iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 8 **Issue:** VII **Month of publication:** July 2020

**DOI:** https://doi.org/10.22214/ijraset.2020.30403

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 8 Issue VII July 2020- Available at www.ijraset.com*

# Detection of Cyberbullying using Machine Learning

Sinchana C[1], Sinchana K[3], Pradyumna C S[5], Janhavi V[2], Deepika S[4],
[1, 2, 3, 4, 5]*Dept. Of Computer Science And Engineering, Vidyavardhaka College of Engineering,* Mysore, Karnataka

***Abstract:*** ***Cyberbullying is a type of tormenting wherein technology is utilized as a medium to menace somebody. As the new blast of the web and other social media platforms are expanding, the quantity of users is additionally expanding and the primary users of online networking are for the most part adolescents and young adults. As much as these social media platforms are utilized for getting new data and for amusement, it is increasingly inclined for bullies to utilizes these systems as helpless against assaults against casualties. Because of the expansion in cyberbullying on casualties, it is deprived to build up an appropriate strategy for the identification and anticipation of cyberbullying. A developing assortment of work is rising on mechanized ways to deal with cyberbullying location. These methodologies use machine learning and natural language processing techniques to identify the characteristics of a cyberbullying exchange and automatically detect cyberbullying by matching Textual data. The primary goal of this task is to distinguish cyberbullying by coordinating both Image and Textual information. The test cases are utilized to characterize the dataset and distinguish the bullying. Machine learning techniques are utilized to proficiently anticipate and identify cyberbullying.***

***Keywords: Cyberbullying, support vector machine, k nearest neighbor, Naïve Bayes, decision tree, Neural Network..***

## I. INTRODUCTION

Social networking sites are extraordinary instruments for interfacing with individuals. Regardless, as Social systems administration locales have become no matter how you look at it, people are finding unlawful and exploitative ways to deal with the use of these systems. We see that people, especially young people what's progressive, energetic adults, are discovering better ways to deal with danger to each other over the Internet. About 25% of guardians in an examination drove by Symantec declared that, most definitely, their child has been related to a cyberbullying event.

Cyberbullying is a kind of badgering using electronic techniques. Cyberbullying is known as internet harassing as well. It has gotten logically ordinary, especially among youths. Cyberbullying is where someone, threat or trouble others on web-based social networking locales. Dangerous torturing behavior can consolidate posting gossipy goodies, threats, sexual remarks, a victim's' own one of a kind information, or pejorative names (i.e., despise talk). Torturing or then again baiting can be perceived by repeated direct and a desire to hurt. Losses may have lower certainty, extended foolish ideation, and an arrangement of enthusiastic responses, checking being scared, baffled, incensed, and debilitated.

The figure 1 shows an example of cyberbullying where a person receives degrading comments on his/her post. Awareness in the United States has risen in the 2010s, due in part to prominent cases.




Figure 1: Example of CyberBullying

A couple of US states and various countries have laws unequivocal to cyberbullying. Regardless, what use are these laws if the cyberbullying cases are growing. Past work has been revolved around the area of cyberbullying after it recently happened. So we have advanced an endeavor to distinguish cyberbullying and alert the pros about these toward the starting time [1]. Our estimation

©IJRASET: All Rights are Reserved

uses a blend of outward appearance acknowledgment and Natural Language Processing (NLP) to perceive cyberbullying.
Outward appearance recognition can be used to recognize and get such an inclination in a picture posted on and social Cyberbullying Detection media. This combined with the treatment of comments on the specific post can give us a strong end concerning whether it is annoying of any kind.
Thus, many machine learning techniques are applied to detect cyberbullying.

## II. BACKGROUND AND RELATED WORK

Based on our initial survey regarding various techniques used for detection of cyberbullying, we feel, most of the approaches make use of only text based approach. The current solutions mostly address them using Bag Of Words technique or TF-IDF techniques and apply a classical Machine Learning Algorithm for classifying the comment as bullying or not [2]. Also, they also try to address a specific category of comments (Example-sexism). The drawbacks of the existing approach is that, it is not generalized enough to use it on any social media comments.

### *A. Content-Driven Detection Of Cyberbullying On The Instagram Social Network*

The aim of the research was to comprehend whether there is a relationship between shared media as posted pictures and subtitles, and the event of cyberbullying occasions. The work was primarily motivated by the following questions:

*1)* Can we further increment the precision in identifying harassing of shared pictures in the Instagram social network by utilizing relevant pieces of information, for example, images' features, image caption, and client metadata, including the number of follows/ - ers.

*2)* Is it conceivable to foresee occurrences of cyberbullying on a bit of shared substance dependent on a mix of logical highlights, i.e., features of the posted image itself, along with the caption and user metadata?

*a)* *Dataset:* A total of 9000 images were collected. In order to obtain contextual information about users' activities and profiles, along with each image, following data was collected:

- The user-created image caption, specific information about the user who posted the content (username, total post count, number of followings and number of followers)
- The text of the 150 most as of late posted remarks (or less, in situations where the total number of remarks for a picture was under 150)

Pictures and relating metadata were chosen randomly from a rundown of well known pictures on the site at the time of the crawl.
The dataset was cut to 3000 pictures by removing pictures with non-English language remarks and safeguarding from this subset the arrangement of pictures having the best number of remarks.
These pictures were then named in two distinct iterations, utilizing Mechanical Turk laborers. In the first place, the two pictures and remarks were introduced, and requested that labelers recognize whether the picture was harassed dependent on the picture's analysis. Next, labelers were approached to name each remark separately as either bullying or non-bullying.

*b)* *Image labeling:* Images were introduced to labelers with their comparing remarks. Labelers were approached to take a stake at the picture, read through the remarks and answer two numerous decision questions. In the first place, we asked whether the remarks incorporated any bullying, and second, on the off chance that an example of tormenting was available, we asked whether that bullying appeared to be because of the substance of the picture. Each picture with remarks was introduced to three particular labelers, and we considered a picture as having been tormented if 2 or 3 labelers reacted positively to it is possible that one of the two inquiries. Every other picture was marked non-bullied.

| Class | Count |
|---|---|
| Bullied Images | 560 |
| Non Bullied Images | 2540 |

Table 2.1: Dataset description

In total, 560 images were considered bullied and 2540 were not. Among those bullied, 19.2% were said to be bullied due to the controversial nature of the image, 21.13% due to the appearance of the subjects of the image, 3% because of the private nature of the image, while the remainder were said to be targeted for "other" reasons (e.g., popularity of the posting user, subjects of the

©IJRASET: All Rights are Reserved

image).

*c)* *Comment Labeling:* Users were solicited to name a subset from the comments, 30 comments each taken from 1120 pictures. Labelers approached the picture, the image's commentary, and showed whether each comment represented bullying.

*d)* *Feature Vector Construction:* A combination of text-based, image-based and meta- features will provide the strongest predictive power in this context.

Feature Set for Comments on Posted Content

- *Bag Of Words* - The "Bag of words" model (BoW) [Harris, 1954] is a baseline text feature wherein the given text is represented as a multi-set of its words, disregarding grammar and word order. We make a word vector, where every segment speaks to a word in the word reference we have created and its worth relates to its recurrence.
- *Offensiveness* - Following past work [Kontostathis et al., 2013] demonstrating that the event of second person pronouns in closeness to hostile words is profoundly characteristic of cyberbullying, we utilize an "offensiveness level" (OFF) feature [Chen et al., 2012]. We first utilize a parser to catch the syntactic conditions inside a sentence. Then for each word in the sentence, a word offensiveness level is calculated as the sum of its dependencies' intensity levels.

  We define the offensiveness level of a sentence:

$$O_s = \sum_w O_w \sum_{j=1}^{k} d_j$$

i ->Eq 2.1: To calculate Offensiveness level

Where Ow = 1 if word w is an offensive word, and 0 otherwise. For word w, there are k word dependencies, and d = 2 if dependent word j is a user identifier, d = 1.5 if it is an offensive word, and 1 otherwise.

- Word2Vec - Word2Vec is a state-of-art model for computing a continuous vector portrayal of individual words [Mikolov et al., 2013], regularly used to compute word likeness or anticipate the co-event of different words in a sentence.

*e)* *A Feature Set for Posted Content:* Their analysis of image content incorporated standard image specific features (i.e., SIFT, color histogram), many of which have been successfully used in other work for similar non descriptive research questions. They additionally consider more sophisticated features extracted with deep learning and leveraged using unsupervised clustering methods.

*f)* *Model:* In the family of supervised learning models, each model performs well for specific situations and ineffectively for other people. Heterogeneity of data, data redundancy interactions among features are considerations when selecting a method. They tested utilizing a multi-layer perceptron, a Bayesian classifier and a Support Vector Machine (SVM) [3][4]. Their best outcomes were acquired utilizing a SVM with a radial basis function (RBF) kernel in OpenCV. Hyperparameters of the SVM were streamlined utilizing the cross-approval gauge of the approval set blunder. They initially adjusted the dataset utilizing subsampling (950 bullying comments, 950 non-bullying). They utilized the standard k-fold validation technique (k = 10) to prepare and assess speculation precision. The concatenated BoW, OFF and Word2Vec feature set proves to be the most powerful combination of comment-based features [5]. As hypothesized, the addition of imaged-based features improves accuracy. As an ensemble, the concatenated feature set BoW, OFF, Word2Vec, Captions provides our strongest result at 95.00%.

*g)* *Results:* The values of the area under the precision/recall curve are: concatenated classifier (0.6573), Captions (0.8209), stacked classifier (0.8537), DLFS (0.7601), stacked classifier with FS (0.8308). (FS - Feature Selection).

### *B. Rule Based And Bag Of Words Model To Detect Cyberbullying*

This arrangement detects language patterns used by bullies and their victims, and create rules to consequently distinguish cyberbullying content.

The information utilized for this venture was collected from the site Formspring.me, an inquiry and-answer arranged site that contains a high level of tormenting content.

The information was named utilizing a web administration, Amazon's Mechanical Turk. So as to test the data collected, two main methods of machine learning is utilized: rule based learning and a bag-of-words approach. The marked information, related to machine learning techniques given by the Weka toolbox was utilized to prepare a PC to perceive bullying content. Both a C4.5 decision tree learner and an instance-based learner had the option to recognize the genuine positives with 78.5% accuracy. The best outcome from the bag-of-words approach yielded a 40% recall and 30.6% rank-464 statistic.

©IJRASET: All Rights are Reserved

Bag-of-words is a keyword based vector space model [6]. This model "perceives that the utilization of binary weights is excessively constraining and proposes a structure wherein partial matching is possible" [7]. Bag-of-words will permit us to create a matrix of the entire lexicon utilized in all training data. We would then be able to utilize this framework to run questions and decide the general closeness of the inquiry to each post in the matrix. When contrasting the rule-based model to the bag-of-words model we need to comprehend the subtleties of the measurements utilized in each arrangement of examinations to figure out which approach gave us better outcomes.

By comparing the results of both the rule-based and bag-of-words methods, it is evident that the rule-based method performed better. The review for the rule-based method is a lot higher, by and large higher than the review accomplished utilizing the bag-of-words method. To get a high review with bag-of-words, the precision must be seriously influenced. With this, it is obvious to see that the rule-based model outflanks the bag-of-words model.

### *C. Other Related Work*

A group of work is developing around the issue of cyberbullying, from different controls. Ongoing papers in developmental psychology and sociology have described the profiles and inspirations of guilty parties, and have examined potential systems of avoidance and intercession [Berson et al., 2002; Hinduja and Patchin, 2013]. Of note, these investigations feature the impact of the two companions and experts on empowering or alleviating cyberbullying practices.

These realities inspire improvement of novel ways to deal with automated detection of cyberbullying in online social networks. Mediation will require ID of launches of the issue and, in a perfect world, may follow from early notice systems when especially powerless substance is posted.

Inside software engineering, specialists have created techniques to consequently recognize cyberbullying, generally concentrating on text mining (for example [Yin et al., 2009; Dinakar et al., 2011; Kontostathis et al., 2013; Chen et al., 2012]). Surrounding the issue marginally in an unexpected way, others [Dadvar et al., 2013] have planned to identify the cyberbullies themselves, utilizing extra client highlights (e.g., geoposition) as well as hybrid machine learning/expert systems.

In existing methodologies, little (if any) attention is paid to setting, for example, the casualties' profiles, directed posted substance and the idea of clients' connections, which may all be vital in activating and cultivating bullying behavior [Sabella et al., 2013; Berson et al., 2002; Hinduja and Patchin, 2013].

[Yin et al., 2009] is the most similar to our work in that they adopt a supervised learning approach to identify cyberbullying utilizing substance and estimation highlights, just as logical highlights of the thought about records. Authors characterize setting by two measurements, both surveying the similitude of an offered post to different posts in its prompt region. Our work is distinct in several ways. Since we address cyberbullying of images in online social networks, our setting is given by highlights of the picture itself, posted captions, and the posting client.

We combine analysis of the text potentially containing abuse with these contextual features, using a combination of supervised and unsupervised learning approaches.

## III. METHODOLOGY

The system design is of 2 approaches. One is the text based approach and another one is the image based approach. In the text based approach, three different datasets have been used. They are formspring dataset, twitter dataset, Wikipedia dataset.

### *A. Data Preprocessing*

1) *Image Based Approach:* For the purpose of exploration, 159 training images from the JAFFE dataset would not be enough. Hence, instead of selecting random 48 x 48 patches, all of them were chosen. As a result, each 64 x 64 was in turn converted into 16 – 48 x 48 sized images. Since the original image size was small, more than 95% of all the facial features were conserved in the resultant patches.
2) *Text Based Approach:* In the text based approach, the raw text data will undergo the process of tokenization. Tokenization helps in tokenizing the sentences into words. Next, the tokenized data will be cleaned. The unwanted data such as URL, tagging, recurring spaces, and mentions will be removed. The cleaned data then will undergo the process of word embedding. That is the vector representation of particular words. The resulting imbalanced data will be oversampled to get the balanced processed data. Here, the replication of minority class takes place to balance the data.

©IJRASET: All Rights are Reserved

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 8 Issue VII July 2020- Available at www.ijraset.com*

## IV. MODEL BUILDING

### *A. Image Based Approach*

In the first step, facial emotion recognition on image using VGG_16 model. Next, from the comment, the API calls are made to get emotion. Finally the comparison of emotions of image and text comment takes place to conclude if the post was bullied or not.

### *B. Image Model - VGG16*

VGG16 (also called OxfordNet) is a convolutional neural network architecture. It was developed by Visual Geometry Group from Oxford, and hence is named after them. Today it is still considered to be an excellent vision model, although it has been somewhat outperformed by more recent advances such as Inception and ResNet. VGGNet consists of 16 convolution layers and is very attractive because of its very uniform architecture. Just like AlexNet, it has only 3x3 convolutions, but many filters. Trained on 4 GPUs for 2–3 weeks. It is presently the highly preferred choice for extracting features from images. The weight configuration of the VGGNet is available publicly and has been used in many of the applications and challenges. It is used as a baseline feature extractor. However, VGGNet includes of 138 million parameters, which can be a challenge in handling. This network is characterized because of its simplicity, it uses only 3×3 convolutional layers stacked on top of one another in increasing depth. Reducing volume size is handled by max pooling. Two fully-connected layers, each with 4,096 nodes are then followed by a softmax classifier.

### *C. API Calls to Get Emotion Expressed Over the Comment*

The publicly available APIs - Paralleldots is used. This model is built using NLP to detect the emotion of the comment. It provide an API to test the emotion of the given sentence.
The APIs take in sentence as input and return the emotion associated with them in terms of a dictionary.
Example: Output of the API Call
{'anger': 0.19738179445266724,
'fear': 0.22391769289970398,
'joy': 0.028553485870361328,
'sadness': 0.5333009958267212,
'surprise': 0.016845988109707832}

### *D. Compare The Emotion Of Image And Comment To Predict The Output*

To check if both the emotions expressed in comments and image are same, certain threshold for text data is used during comparison. A check is done if both of them express same emotion and decide based on this.

### *E. Text Based Approach*

LSTMs are used for all the text based models. 4 architectures of LSTMs were developed and were used to conclude if the text comment is cyberbullied or not.

### *F. LSTM*

Long Short Term Memory networks –are also just called “LSTMs”. They are a special kind of RNN. They are capable of learning long-term dependencies [10]. Hochreiter & Schmidhuber (1997) introduced them and were further refined, developed and were popularized by many people in following work. The are now widely used as they work very well on a large variety of problems. LSTMs are specially designed to prevent the long-term dependency problem. They are special because of their behavior of remembering information for long periods of time. All RNNs have the form of a chain. It is compose of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer. LSTMs also have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way. LSTMs have 3 Gates namely Forget Gate, Update gate and Output gate as shown in the figure below.

©IJRASET: All Rights are Reserved

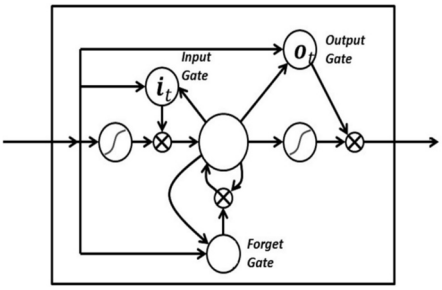



Figure 3.1: LSTM Gate.

The memory is either kept persistent or changed according to Update and Forget gate. Output gate controls activation of each output activation. These gates in LSTM allow the model to remember long-term dependencies in the data and hence are very efficient. On the contrary, the LSTM networks are complicated compared to their corresponding simple RNN networks. Also LSTM is computationally expensive to train over data. So care should be taken when using LSTM, main care is with respect to resources. LSTM require lot of resources.

Since LSTMs require lot of resources we made use of simple architectures that can run easily on our systems. We used small lstm networks containing embed size number of units. We created four models totally and at the end we consider the final output based on voting.

*G. Embedding Layer*

A word embedding is a class of approaches for the representation of words and documents by using a dense vector representation. It is an improvisation of the traditional bag-of-word model encoding schemes where large sparse vectors were used for the representation of each word or to score each word within a vector to represent an entire vocabulary [9]. These representations are rare because vocabularies are broad and a particular word or document is often represented by a large vector that contains zero values.

Instead, in an embedding, words are represented by dense vectors where a vector represents the projection of the word into a continuous vector space. The position of a word within the vector space is learned from text and is based on the words that surround the word when it is used. The position of a word in the learned vector space is referred to as its embedding. Two popular examples of methods of learning word embeddings from text include:

*1)* Word2Vec.

*2)* GloVe.

In addition to these carefully designed methods, a word embedding can be learned as part of a deep learning model. This can be a slower approach, but tailors the model to a specific training dataset. Word embeddings provide a dense representation of words and their relative meanings. They are an advancement over sparse representations used in simpler bag of word model representations [8]. Word embeddings can be learned from text data and reused among projects. They can also be learned as part of fitting a neural network on text data.

*H. Dropout Layer*

Dropout is a technique which is used for improving over-fit on neural networks. Dropout should be used along with other techniques like L2 Regularization. Usually, during training half of neurons on a particular layer will be deactivated. This improve generalization because force your layer to learn with different neurons the same "concept". During the prediction phase the dropout is deactivated.

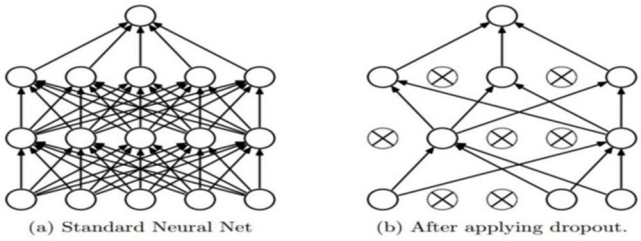



Figure 3.2: DROPOUT usage.

©IJRASET: All Rights are Reserved

Dropout is a technique or a procedure where some randomly selected neurons are ignored during the process of training. They are randomly "dropped-out".This means that their contribution to the activation of downstream neurons is being removed temporarily on the forward pass and any weight updates are not applied to the neuron on the backward pass. While the neural network learns, neuron weights settle into their context within the network. Weights of neurons are tuned for specific features thereby providing some specialization. Neighboring neurons will then rely on this specialization, which if taken too far may end in a fragile model too specialized for the training data. This reliant on context for a neuron during training is referred to complex co-adaptations. Normally some deep learning models use Dropout on the fully connected layers, but it is also possible to use dropout after max-pooling layers, which results in creating some kind of image noise augmentation.

## V. DESIGN

First, detection of cyberbullying using only comments that one enters takes place. Next, use of the image uploaded and the comments that one enters to detect cyberbullying. Then, we build fine-tuned user interface to make this user-friendly.

First a classifier to find the class (Bullying or non bullying) of the each comment is built. For that 4 architectures of LSTMs are built and are used to conclude if the text comment is cyberbullying or not.

Next is the image based approach. An emotion recognition architecture withVGG-16 is made. This captures the emotion in the picture that a user uploads and has a say in the final result (Bullying or not).

After obtaining the prediction from VGG-16, it is combined with the prediction of the text based approach on the comments that are entered and output a final result.

Here are some of the figure showing front end element for IBA and TBA.

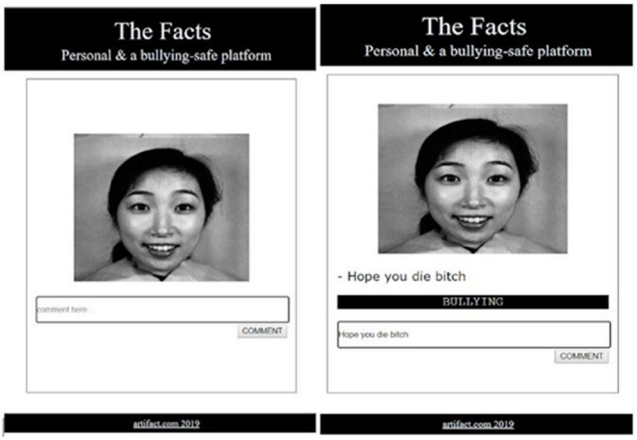



Figure 3.3: IBA element and output sample.

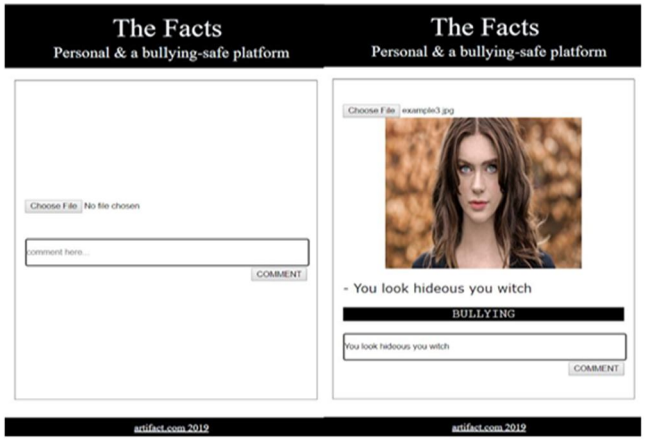



Figure 3.4: TBA element and output sample.

©IJRASET: All Rights are Reserved

## VI. RESULT AND DISCUSSION

IBA and TBA result analysis and testing is discussed in detailed below. The analysis and the testing gives a clear picture of the result.

### *A. IBA Result Analysis*

For our image based approach we can't give a presentation metric as it is unsupervised. However, we provide an analysis of the behavior of our algorithm. Some of the important points are:

*1)* This methodology possibly works when we have a picture with an individual's face.

*2)* This model performs acceptably well in situations where our TBA fizzles. This is on the grounds that TBA utilizes the context in just the comments, whereas the IBA utilizes the hidden context which is now and then helpful under certain circumstances.

### *B. TBA Result Analysis and Testing*

All the 4 models were individually trained and unit-tested. Following were the results: In each model we will see performance of models in train data and test data.

*1) Model-1*

*a) Analysis on Train Data:* Accuracy: 80.50% Precision: 0.32 Recall: 0.17 F1-Score: 0.22

Confusion Matrix

| Actual | | Predicted | |
|---|---|---|---|
| | | Non Cyberbullying | Cyberbullying |
| | Non Cyberbullying | 10057 | 774 |
| | Cyberbullying | 1747 | 356 |

Table 4.1: Confusion matrix of model-1 train data

*b) Analysis on Test Data:* Accuracy: 81.29% Precision: 0.30 Recall: 0.14 F1-Score: 0.19

Confusion Matrix:

| Actual | | Predicted | |
|---|---|---|---|
| | | Non Cyberbullying | Cyberbullying |
| | Non Cyberbullying | 1134 | 74 |
| | Cyberbullying | 194 | 31 |

Table 4.2: Confusion matrix of model-1 test data

*2) Model-2*

*a) Analysis on Train Data:* Accuracy: 97% Precision: 0.97(Sexism), 0.97(Racism) Recall: 0.97(Sexism), 0.99(Racism) F1-Score: 0.97(Sexism), 0.98 (Racism)

Confusion Matrix

| Actual | | Predicted | | |
|---|---|---|---|---|
| | | None | Sexism | Racism |
| | None | 9473 | 271 | 184 |
| | Sexism | 222 | 8189 | 0 |
| | Racism | 27 | 1 | 5211 |

Table 4.3: Confusion matrix of model-2 train data

©IJRASET: All Rights are Reserved

*b) Analysis on Test Data:* Accuracy: 91.75% Precision: 0.89(Sexism), 0.90(Racism) Recall: 0.96(Sexism), 0.98(Racism) F1-Score: 0.92(Sexism), 0.94 (Racism)

Confusion Matrix

| Actual | Predicted | | | |
|---|---|---|---|---|
| | | None | Sexism | Racism |
| | None | 939 | 109 | 60 |
| | Sexism | 36 | 904 | 0 |
| | Racism | 9 | 2 | 561 |

Table 4.4: Confusion matrix of model-2 test data

*3) Model-3*

*a) Analysis on Train Data:* Accuracy: 84% Precision: 0.87 Recall: 0.51 F1-Score: 0.64

Confusion Matrix

| Actual | Predicted | | |
|---|---|---|---|
| | | Non Cyberbullying | Cyberbullying |
| | Non Cyberbullying | 89338 | 2764 |
| | Cyberbullying | 17931 | 18706 |

Table 4.5: Confusion matrix of model-3 train data

*b) Analysis on Test Data:* Accuracy: 84% Precision: 0.87 Recall: 0.51 F1-Score: 0.64

Confusion Matrix:

| Actual | Predicted | | |
|---|---|---|---|
| | | Non Cyberbullying | Cyberbullying |
| | Non Cyberbullying | 9866 | 306 |
| | Cyberbullying | 2037 | 2096 |

Table 4.6: Confusion matrix of model-3 test data

*4) Model – 4*

*a) Analysis on Train Data:* Accuracy: 98.12% Precision: 0.98 Recall: 0.73 F1-Score: 0.83

Confusion Matrix:

| Actual | Predicted | | |
|---|---|---|---|
| | | Non Cyberbullying | Cyberbullying |
| | Non Cyberbullying | 10440 | 11 |
| | Cyberbullying | 199 | 535 |

Table 4.7: Confusion matrix of model-4 train data

©IJRASET: All Rights are Reserved

*b) Analysis on Test Data:* Accuracy: 94.37% Precision: 0.66 Recall: 0.35 F1-Score: 0.45

Confusion Matrix:

| | | Predicted | |
|---|---|---|---|
| | | Non Cyberbullying | Cyberbullying |
| Actual | Non Cyberbullying | 1816 | 24 |
| | Cyberbullying | 87 | 47 |

Table 4.8: Confusion matrix of model-4 test data

These models perform well independently. At the point when joined together the presentation of model to some degree debases because of the various appropriation of information they were prepared on. They perform acceptably well in the majority of the cases. We tried creating an ensemble model, which is to prepare the yield of these models into another classifier. In any case, gathering model had just 60% exactness, along these lines we opted to do voting among the model outputs in-order to get the final output.

## VII. CONCLUSION

Taking everything into account, it is seen that utilizing pictures in cyberbullying discovery can upgrade the choice of AI calculations. By utilizing the hidden context and image-comment relation we can use certain valuable data that can help in location of cyberbullying. In any case, pictures without anyone else can't be effectively used to identify cyberbullying, henceforth it is hard to assemble a framework that forestalls digital cyberbullying before it even happens dependent on the questionable idea of pictures.
Most definitely, in our text based approach, we see that the presentation of models rely upon the nature of dataset that is utilized for preparing. Likewise, the class imbalance problem in a large number of these datasets causes the model to give skewed predictions. To defeat this oversampling can be utilized. Finally, we saw that utilizing ensembling procedures can debase by and large execution because of the differing idea of the preparation datasets. Models trained on datasets tend to overfit to those datasets and when ensemble model is prepared on their forecast, it performs ineffectively because of the random distribution of model predictions.
Our image based approach makes a few suppositions, which may not be material in reality situation. This issue can be overcome by utilizing a reinforcement learning algorithm. These are working particularly well in the area of language modeling. This can be an indication that they can also work well for classification problems such as this.

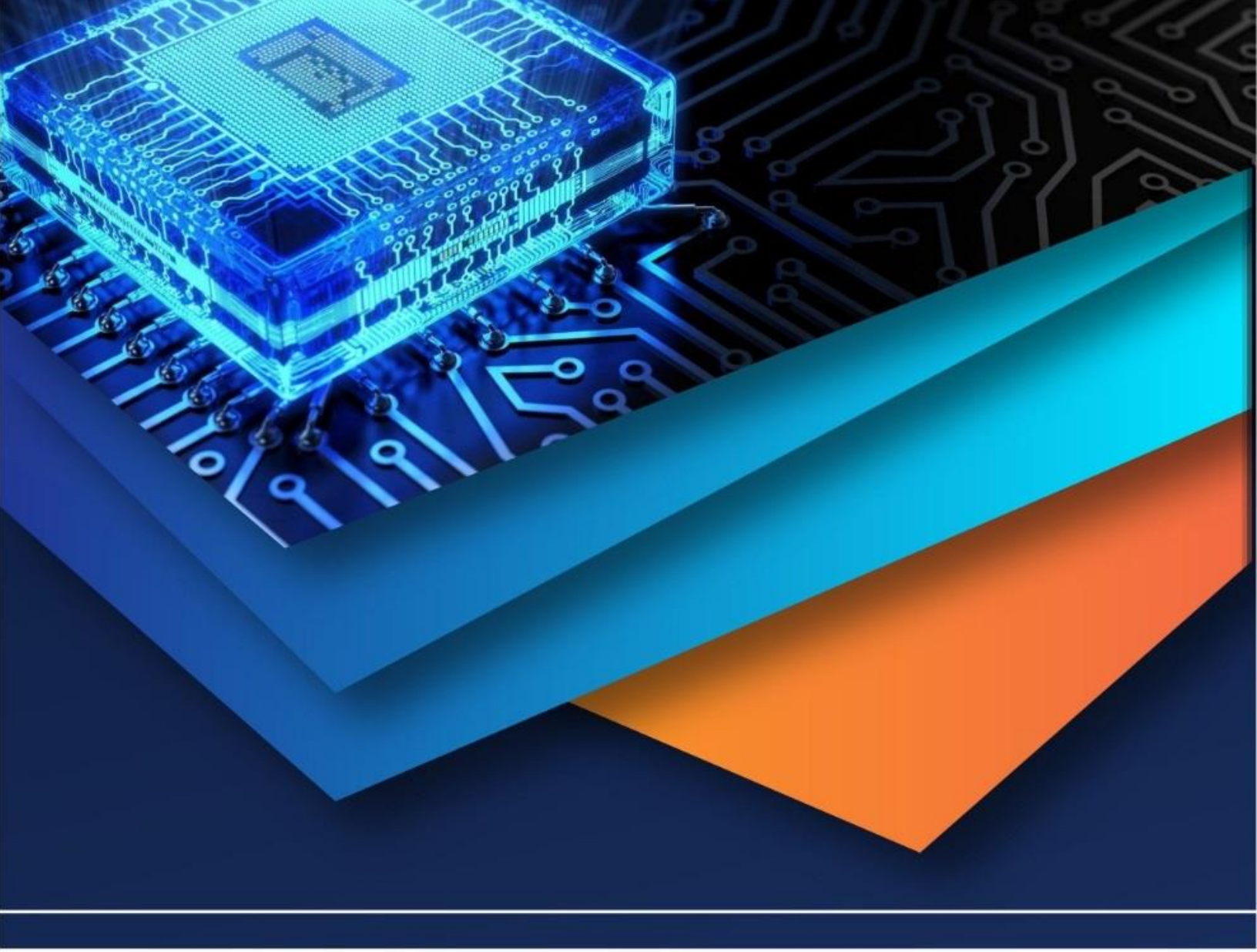
## VIII. ACKNOWDEGEMENT

The authors express their appreciation for the support provided by our mentors and faculty members who have guided us during the research and helped us achieve the desired results.

## REFERENCES

[1] Sinchana C, Sinchana K, Pradyumna C S, Deepika S, “Detection of Cyberbullying using Machine Learning”, International Journal of Advance Research, Ideas and Innovations in Technology, Volume 6, Issue 4, July 2020, ISSN: 2454-132X.
[2] John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad, Eslam Amer, “Social Media Cyberbullying Detection using Machine Learning”. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, No. 5, 2019.
[3] Kelly Reynolds, April Kontostathis and Lynne Edwards, “Using Machine Learning to Detect Cyberbullying”, International Journal of Advanced Computer Science and Applications (IJACSA) October 2019.
[4] Cynthia Van HeeI, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, VeÂronique Hoste, “Automatic detection of cyberbullying in social media text”. PLOS ONE https://doi.org/10.1371/journal.pone.0203794 October 8, 2018
[5] Ghada M. Abaido, “Cyberbullying on social media platforms among university students in the United Arab Emirates”. International Journal of Adolescence and Youth, 26 Sep 2019.
[6] Elizabeth Byrne and Lauren Pfeifer, " Cyberbullying and Social Media: Information and Interventions for School Nurses Working With Victims, Students, and Families", The Journal of School Nursing, 2018, Vol. 34(1)
[7] Mohammed Ali Al-Garadi, Mohammed Rashid Hussain, Nawsher Khan, Ghulam Murtaza, Henry Friday Nweke, Ihsan Ali, Ghulam Mujtaba, Haruna Chiroma, Hasan Ali Khattak and Abdullah Gani, “Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges”, IEEE, VOLUME 7, June 11, 2019.
[8] https://www.bullying.co.uk/cyberbullying/effects-of-cyberbullying/ (Effects of cyberbullying)
[9] https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa (WordEmbedding)
[10] https://machinelearningmastery.com/handle-long-sequences-long-short-term-mem ory-recurrent-neural-networks/ (How to Handle Very Long Sequences in LSTM neuralnetworks).

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)