ISSN: 2454-9940

# INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 17, Issue 1, 2023

# A Machine Learning Framework for Detecting and Mitigating Online Harm on Social Media Platforms

**Mrs.V.Tejaswi[1]., T.Kaveri[2]**

1 Assistant Professor, Department of CSE, Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

2, B.Tech CSE (19RG1A0556),

Malla Reddy College of Engineering for Women., Maisammaguda., Medchal., TS, India

**Abstract -**

***With more people using the internet, social media has seen a surge in recent years and is now the main networking platform of our day. However, it also has negative effects on society and an individual's mental health, including trolling, abuse, harassment, and scams conducted online. Cyberbullying has a negative psychological and physical impact on a person, especially on females and kids, and may even lead to suicidal thoughts. The negative effects of cyberbullying on society are enormous. Numerous incidents of internet bullying, including disclosing personal information, assaulting someone online, and racial prejudice, have happened in various places. Therefore, it is necessary to identify bullying on social media platforms, and this is a global problem. Our goal is to use natural language processing (NLP) and machine learning (ML) to identify the most efficient method for detecting online harassment by comparing several approaches.***

***Key words: NLP, social applications, machine learning, and cyberbullying.***

## 1.INTRODUCTION

We may share our personal lives with others and express our ideas and feelings using socializing applications on the internet [1]. With the aid of an internet connection, we may browse social media on our phones, laptops, PCs, tablets, etc. Online media that is most well-known includes Facebook, Twitter, Instagram, Tik-Tok, and so on. Web-based media is involved in a number of fields these days, including coaching [2], entrepreneurship [3], and even for the noble purpose [4]. Online media is also helping the global economy by creating a lot of new job opportunities [5].

[1]https://www.facebook.com/

[2]https://twitter.com/

[3]https://www.instagram.com/

Web-based media has many advantages, but it also has certain drawbacks. Malicious clients use this medium to stage deceptive and exploitative performances in an effort to offend and damage their reputation. Cyber bullying is becoming a major problem with web-based social applications. Digital badgering, often known as cyber bullying, refers to an online provocation or taunting tactic. Other names for cyber bullying and digital badgering include online tormenting. The proliferation and development of social media platforms such as Facebook and Twitter has made cyberbullying a common occurrence, particularly among youngsters.

In America, around half of the youth have been victims of cyber bullying [6]. The victim is affected cognitively by this distressing [7]. Given the difficulty of suffering from cyber bullying injuries, the victims choose heedless behaviors such as self-destruction [8]. Therefore, it's critical to recognize and steer clear of cyber bullying in order to protect children. In light of AI's ability to determine whether a text is connected to cyber bullying or not, we suggest a cyber bullying recognition system based on the circumstances. In order to determine which AI algorithm is most effective in accurately detecting cyber bullying, we have investigated a number of them in our study. We conduct analyses using datasets derived from remarks and comments on Twitter. We use two unique component vectors in our execution investigation: TF-IDF and BOW. The results indicate that TF-IDF highlight provides better accuracy than BOW, while SVM provides better performance than several other AI computations we

used in our study. The following pattern guides the coordination of the remaining research. The linked works are delineated in Section 2. The intricacies of the provided AI comparison are presented in Section 3. The final study findings are shown in Section 4. The last section of the report concludes with some possible future research topics.

## 2. CONNECTED WORKS

A few trades use digital tormenting identification based on artificial intelligence. To address the issue of determining the context and writer's purpose of the text, a controlled artificial intelligence computation based on a bag of words was presented [9]. The accuracy of this computation is just 61.9%. A project named Ruminati [10] was headed by MIT (Massachusetts Institute of Technology) and used SVM to identify cyber bullying in comments on Twitter. The scientist added social dimensions, tying place and reason together. The project's outcome increased the application of probabilistic displaying accuracy to 66.7%. Reynolds et al. [11] find a cyber bullying method that is even more successful, with a 78.5% accuracy rate. The creator uses an occasion-based mentor and decision tree to achieve this precision. The researcher [12] has used elemental characteristics, emotions, and opinions to enhance the identification of online abuse.

A few deep learning-based models were also trained to recognize instances of cyber bullying. By using real data, a model based on profound neural networks is employed to identify cyber bullying [13]. The inventors conducted a thorough investigation into cyber bullying before using the collected data to teach the AI how to recognize bullying automatically. For the purpose of recognizing talk that is derogatory, Badjatiya et al. [14] have presented a method that uses deep neural organization models. Online bullying is detected using a model based on convolutional neural organization [15]. Where comparable words have comparative installing, the authors used word insertion. Cheng et al. [16] report the first investigation of online bullying detection in a multi-modular scenario by collaboratively using web-based media content. However, this test is challenging due to the complicated combination of sophisticated models and conversational modes, as well as the cross-modular relationships across several methods and underlying links between various web-based interactive discussions. In order to overcome the challenges, they suggest the Bully online harassment detection framework, which modifies the information on social media applications in several ways as a varied group before attempting to perform hub implanting representations onto that data.

Many articles about cyber bullying concentrated on analyzing writing over the last few years. That being said, cyber bullying is evolving and is no longer limited to textual communication. Text detection techniques are not sufficient to get the variety of distressing data of friendly phases.

In order to keep up with the latest forms of harassment, Wang et al. [17] suggested a multi-modes detection system that synchronizes several data kinds, such as gifs, photos, nasty remarks, and time via the media. To be more precise, they eliminate written qualities while also encoding other kinds of data, such as gifs and photographs, and using progressive consideration organizations to capture the informal organization meeting capability. In order to confront these new forms of online bullying in methods other than texting, the designers modeled the multi-modes harassment discovery framework.

Today, it's normal practice to use neural networks to assist in the detection of web-based harassment. By using Long-Short-Term-Memory layers, these neural networks were first developed exclusively for or in relation to other kinds of Layers. Another neural network approach, introduced by Buan et al. [18], may be used to word-based media to determine the presence or absence of online bullying. The concept is based on existing models that combine co evolutionary layers with the strength of long-short-term-memory layers. In addition, the inclusion of stacked center layers in the architecture demonstrates how their evaluation enhances the performance of the neural network. We also keep in mind another kind of enactment technique for our plan, which is categorized as "SVM like initiation." The "SVM like accuracy" is attained by using a misfortune work in conjunction with the weight L2 regularization of a straight actuation work in the actuation layer.

By creating an AI system with three distinct features, Raisi et al. [19] address the computational problem associated with badgering detection in social organizations. (1) In this sort of situation, there is less need for monitoring since the crucial phrases that separate bullying from non-bullying are provided by experts. (2) A maximum of two students who work together as co-trainers, with one studying the text's linguistic content and the other the social construction component. In the process of creating nonlinear deep models, this coordinates decentralized word and graphic hub representations. developing a real capability that combines light supervision and co-preparation, and the model is trained.

## 3. AIM OF THE PROJECT

ISSN 2454-9940

www.ijasem.org

Vol 17, Issue 1, 2023

The purpose of this research is to examine five machine learning algorithms and determine which one is the most accurate in classifying a piece of data as either cyber bullying or not.

## 4. METHODOLOGY

We will use web technologies and Python to construct this project. We will design and construct the project's web interfaces using HTML and CSS. Next, after the online interfaces are ready, we will search for and download the dataset that has to be classified. We will pre-process the data after downloading it, and then we will upload it to Tf-Idf. Next, we will use Python to create the codes for the machine learning algorithms (DNN Model, SVM, Random Forest, Decision Tree, and Naive Bayes). Thus, the backend in this case is Python, while the frontend uses HTML, CSS, and other elements. There are a lot of superfluous symbols and words in the real-world messages and content. Emojis and symbols, for example, are not required to identify cyber bullying. Therefore, in order to identify bullying material, machine learning methods are performed first and these are eliminated. The goal of this step is to eliminate any extraneous characters, such as links, emojis, numbers, and symbols. And after the preparation of those two crucial textual elements:

•Bag-of-Word: Texts will not be directly processed by machine learning techniques. Therefore, before using a machine learning method on them, we must transform them into another format, such as integers or vectors. In this manner, Bag-of-Words (BOW) transforms the data so that it is prepared for use in the subsequent round.

•TF-IDF: This is a crucial element that has to be taken into account. Term Frequency-Inverse Document Frequency, or TF-IDF, is a statistical indicator of a word's significance in a document.

### 4.1. Understanding Machine Learning

We will examine and determine which of the five effective machine learning algorithms—Random Forest, Decision Tree, Naive Bayes, Support Vector Machines, and Deep Neural Networks Model (DNN)—is the most accurate by using it in this model. Using open datasets, the five algorithms are compared to find the method with the greatest accuracy.

#### 4.1.1: Decision Tree

It is possible to use this tree classifier for both arrangement and relapse. It may help address the decision and choose both. The architecture of a decision tree is similar to a tree; the condition is represented by the parent/root hub, and the option of the condition is represented by the descending parent hub, which is the leaf/branch hub. For example, if the coin represents the root hub, then the coin's branch hub will represent the head and tail of the coin. The expected motivation for a client to enter is provided by a relapse tree.

#### 4.1.2. Unsupervised Bayes

In light of the Bayes hypothesis, this AI computation is fruitful. It makes predictions on the likelihood that an event will occur based on past events. And the Naïve Bayes classifier was created when we added naïve assumptions to it.

Under the naïve Bayes assumption, every event is assumed to be independent of the others and to contribute equally to the outcome. It works best when used for text classification, which calls for a high dimensional training dataset. It cannot be applied to events that have a connection with one another since, as was previously said, it treats every event as separate.

#### 4.1.3. The Forest of Random

A variety of choice tree classifiers make up this classifier. It is produced using a subset of data, and the majority ranking—that is, the more votes—determines the final output of that data. Because it comprises several decision trees that combine to create a forest, thus the term "random forest," it is slower than the decision tree that we previously examined. The more decision trees that are included in the random forest, the more accurate the result will be. It displays the outcome without the need of formulae or rules, as contrast to decision trees.

#### 4.1.4: Support Vector Computer

A controlled artificial intelligence algorithm that may be used to both order and relapse decision trees is called Support Vector Machine (SVM). In n-layered space, it has extraordinary class recognition capabilities. Accordingly, SVM generates a more accurate result comparatively to other computations much faster. SVM eventually creates a collection of hyper planes in an infinite layered space, and it is implemented in a way that transforms an information space into the required structure. For example, the

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 17, Issue 1, 2023

conventional spot result of any two cases is involved in the Linear Kernel as follows:

$$K \quad (y, yi) = \text{aggregate}\ (y * yi)$$

#### 4.1.5. DNN Model

It is made up of several layered calculations that are carried out concurrently. A neural network consists of three levels: input, output, and hidden layers. If a neural network contains two or more hidden layers, it is referred to as a deep neural network. It may be seen as an enhanced ANN (Artificial Neural Network). Because this model is more accurate than other algorithms, it has lately gained a lot of popularity.
An input vector must be gathered in order to train the dataset using the DNN model. Two passes—a forward pass and a backward pass—make up the training. A non-linear activation layer in forward pass is computed one by one from the input layer to the output layer. When computing the error function in a backward pass, we go from the output layer to the input layer in reverse order.

## 5. EXPERIMENT AND RESULTS

In order to determine which machine learning algorithm had the highest accuracy out of the five, we used five in this study and thoroughly examined each one. We trained all of the algorithms on the same dataset in order to compare them more closely and determine which one is the best. The five methods are DNN, SVM, RF, DT, and NB. We will first apply each of these algorithms on a different dataset, and then we will talk about the results that each algorithm produces to determine which is the best.

### 5.1. Information Base

The dataset used in this investigation was obtained from the kaggle.com website [27]. Bullying text and non-bullying text are the two sorts of sets that are present in the dataset. Finding every instance of bullying text is the aim. Texts that are not bullying: These are remarks about someone's job that are politely critical of them rather than degrading or harmful. For instance, remarks like "This girl is cute" are both degrading and in some ways kind. Bullying text refers to remarks that are harmful and abusive in character, or that propagate racism, body shaming, casteism, slut shaming, and other similar practices. Texts like "This bitch is ugly" and "you should die" are examples of explicit bullying that may have a very negative impact on someone's mental health.

In light of this, troubleshooting methods are built using Python machine learning libraries.

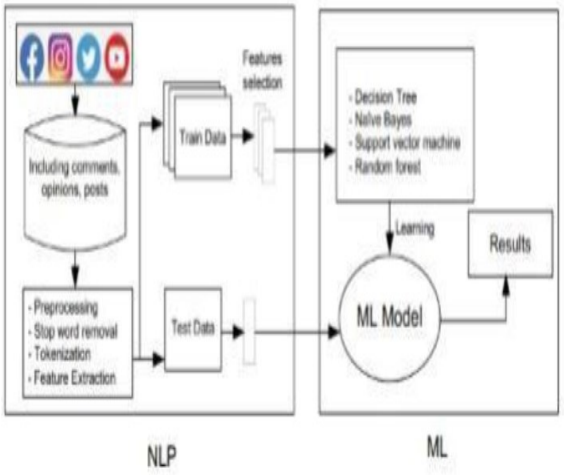
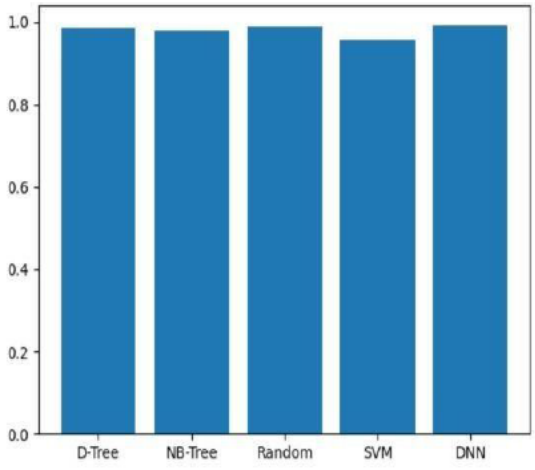



**Fig -1**: created a strategy to identify cyber bullying

### 5.2. The precision of various algorithms

The provided graph is shown in Chart-1. To determine which of the five algorithms is the best, we are comparing them to one another. The Mat plot library was used to plot this graph. Among the five machine learning algorithms, we found that DNN Model performs the best, followed by Random Forest in second place, Decision Tree in third place, Naïve Bayes in second place, and SVM in the lowest accuracy. Thus, based on this finding, we can confidently state that the DNN Model outperforms all other algorithms in terms of cyber bullying detection.



ISSN 2454-9940

www.ijasem.org

Vol 17, Issue 1, 2023

| | |
|---|---|
| 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

Because social media is becoming more and more popular, young people are using it more often, and as a result, there are a rising number of cyber bullying incidents on these platforms. To prevent the negative impacts of cyber bullying before it's too late, an automated technique of recognizing it must be developed. Because the effects of cyber bullying may sometimes be just as severe as the victim of the bullying taking their own life. In light of the significance of developing a system that can identify instances of cyber bullying and online harassment, we will examine many machine learning algorithms and compare their efficacy in order to determine which one is most likely to accurately forecast data on a certain collection of examples. Upon examining all five methods and their outcomes, we conclude that the DNN model, with an accuracy of 0.990145, outperforms the other two algorithms in terms of cyber bullying detection. The random forest technique, with an accuracy of 0.986897, is the second-best performing algorithm.

Thus, SVM is the least accurate algorithm among all of them, and we may use either of these two to identify cyber bullying and online harassment with the maximum accuracy.

## REFERENCES

[*1] C. Fuchs, social media: A critical introduction. Sage, 2017.*

*[2] N. Selwyn, "Social media in higher education," The Europe 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.*