysis (LSA) has been used as a feature selection method to reduce the size of the input data. In addition to standard feature extraction methods such as tokenization, N-gram, TF-IDF was applied to detect the important notes. We made three different machine learning models, Random Forest, Logistic Regression, and Support Vector Machines (SVMs) to perform the calculation, analysis, forecasting, and a teasing comment.

## 3 Proposed Methodology

In this paper, a method to detect cyberbullying on social media is proposed that is not just based on the sentimental analysis but also considers the syntactic, semantic, and sarcastic nature of the sentence before classifying it as hate speech. To achieve our goal we start with the traditional sentiment analysis where we perform contextual mining of text to identify and extract the subjective information in the source material to understand the opinion, emotion, or attitude towards the topic. Later we introduce a group of “social” features that can highly affect and guide the cyberbullying detection process. We have divided all the features we have extracted into five categories:

- Sentimental Features
- Sarcastic Features
- Syntactic Features
- Semantic Features
- Social Features

All these features have been categorized based on the literature survey of the existing systems and each feature uniquely identifies the text. Choosing informative, descriptive, and independent features is a crucial stage for the effectiveness of the algorithms in pattern recognition and classification problems.

In sentimental features, we try to evaluate the sentiment( positive, negative ) of a given text document. The research shows that human analysts tend to agree around 80-85% of the time and that is the baseline we have tried to consider while training our sentiment scoring system.

In sarcastic features, we try to consider the context incongruity. Incongruity occurs when a nonverbal behavior contradicts a person's word. A text may contain half of the objects in a congruent context which can be considered as expected context, whereas for the other half, objects were embedded in incongruent contexts. This can be a major factor in cyberbullying detection because the hidden nature of the sarcastic comment won’t be detected in sentiment analysis because of the context incongruity. We also consider pragmatic features like emojis, mentions, etc. while detecting the sarcastic nature of the source material.

While considering the syntactic features we have identified in the lists of insults, we also monitor and take into consideration the number of such bad words or insults present in a single sentence and accordingly map a density to it. We have also validated the badness of the entire sentence based on certain parameters like density range. The emphasis of uppercase characters while making hate statements is also taken into consideration while generating syntactic features because it can be referred to as an act of shouting or attacking over social media platforms. Similarly, the use of special characters or patterns formed by them is also brought into consideration while deriving syntactic features.

Semantic Features can be used to determine the lexical relation which exists between two words in a language. The meaning of the word can be represented by Semantic features. Here we have tried to identify the trigram and the bigrams that occur while referencing something in the text format. Here usually the negation of the sentence is considered along with the mapping of different pronouns that can be implicitly or explicitly used to refer to another individual while harassing someone over social media.

Social features refer to the social behavior of the victim or the bully itself. The post itself won't be sufficient to detect the nature of the text. We have considered patterns in the behaviors of the bullies and identified a few features. We have considered the direct tagging of the victim while using hate speech. We also try to gain information regarding the context of the post based on the previous interactions between the bully and the victim. Profiling of the author can be done to discover its past interactions and involvement in similar malicious activities over social media platforms.

We proposed a cyberbullying detection model based on transformers. Similar to RNN, transformers can also be used to solve a wide variety of NLP(Natural Language Processing) problems like translation and text summarization as they can take sequential data as input. 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 is a bidirectional model that is pre-trained on unlabeled texts from both left and right directions to understand the meaning of both contexts. BERT is a powerful model for NLP tasks because of the use of semi-supervised learning. We can use this model to create a state-of-the-art machine learning model for a specific task by applying an additional task-specific layer on top of the BERT architecture. BERT is a Bidirectional model which means it aims to understand the meaning of the word from both the left and the right context to derive a better meaning during the training phase.

We saw a **bat**.
This **bat** was given to me by my father.

Here in the first sentence if we focus on the context of the underlined word "bat" from the left till the word, it refers to the nocturnal animal. Whereas if we focus on the context of the word "bat" in the second sentence from right till the word it refers to the bat from the game of cricket. Thus a machine can face problems predicting the actual meaning of the word without considering both the context. This problem is solved by BERT as it is a bidirectional model.

BERT model requires its input in a preprocessed form as per the rules made by its developer. These rules have helped the model to achieve better performances. All inputs are embedded as a combination of the other 3 embeddings and given as an input to the model:

- Position embedding: BERT reads and uses existing embedding to express word order in a sentence.
- Segment Embedding: BERT can also take more than one sentence as input functions. It uses this embedding to understand the difference between two different sentences.
- Token Embeddings: This is the embedded text token from Word Piece token vocabulary.

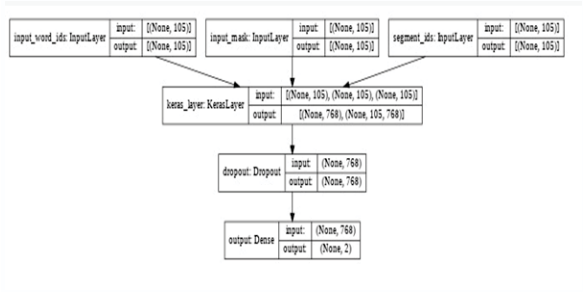


**Fig.1.** BERT model based on sentiment analysis

The Fig.1 depicts the BERT model we developed for sentimental analysis. The three separate embeddings are summarized together to give input to the internal layers of the model. The Fig.2 demonstrates the flow of the sentimental analysis process. To begin with, the final CLS token will output a matrix of hidden size. Furthermore it will be passed to a classifier layer. In conclusion the classifier layer will determine the sentiment of the input text.

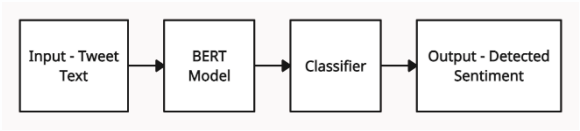


**Fig.2.** BERT model flow chart based on sentiment analysis

## 4 Result Analysis

The Fig.3 represents the input processing and prediction result that we performed during our testing. We used a tweet from Twitter with the trace of bullying and applied it to our model. Fig 4 shows the classification report based on our testing data. Here labels 0 and 1 represent Bullying and Non-Bullying respectively. The Fig.5 represents the confusion matrix based on the result of our testing data. Table 1 represents the accuracy of the SVM and Naive Bayes that is 71.25% and 52.70% respectively, when applied on the same dataset from [3]. Table 2 represents the accuracy of the BERT model on the same dataset. The result shows better accuracy when using the BERT model for sentiment analysis on the Twitter dataset. Our proposed model gave a better accuracy of 91.90% when applied to the Twitter dataset for the sentimental analysis which can be considered as a greater result when compared to the traditional machine learning models used on similar datasets.


**Fig.3.** Implementation Result

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.65 | 0.57 | 0.61 | 84 |
| 1 | 0.74 | 0.80 | 0.77 | 129 |
| accuracy | | | 0.71 | 213 |
| macro avg | 0.69 | 0.68 | 0.69 | 213 |
| weighted avg | 0.70 | 0.71 | 0.71 | 213 |

**Fig.4.** Classification report


**Fig. 5.** Confusion Matrix

**Table 1.** Accuracy of SVM and Naive Bayes from [3]

| Classifier | Accuracy in percentage |
|---|---|
| Naïve Bayes Classifier | 52.70 |
| Support Vector Machine | 71.25 |

**Table 2.** Accuracy of BERT Model

| Classifier | Accuracy in percentage |
|---|---|
| Pre-Trained BERT (testing) | 70.89 |
| Pre-Trained BERT (training) | 91.90 |

## 5 Conclusion

We proposed a semi-supervised approach in detecting cyberbullying based on the five features that can be used to define a cyberbullying post or message using the BERT model. While considering just one of the features which was sentimental features the BERT model achieved 91.90% accuracy when trained over dual cycles which outperformed the traditional machine learning models. The BERT model can achieve more accurate results if provided with a large dataset. We can try to achieve even better results in the cyberbullying detection process if we consider all the features that we have proposed in this research paper. Based on all the features an application can be created to detect the bullying traces and thus help in detecting and reporting such posts. A combination of other models on top of the BERT model can also be used in the future to create a state-of-the-art model for the specific NLP tasks in detecting cyberbullying.

## References

1. M. Di Capua, E. Di Nardo and A. Petrosino, *Unsupervised cyberbullying detection in social networks*, *ICPR*, pp. 432-437, doi: 10.1109/ICPR.2016.7899672. (2016)
2. 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)
3. 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)
4. Trana R.E., Gomez C.E., Adler R.F. (2021) *Fighting Cyberbullying: An Analysis of Algorithms Used to Detect Harassing Text Found on YouTube*. In: Ahram T. (eds) Advances in Artificial Intelligence, Software and Systems Engineering. AHFE 2020. Advances in Intelligent Systems and Computing, vol **1213**. Springer, Cham. https://doi.org/10.1007/978-3-030-51328-3_2. (2020)
5. N. Tsapatsoulis and V. Anastasopoulou, *Cyberbullies in Twitter: A focused review*, SMAP, pp. 1-6, doi: 10.1109/SMAP.2019.8864918. (2019)
6. G. A. León-Paredes et al., *Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language*, CHILECON pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684. (2019)
7. P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, *A Framework for Hate Speech Detection Using Deep Convolutional Neural Network*, in IEEE Access, vol. **8**, pp. 204951-204962,, doi: 10.1109/ACCESS.2020.3037073. (2020)
8. 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-000137. (2020)
9. Jamil, H. and R. Breckenridge. *Greenship: a social networking system for combating cyber-bullying and defending personal reputation*., ACM : n. pag. (2018)
10. Rasel, Risul Islam & Sultana, Nasrin & Akhter, Sharna & Meesad, Phayung, *Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach*. 37-41. 10.1145/3278293.3278303. (2018)