Detection of Cyberbullying on Social Media

Neha Hudda
Dept. of Information Technology
Meerut Institute of Engineering & Technology
Meerut, Uttar Pradesh, India

Akanksha Mishra
Dept. of Information Technology
Meerut Institute of Engineering & Technology
Meerut, Uttar Pradesh, India

Sunil Kumar, PhD
Dept. of Information Technology
Meerut Institute of Engineering & Technology
Meerut, Uttar Pradesh, India

ABSTRACT
The evolution of the internet gave us capabilities to be interconnected by each other over the globe, but every entity that has evolved in mankind also comes with some demerits. The evolution of the web from just being a simple search index to blogging and social networking, it gives everyone a freedom of speech which makes a chaotic situation about polling and changing perceptions of the people it may cause a social dilemma where an individual has unconsciously maintained 2 personalities. However, the evolution in computer analytical field gave us the privilege of using various available tools and techniques for analyzing the social media and create a policy on the freedom of speech, Our paper focuses on the solution for regulating the policy and guidelines of the social media platform with the help of machine learning techniques and natural language processing. Nowadays there are more and more people who are having a perception of harassing others on social networks, these people may have a different behavior towards others in real life. We have developed a solution to filter out the content on social media, first filtering out the slangs and harassing comments, then it can be used on the blog posts, or even on the images using computer vision techniques. It is necessary to use the language pattern and grammar to maintain a high order overview by which toxicity can be judged. The solution in this particular situation are approached best by Neural Networks, we are using a better Neural Network which is more effective and replace the sequential nature of the neural network, i.e. BERT and it makes use of transformers which is just an attention-based mechanism that can learn contextual relations between words in a text, to identify the text we will be using a Convolutional Neural Network which can help in NLP.

General Terms
Cyber security, detection of slangs, enhanced monitoring on cyber bullying

Keywords
Neural Network, Convolution Neural Network, NLP, BERT

1. INTRODUCTION
The evolution of the social media platform results in the increase of users on the platforms also these users are not well authenticated, means that i can make more than one account with the same details, or the social entity of others can easily be spoofed, as there are no validation check for the person entering in the social world. It is also not efficient to handle the validation of millions of people who are making their first step in the social media world, due to this spoofing and other society lawbreakers social media platforms have become more aggressive in nature, platforms where anyone can post or react on the content. This leads to the major problem of cyber harassment and cyberbullying among common people and social deframe among influencers and other activist, The factor by which cyber bullying is stepping up a step is that it impacts more on social media than an in person bully, this is because everyone on the platform can see and post another comment on the same inappropriate content.

Social media gives us a privilege to maintain a good personality among the peoples, but it can be destroyed by just a simple de framed statement, the statement is the main entity we are focusing on. These statements are simply just like a bully, in this research paper we will be trying to propose a model which can effectively distinguish the toxic statement on the platform. However the main problem for us is to identify which comment or statement is worth noticing and which is not, the solution for this has been derived from the well researched field of computer vision in which computer itself is capable of detecting certain patterns which are then analyzed for detecting a certain parameter in any dataset, CNN is most commonly used on the images but in this situation we are using it on the text or sentence classification we will be using kimCNN for this purpose. After this we need to measure the amount of toxicity in the comment for this to check we have set certain marks that are categorized in: severe_toxic, insult, toxic, threat, identity_hate, obscene. Out of these our model will be calculating the probability for each of the parameters which make it more flexible to identify on which set the statement is, furthermore these sets can be used to completely eliminate the comment or give this to partially change the comment or statements. The advanced BERT Recurrent Neural Network is used to map the words with the vector features which can be used to calculate the context, for which we need to feed the model with comments. After this we will be training our model using KimCNN by the help of PyTorch library which results in the model that is made up so that it can calculate the toxicity of the sentence.

Our aim is to leverage the machine learning algorithms and develop a system that will accurately identify the comments and can generate a toxicity score describing the nature of that particular statement and compare its results with other models using several Scoring techniques.

2. RELATED WORK/ LITERATURE REVIEW
There are many approaches that are being considered to solve this problem however many of them are used to be pretty primitive while others having high accuracy in prediction of explicit text as shown by Nandhini et al. [1] by using popular Naive Bayes approach of machine learning the data set that is being used to fed the model is from MySpace.com, this data set is also used in further models and the algorithm is also modified for that particular dataset, other operations related to detection of speech which have achieved the accuracy in the range of 80

to 90 percent, by this modification now the naive bayes algorithm is modified further by Romsaiyud et al.[2] in which they are able to enhance this algorithm by extracting the words and examining the loaded patterns which further achieve a higher accuracy of 90 to 95 percent. In this approach various data sets are being used from Myspace, Kongregate and Slashdot which makes the model more predictable and accurate.

However these approaches have a major problem which is not taken into consideration that the process doesn't work in parallel. Other approaches use stream buffers of data coming from a platform and then before feeding it to the model they have manually classified all the dataset then feeded to the classification model, also this time it is using the sentiment analysis technique for determination of the results[3].

Moreover support vector Machine (SVM) is also used to yield the result giving us a massive improvement of 95 to 98 percent accuracy as shown in[4] the data-set in this problem is used from kaggle but the actual numbers of the data set is not mentioned, so the result may be dangling on particular test inputs.

Another approach that is considered is done by Dinakar et al. [5] which is aimed specially to detect explicit language which are in the context of sexuality, race, culture, intelligence and gender, in this approach they have processed the dataset from youtube comment section and after applying both SVM and Naive Bayes classifiers the results yielded accuracy of about 66 percent.

Deep learning neural networks were also tested to solve the exact same problem, as done by Zhang et al[6] they have used the Convolutional Neural Network and advanced it to the Pronunciation based Neural Network which itself generate the result by previous patterns and knowledge until it find the right result, the data set that they are using consists at most of 13 thousand description that are being taken from various sources.

3. PROPOSED MODEL

A Prerequisites : our proposed approach is completely based on deep neural networking and Machine learning algorithm technique. In here we have used some mathematical concepts like gradient descent, probability and using keras model layer for providing desirable and most precise output.

Our main goal is to develop an automatic cyberbullying detection system that can be used to detect the activities that are harmful for users on the social media, it can be a threat or any misleading comment, after detecting we will be classifying the same activities and try to predict on which factors it will be affecting the user.

This technique can be applied to the large amount of texts that are streaming from the live chat.

The text stream is then classified under clusters for analysis of the certain parameter like abusive text using Naive Bayes classification algorithm which is used to make a predictive model from our classifier algorithm

First we need to clean the data from the dataset, that is done by pre-processing the data and making it unique by removing the duplicates and special characters from the dataset.

Then we have to classify the messages and predict the text from bullying messages to prevent cyberbullying.

the tasks that are needed to be performed while developing the model are the main stages in which we need to make changes and manipulate the data by preprocessing it:

Making Token : In this stage we need to take care of the input phrases or the paragraphs that are used as an input to the model, These input are used to separate the words in a list

LowerCase Text : after making the tokens of all the input phrases or paragraphs the words needs to come under one Casing that is for convenience lower case

Cleaning Text : This is the important stage where we need to clean the text from escape sequences like new line character, end of file character and many more.

Text Correction : The input text needs to be corrected as the probability is less to get text which is having 100 percent correct semantic and grammar, we need to provide some feature that will auto correct to an extent[7].

The dataset that is going to be taken in consideration is the kaggle dataset which is collected and mentioned by renowned author Kelly Renolds, this particular dataset consist of in general 12 thousand conversations messages which comes from Formspring a general purpose forum platform. This dataset consists of comments that are being treated as questions and their cyber threat status as answers, it basically is a mapping between the comments and a boolean value, which indicates whether this comment is a cyber threat or not.

Table. Statistics of The Dataset

Total no. of Conversation	1624
Cyber threat conversations	812
Normal conversations	812
Distinct words count	5714
Tokens appeared	48928
Maximum size of Text	782 character
Minimum size of Text	64 character

The classification elimination phase is the following stage in this suggested model. Here, the textual information is modified.

To fit a format that machine learning algorithms can use. First, we utilized TFIDF to remove the input data's features and add them to the list's characteristics. The purpose of the key TFIDF is to alter the word weights inside a sentence or piece of text. Using the feeling inspect technique, we remove the sentences dualities and add them as an attribute to the features list that contains the TFIDF features. Due to the polarity of the statements, they can be categorized as either positive or negative. For that reason, we need to distinguish between using Text Blob as a former training tool and a present movie review. In addition to sentiment disparity extortion and feature expression using TFIDF, the suggested approach leverages grammar to take various word combinations into account.The categorization stage in the suggested approach comes next, where the takeaway is added to a grouping algorithm to train, test, and employ the ranking in the phase.

We employed the SVM (Support Vector Machine) and neural network classifiers. Three layers make up the neural network: input, secret, and output layers.

There are 128 nodes in the input sheet. It has 64 neutrons in the buried layer. A Boolean output is the layer in the output.

4. RESULT

After appealing the data file, we extract the features using the same process as in Section III. The data set was then divided into (0.8,0.2) allowances for train and testing. The classifiers are evaluated using precision, recall, and precision as well as f-score[9]. SVM and Neural Networks (NN) were included since they are some of the best exhibit characteristics in the literature. Various tests were conducted using various n-gram language models. Particularly, we consider two gram, three gram, and four gram through classifiers progress of the prototype build. The accuracy of both SVM and NN are summarized in Table I.The SVM had the maximum percentage of accuracy while using four Gram, with an accuracy of 90.3%, while the NN exhibited the highest correctness when using three Gram, with an accuracy of 92.8%.The results of both algorithms accuracy tests show that the mean of all n-gram models of NN is 91.76% and The average across all SVM n-gram models is 89.87%.

This analysis clearly shows the minute difference between various techniques that are used to predict the result, all this can be compared with further information as shown in table II and table III.

Table I. The Accuracy of SVM and NN in Different Language Model

Classifier	II-gram	III-gram	IV-gram	Average
SVM	88.67%	89.9%	90.3%	89.87%
Neural Network	90.9%	92.8%	91.6%	91.76%

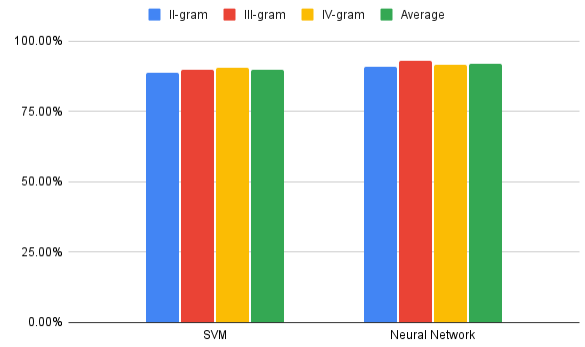

Fig 2. graphical comparison of accuracy

Besides accuracy, Tables II and III display the evaluation of the two features for each language model in terms of accuracy and memory, respectively.[9] Table five which shows the fscore of both traits across specific linguistic frameworks, illustrates the adjustment between recall and correctness. The f-score for the SVM and NN pair is condensed in Table IV. With an f-score of 91.27%, SVM algorithm has the lowest error f-compute using four Gram, while the NN had the highest f-calculate using 2-Gram with a score of 94%[10] It has been determined the median f type score for all NN n type gram framework is 92.8%, whereas the average f-score for all SVM n-gram models is 89.6.For the SVM and neural network grouping in Fig. 3, minimize the f type score. The average outcomes and average f-score both demonstrate how much more accurate NN executions are than SVM.

Table II. : Recall comparison

Classifier	II-gram	III-gram	IV-gram	Average
SVM	88.67%	90.3%	90.8%	90.1%
NN	91.6%	91.5%	92%	91.76%

Table III. Precision comparison

Classifier	II-gram	III-gram	IV-gram	Average
SVM	88.67%	89.34%	90%	89.6%
NN	94%	93.23%	92.6%	92.8%

Table IV. F-score comparison

Classifier	II-gram	III-gram	IV-gram	Average
SVM	88.67%	88.98%	91.27%	90.87%
NN	93.2%	90.9%	92.8%	92.9%

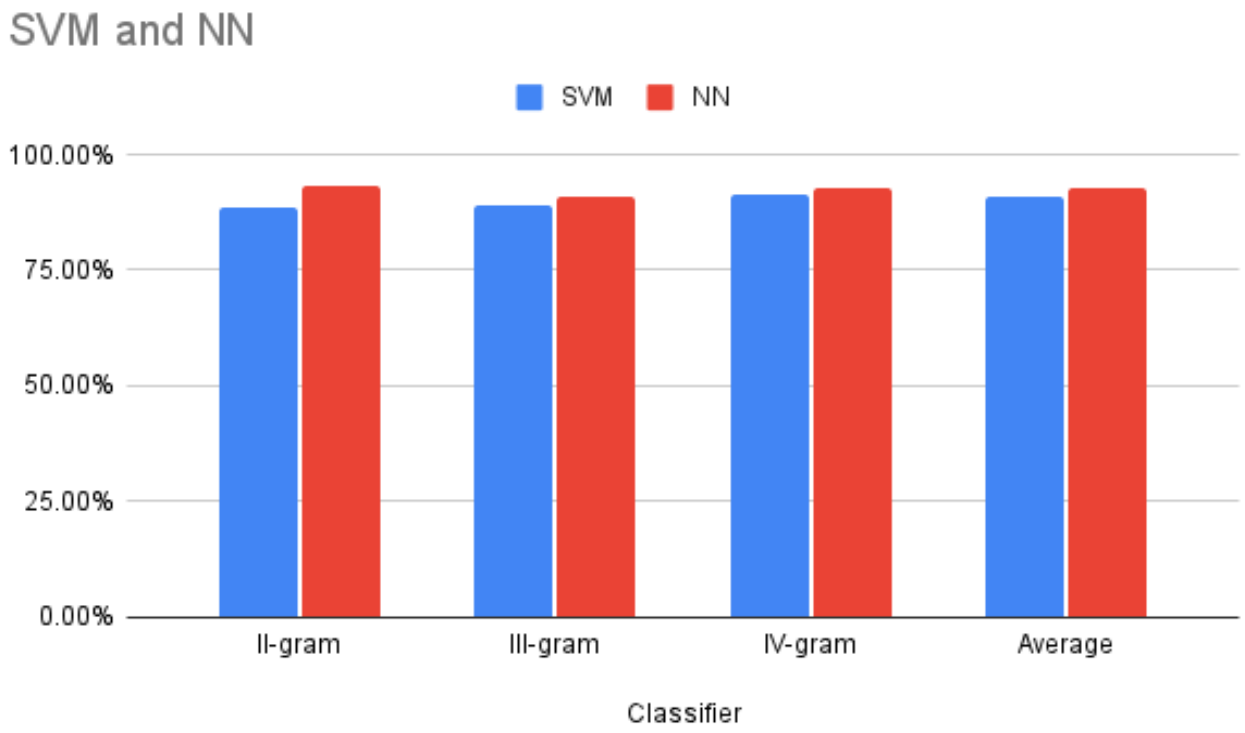

Fig 3. F score comparison

5. CONCLUSION

Using machine learning and artificial intelligence approaches, we developed a perspective in this research study for addressing cyberbullying. We checked our prototype on these double features support vector machine and also with Neural Network model furthermore for characteristics removal, we employed and leveraged the benefits of TF IDF also with the help of algorithms especially designed for graphical analysis.The traits were calculated on the various pre-configured n gram models designed for languages. This let us successfully accomplish 92.8 percent perfection with the help of Neural Network by using majorly three grams but also with 90.3 percent accuracy using a support vector machine with four dowager by just leveraging both the TF IDF and view study. We maintain our Neural Network by setting up discharge as it precisely obtains an average f-score of 91.9 percent, it is considered to be more superior to the SVM classifier because the SVM achieves an median f-score of 89.8 percent. Furthermore,We looked at how our Neural Network performed in terms of accuracy and f-score compared to another relevant study that employed a comparable data file. By reaching this validity, our study will undoubtedly enhance the ability to detect cyberbullying and assist users in doing so securely.

The amount of the training data, however, restricts the ability to identify cyberbullying patterns.

Larger cyberbullying data are therefore required to improve performance. Because deep learning approaches have been shown to outperform machine learning, they will therefore be suitable for larger data sets.

6. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

7. REFERENCES

[1] B Nandhini and JI Sheeba. Cyberbullying detection and classification using information retrieval algorithms. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), page 20. ACM, 2015.

[2] 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)

[3] Shane Murnion, William J Buchanan, Adrian Smales, and Gordon Russell. Machine learning and semantic analysis of in-game chat for cyberbullying. Computers & Security, 76:197–213, 2018.

[4] Sani Muhamad Isa, Livia Ashianti, et al. Cyberbullying classification using text mining. In Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on, pages 241–246. IEEE, 2017.

[5] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Transactions on Interactive Intelligent Systems (TiiS), 2(3):18, 2012.

[6] Xiang Zhang, Jonathan Tong, Nishant Vishwamitra, Elizabeth Whit- taker, Joseph P Mazer, Robin Kowalski, Hongxin Hu, Feng Luo, Jamie Macbeth, and Edward Dillon. Cyberbullying detection with a pronunciation based convolutional neural network. In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 740–745. IEEE, 2016.

[7] Nektaria Potha and Manolis Maragoudakis. Cyberbullying detection using time series modeling. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, pages 373–382. IEEE, 2014.

[8] Rui Zhao, Anna Zhou, and Kezhi Mao. Automatic detection of cyberbullying on social networks based on

bullying features. In Proceedings of the 17th international conference on distributed computing and networking, page 43. ACM, 2016.

[9] Q. You, Z. Zhang, and J. Luo, "End-to-end convolutional semantic embeddings," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5735–5744, Salt Lake City, UT, USA, June 2018.

[10] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasert-silp, Piyaporn Nurarak, and Pirom Konglerd. Automated cyberbullying detection using clustering appearance patterns. In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242–247. IEEE, 2017.

[11] 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)