International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Cyberbullying Detection on Social Networks Using Machine Learning Approaches

Adya Bansal, Akash Baliyan, Akash Yadav, Aman Kamlesh, Hemant Kumar Baranwal

Dept. of Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, U.P. India

---

Abstract - *The user on social media has saw a boom in recent time with increase in number of users of the net and emerges as the major networking platform of our era. But it also has its own repercussions on society and the mental health of a person such as online abuse, harassment, scamming, private information leaks and trolling. Cyberbullying affects a person both physically and mentally, particularly for girls and students, and sometimes escalated to their suicide. Online harassment has a huge bad impact on society. No of cases have occurred in different parts due to online bullying, such as sharing private information, abusing someone online, and racial discrimination. So, there is a need for identification of bullying on social apps and this has become a major concern all over the world. Our motive for this research is to compare different techniques to find out the most effective technique to detect online harassment by merging NLP (natural language processing) with ML (machine learning).*

Key Words: ***Cyber bullying, Machine Learning, NLP, Social apps.***

1.INTRODUCTION

Online apps for socializing are a place where we can express our thoughts and opinions and can even share our personal lives with another person [1]. We can access social media with the help of internet connection in our phones, laptops, PCs, tablets etc. The most well-known online media incorporates Facebook1, Twitter2, Instagram3, Tik-Tok4, etc. These days, web-based media is engaged with various areas like Coaching [2], Entrepreneurship [3], and furthermore for the respectable objective [4]. Online media is likewise improving the world's economy through setting out many new position open doors [5].

[1]https://www.facebook.com/
[2]https://twitter.com/
[3]https://www.instagram.com/

Albeit web-based media has a great deal of benefits, it likewise has a few downsides. Utilizing this media, malignant clients lead exploitative and deceitful demonstrations to offend and harm their notoriety. As of late, cyberbullying emerges as the significant web-based social apps issue.

Cyberbullying or digital badgering alludes to an electronic strategy for tormenting or provocation. Cyberbullying and digital badgering are otherwise called internet tormenting. With the boom and advancement of the social apps like Twitter, Facebook, cyberbullying has become normal in the life, especially in the life of students.

Roughly half of the young people in America experience cyberbullying [6]. This tormenting intellectually affects the casualty [7]. The casualties pick reckless behaves like self-destruction in light of the fact that the injury of cyberbullying which is difficult to be suffered [8]. Subsequently, the identifying and avoidance of cyberbullying is essential to secure youngsters.

According to situation, we recommend a cyberbullying recognition system in view of AI to recognize whether or not a text connects with cyberbullying. We have explored a few AI algorithms in our research to find out the best one among them which can be used to detect cyberbullying more precisely. We direct examinations with datasets from twitter comments and remarks. In execution investigation, we utilize two distinctive component vectors BOW and TF-IDF. The outcomes show that TF-IDF highlight gives preferable exactness over BOW where SVM gives preferred execution over some other AI calculations used for our research.

Rest of the research is coordinated in the following pattern. Section 2 outlines the connected works. Section 3 presents the subtleties of the given AI comparison. Section 4 shows the final results of research. Section 5 finishes up the paper and features some potential work which can be done in future.

2. CONNECTED WORKS

There are a few deals with AI based digital tormenting identification. A regulated AI calculation was proposed utilizing a sack of words way to deal with identifying the context and intention of the writer of the text [9]. This calculation shows scarcely 61.9% of exactness. MIT (Massachusetts Institute of Technology) directed a venture called Ruminati [10] utilizing SVM to detect cyberbullying of Twitter remarks. The scientist joined location with rational thinking by adding social parameters. The after effect of this undertaking was improved to 66.7% precision for applying probabilistic displaying. Reynolds et al. [11] discover an even more effective cyber bullying technique with an increased accuracy of 78.5%. The decision tree and occasion-based mentor is utilized by creator to accomplish this accuracy. To improve online harassment recognition, the creator of research [12] has utilized characters, feeling and opinion of element.

A few profound learning-based models were additionally acquainted with identify the cyberbullying. Profound Neural Network-based model is used for detecting online bullying by utilizing genuine information [13]. The creators first investigate cyberbullying methodically then used that data to learn the AI about automatic detection of bullying. Badjatiyaet al. [14] has introduced a technique involving profound Neural organization models for identifying disdain discourse. A convolutional neural organization-based model is used to identify online bullying [15]. The creators utilized word inserting where comparative words have comparative installing. In a multi-modular setting, Cheng et al. [16] paper the original study of online bullying identification by cooperatively taking advantage of web-based media information. This test, nonetheless, is difficult because of the intricate blend of both cross-modular relationship among numerous strategies and underlying connections between different web-based interactive conversations, arrangement of various complex models and modes of conversations. They propose Bully, online harassment detection framework to come out of the difficulties, which first change multi-modes of social apps information as a diverse organization and afterward attempts to do hub implanting portrayals onto that information.

Numerous writings about online harassment focused on examination of what is written throughout recent many years. But Cyberbullying, be that as it may, is now changing now it not only present in the form of text only. The assortment of tormenting information of friendly stages can't be achieved only by text detecting algorithms only.

Wang et al. [17] recommended a multi-modes detection framework that coordinates multiple types of data, for example, gif, images, vulgar comments, time via online media to adapt to the most recent kind of harassment. Specifically, they remove printed attributes, yet additionally apply progressive consideration organizations to catch the informal organization meeting capacity and encode various types of data including gifs, pictures. The creators model the multi-modes harassment discovery structure to for addressing these new kind of online bullying ways other than text.

Utilizing Neural Networks to work with the identification of web-based harassing become a common norm today. These Neural Networks originally founded exclusive for or related to other types of Layers by use of Long-Short-Term-Memory layers. Buan et al. [18] presented another model for the Neural Network that can be used in words-based media to identify whether there is online bullying or not. The idea is based upon current models that combine the strength of Long-Short-Term-Memory layers with Coevolutionary layers. Along with this, the design includes the use of stacked center layers, which shows that their review improves the Neural Network's performance. A different type of enactment strategy is also remembered for our plan, that is classified" SVM like initiation" By involving the weight L2 regularization of a straight actuation work in the actuation layer along with utilizing a misfortune work, the" SVM like accuracy" is achieved.

Making an AI framework with three unmistakable highlights, by Raisi et al. [19] solves the issue of computation connected with badgering identification in interpersonal organizations. (1) In this type the key expressions which is given by expert which distinguish bullying from non-bullying, with minimal supervision required. (2) A total no. of two students who co-train each other, in which one student study the content of the language of text and the second student looks at social construction aspect. (3) On preparing nonlinear profound models, this coordinate decentralized word and chart hub portrayals. Upgrading a genuine capacity that consolidates co-preparing along with weak supervision, and the model is trained.

Cyberbullying has as of late been identified by clients of online interpersonal organizations as significant medical of issue of public and the production of compelling discovery model with extensive scientific merit. Al et al. [20] have presented an assortment of specific Twitter-inferred highlights including conduct, client, and tweet content. They have assembled a directed AI answer for the discovery of cyberbullying on Twitter based organization. As shown by their appraisal, in view of highlights proposed by them, their set up recognition framework working under the recipient trademark acquired results with a locale bend of 0.943 and an f-measure of 0.936.

For those impacted, cyber bullying can create serious mental issues in the person. So, there is a need to plan Automated approaches for the recognition of cyber bullying. Albeit late cyber bullying identification endeavors have set up brand new

algorithms for text handling to identify cyber harassment, there are as yet couple of endeavors when it comes to recognize cyber bullying of visual means i.e., using images etc. Singh et al. [21] revealed that images components support highlight vectors in identification of online harassment in view of early investigation of a public, named online bullying dataset, which can help in predicting online bullying. When cyber bullying is turning out to be an ever-increasing number of normal in interpersonal organizations, it is the fate of outrageous significance to promptly recognize and responsively react upon this. The work in [22] explored how Fuzzy Fingerprints, a new method with recorded viability in similar undertakings, works while distinguishing literary harassment in social sites.

3. AIM OF THE PROJECT

The aim of this project is to classify a set of data as cyberbullying or not and to compare five machine learning algorithms and find the most accurate one for the classification.

4. METHODOLOGY

We will develop this project with the help of python and web technology. Using html and CSS, we will design and develop the web interfaces for the project. Then after preparing the web interfaces, we will search and download the dataset that we need to classify. After downloading the dataset, we will pre-process the data and then transfer to Tf-Idf. Then we will generate codes for the machine learning algorithms (Naive Bayes, Decision Tree, Random Forest, SVM, DNN Model) using python. So here, we are using python as backend and for frontend html, CSS etc.
The real-world posts or text contain number of unnecessary symbols or texts. For instance, emojis and symbols are not needed to detect cyber bullying. Hence, first they are removed and then machine learning algorithms are applied for the identification of bullying text. In this phase, the task is to remove unnecessary characters like symbols, emojis, numbers, links etc. And after those two important features of the text is prepared:

•**Bag-of-Word:** The machine learning algorithms are not going to work directly with texts. So, we have to convert them into some other form like numbers or vectors before applying machine learning algorithm to them. In this way the data is converted by Bag-of-Words (BOW) so that it can be ready to use in next round.

•**TF-IDF**: One of the important features to be considered is this. TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure to know the importance that a word carries in a document.

4.1. Machine Learning

In this model we will apply five efficient algorithms used in machine learning namely- Random Forest, Decision tree, Naive Bayes, SVM and Deep Neural Networks Model (DNN) to compare them and find the most accurate algorithm among them. The algorithm having highest accuracy is discovered among the five algorithms using public datasets.

4.1.1. Decision Tree

This tree classifier can be utilized in both arrangement and relapse. It can assist with addressing the choice and choose both. Decision tree has a design which resembles a tree like structure in which the parent/root hub is a condition, and descending parent hub is leaf/branch hub which is a choice of the condition. For ex. If the root hub is the coin than its branch hub will be the outcome of the coin i.e., head and tail. A relapse tree yields the anticipated incentive for a tended to enter.

4.1.2. Naive Bayes

This is a productive AI calculation in view of Bayes hypothesis. It predicts about the probability of occurring of an event based upon the event which already occurred previously. And when we add naïve assumption into it became the Naïve Bayes classifier.
In naïve bayes assumption we consider that each event is independent of each other and is going to make an equal contribution to final result. The best use of it is the classification of text which required a high dimensional training dataset. As stated earlier it consider that each event is independent so it cannot be used for events having relationship between them.

4.1.3. Random Forest

It is a classifier which comprises of different choice tree classifiers. It is created by using subset of data and the final output of that data is based on majority ranking means the higher votes. It is slower than the decision tree which we have discussed earlier as it contains not only one but a large number of decision tree which comes together to form a forest and hence the name random forest. The greater number of decision tree are going to be present in random forest the more precise the output is going to be. Unlike decision tree it doesn't use any set of rules or formulas to show the output.

4.1.4. Support Vector Machine

Support Vector Machine (SVM) is a regulated AI calculation which can be applied in both order and relapse the same a choice tree. It can recognize the classes extraordinarily in n-layered space. Along these lines, SVM produces a more precise outcome than different calculations significantly quicker. By and by, SVM develops set an of hyper planes in a limitless layered space and SVM is executed with part which changes an information space into the necessary structure. For instance, Linear Kernel involves the ordinary spot result of any two examples as follows:

K (y, yi) = aggregate (y ∗ yi)

4.1.5. DNN Model

It consists of many layered computations which is performed together. A neural network has layers known as: hidden layers, input layers and output layers and in case if the hidden layer is two or more the two than we can call it as deep neural network. It can be considered as the improved version of ANN (Artificial Neural Network). This model has recently become very popular due to its accuracy over other algorithms.
While training the dataset on DNN model an input vector is need to be collected. The training consists of two passes forward pass and backward pass. In forward pass a non-linear activation layer is calculated from input to output layer one by one. In backward pass we move in reverse order from output layer to input layer while calculating the error function.

5. EXPERIMENT AND RESULTS

In this research we have used five machine learning algorithm and studied them carefully to find out the algorithm with best accuracy among all five of them. To find out which algorithm is best we trained each one of them on same dataset so that we can compare them with each other more precisely. The five algorithms are DNN, SVM, RF, DT and NB so first we are going to use these algorithms on dataset one by one and then we will discuss the result shown by each one of them to find out which one is best among them.

5.1. Dataset

The dataset used for this study is downloaded from website called kaggle.com [27]. The dataset contains two types of set which are bullying text and non-bullying text. The goal is to identify all the bullying text.

•**Non-bullying Text:** The text which is not demeaning or hurtful but is a legit compliment or respectful criticism of the work of an individual. For example, comments such as "This girl is cute" are somewhat humane and demeaning.
•**Bullying text:** The comments which is hurtful and abusive in nature or are promoting racism, body shaming, casteism, slut shaming etc. comes in the category of bullying text. For example, "This bitch is ugly", "you should die" are the text which is straight up bullying someone which can affect their mental health severely.
So, with the use of python machine learning packages algorithms known as troubleshooting algorithm is implemented.

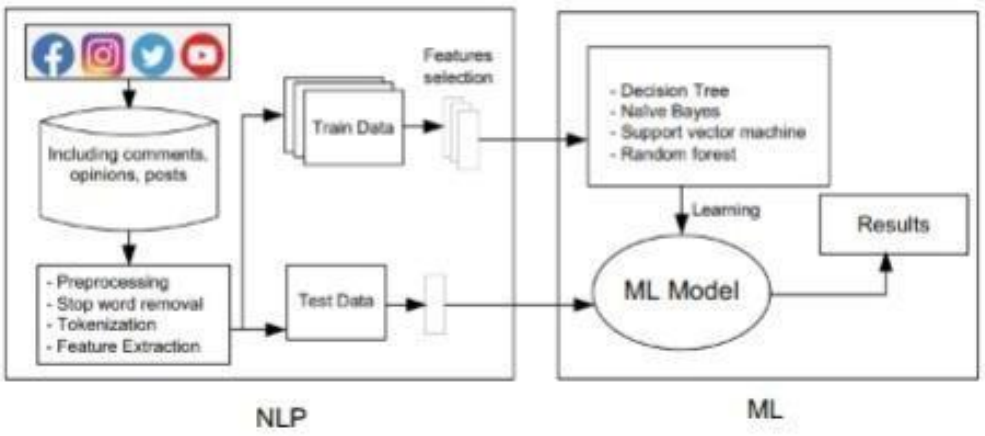
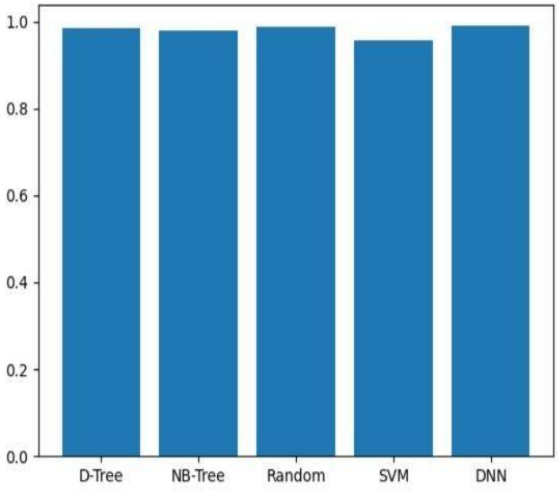


Fig -1: Designed model for detecting cyberbullying

5.2. Accuracy of different algorithms

In the given graph shown in Chart-1. We are comparing all the five algorithms with each other to find out which algorithm is best among all five. This graph has been plotted with the help of Mat plot library. We observed that among the five machine learning algorithms, DNN Model outperforms the others while Random Forest comes second, Decision Tree- third, Naïve Bayes comes second last and SVM is the least accurate. So according to this result we can safely say that it is better to use DNN Model for detecting the cyberbullying than any other algorithm.



Decision Tree	0.985731
NB Tree	0.979898
Random Forest	0.986897
SVM	0.956987
DNN	0.990145

Chart -1: Accuracy graph of ML algorithms

6. CONCLUSIONS

As the users of social media is increasing day by day along with-it cyber bullying related cases is also increasing on social media with its growing popularity and the increasing usage of social media by young people. It is necessary to devise an automated method of detecting cyberbullying in order to avoid the harmful effects of cyberbullying before it's too late. As sometimes the consequences of cyberbullying can be as bad as the suicide by the person that is bullied. So, keeping in mind the importance of a system which can detect the cyberbullying and online harassment, we are going to study different ML algorithms and their effectiveness in comparison with each other to predict their accuracy on a given data set to find out the best among them. After studying all the five algorithms and their results we come to a conclusion that DNN model perform best in detecting cyberbullying with an accuracy of 0.990145 and along with the second-best performing algorithm comes out to be random forest algorithm with an accuracy of 0.986897.
So, we can use any of these two algorithms to detect the cyberbullying and online harassment to get the highest accuracy while SVM is the least accurate among all of them.

ACKNOWLEDGEMENT

All the work and research done in this project is supported by M.I.E.T(Meerut Institute of Engineering and Technology), Meerut.

REFERENCES

[1] C. Fuchs, social media: A critical introduction. Sage, 2017.
[2] N. Selwyn, "Social media in higher education," The Europa world of learning, vol. 1, no. 3, pp. 1–10, 2012.
[3] H. Karjaluoto, P. Ulkuniemi, H. Keinanen, and O. Kuivalainen, "An- ¨ tecedents of social media b2b use in industrial marketing context: customers' view," Journal of Business & Industrial Marketing, 2015.
[4] W. Akram and R. Kumar, "A study on positive and negative effects of social media on society," International Journal of Computer Sciences and Engineering, vol. 5, no. 10, pp. 351–354, 2017.
[5] D. Tapscott et al., The digital economy. McGraw-Hill Education, 2015.
[6] S. Bastiaensens, H. Vandebosch, K. Poels, K. Van Cleemput, A. Desmet, and I. De Bourdeaudhuij, "Cyberbullying on social network sites. an experimental study into bystanders' behavioral intentions to help the victim or reinforce the bully," Computers in Human Behavior, vol. 31, pp. 259–271, 2014.
[7] D. L. Hoff and S. N. Mitchell, "Cyberbullying: Causes, effects, and remedies," Journal of Educational Administration, 2009. [8]
[8] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," Archives of suicide research, vol. 14, no. 3, pp. 206–221, 2010.
[9] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.
[10] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in In Proceedings of the Social Mobile Web. Citeseer, 2011.
[11] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in 2011 10th International Conference on Machine learning and applications and workshops, vol. 2. IEEE, 2011, pp. 241–244.
[12] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using twitter users' psychological features and machine learning," Computers & Security, vol. 90, p. 101710, 2020.
[13] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," in European Conference on Information Retrieval. Springer, 2018, pp. 141–153.
[14] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.
[15] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, 2018.
[16] L. Cheng, J. Li, Y. N. Silva, D. L. Hall, and H. Liu, "Xbully: Cyberbullying detection within a multi-modal context," in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 339–347.
[17] K. Wang, Q. Xiong, C. Wu, M. Gao, and Y. Yu, "Multi-modal cyberbullying detection on social networks," in 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020, pp. 1–8.
[18] T. A. Buan and R. Ramachandra, "Automated cyberbullying detection in social media using a sum activated stacked convolution lstm network," in Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis, 2020, pp. 170–174.

[19] E. Raisi and B. Huang, "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2018, pp. 479–486.

[20] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," Computers in Human Behavior, vol. 63, pp. 433–443, 2016.

[21] V. K. Singh, S. Ghosh, and C. Jose, "Toward multimodal cyberbullying detection," in Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2017, pp. 2090–2099.

[22] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, "Using fuzzy fingerprints for cyberbullying detection in social networks," in 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2018, pp. 1–7.

[23] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21, no. 3, pp. 660–674, 1991.

[24] I. Rish et al., "An empirical study of the naive bayes classifier," IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, pp. 41–46, 2001.