ISSN No:-2456-2165

Cyber Barrier

[1]Aswathy P Shyju
Student, Computer science and Engineering Department
Sahrdaya College of Engineering & Technology
Thrissur, India

[2]Austin Larson
Student, Computer science and Engineering Department
Sahrdaya College of Engineering & Technology
Thrissur, India

[3]Femy P Joy
Student, Computer science and Engineering Department
Sahrdaya College of Engineering & Technology
Thrissur, India

[4]Gopika K S
Student, Computer science and Engineering Department
Sahrdaya College of Engineering & Technology
Thrissur, India

[5]Anusree K
Asst. Professor, Computer science and Engineering Department
Sahrdaya College of Engineering & Technology
Thrissur, India

Abstract:- Social media has become an integral part of our lives. It gives us the freedom to express ourselves and to communicate with people around the globe. But currently, the platform is being exploited for cyberbullying and personal harassment. Because of the increasing expansion of social media and its integration into ordinary living, cyberbullying has become extremely common. Being the victim of a cyberbully could have a severe emotional and psychological impact on an individual. It can make anyone feel vulnerable and exploited. One of the main challenges faced by cyberbullying detection is the lack of labeled data. Keeping this in mind, a Semi-supervised learning model is proposed to detect and prevent cyberbullying on social media platforms. This model uses partially labeled training data, with a small amount of labeled data and a larger amount of unlabeled data.

Keywords:- *Cyberbullying; Label Propagation; Natural Language Toolkit (NLTK); Semi-supervised learning; Social media; Sentiment analysis; Support Vector Machine (SVM)*

I. INTRODUCTION

With the rise in popularity of social media, a new trend of cyberbullying has emerged. It is defined as an offensive, calculated move made over time by a group of people via electronic forms of communication against a victim who is unable to defend themselves. Creating fabricated rumors and disclosing personal details to hurt and disgrace the victims. Cyberbullying can take multiple shapes, from text to photos and videos. The inability to comprehend cyberbullying practice heightens the potential risks of victims' invasions. Furthermore, provided the virtual world in which cyberbullies exist, the effects of cyberbullying are seemingly unlimited, since they may engage in these acts without consideration for time, apps, or sites. The victims' behaviors alter as well, affecting their sentiments, self-confidence, and insecurity. Cyberbullying includes not only the fabrication of a false identity and the posting of any degrading picture or video, as well as the circulation of negative information about an individual, but also intimidation. The repercussions of social media cyberbullying are devastating, some even led to the death of unfortunate victims.

In a nutshell, cyberbullying requires a comprehensive remedy. Cyberbullying must be eradicated. This problem can be handled by detecting and avoiding it. The key objective of our paper is to develop a semi-supervised learning model for detecting social media abuse so that no one has to experience it. The proposed methodology is implemented on a social media bullying dataset gathered from sources such as Twitter.

II. RELATED WORK

A. A hierarchical approach for timely cyberbullying detection

Nazar, D. Zois and M. Yao [1] "A hierarchical approach for timely cyberbullying detection," this paper introduces a hierarchical approach for timely cyberbullying detection. The approach consists of the following steps: First characterizes an individual message as aggressive or not by evaluating the optimum least number of informative features extracted from the message. Then the prior probability is updated and if the average Bayes risk is less than the expected cost of continuing to review the message then the reviewing the message will be stopped otherwise it will continue to review the next message. When the algorithm decides to finish examining messages, it uses the optimum decision rule to classify the session and, if the session is bullying, it issues a cyberbullying alarm. This approach operates in a continuous loop, for each session as new messages are posted and until a decision is reached. The approach was successfully evaluated on the Instagram dataset and the expected accuracy is 74%.

B. Identification and Classification of Cybercrimes using Text Mining Technique

S. Andleeb, R. Ahmed, Z. Ahmed, and M. Kanwal [2] "Identification and classification of cybercrimes using text mining technique," Framework provides promising results in classification and recognition of cyberbullying using data mining techniques by considering the detection of cyberbullying from the dataset. It mainly consists of two modules: Preprocessing and feature extraction, Classification, and recognition. The classifier is trained using features extracted from the first module for training data and testing data is used for making predictions for data containing cyberbullying activity. On the training dataset, the accuracy level is tested for a suitable level of satisfaction. Then the classifier is used to test the data whose labels are unknown, otherwise, the model is calibrated and the procedure is rehashed. It uses three types of features namely demographic, behavioral, and textual features to identify cyberbullying words. Two classifiers are used during the study (Bernoulli's nb and SVM), while the overall accuracy of SVM linear was considerably found good as compared to the previous study carried over the same dataset.

C. Real-Time Detection of Cyberbullying in Arabic Twitter Streams

D Mouheb, M. H. Abushamleh, M. H. Abushamleh, Z. A. Aghbari and I. Kamel [3] "Real-time detection of cyberbullying in Arabic Twitter streams," This proposes a system to detect cyberbullying in Arabic tweets in real-time by using Twitter API. It filters the posts based on the offensive words, classifies them based on the strength of the offensive words. We used Twitter as the main platform, Twitter API is used to retrieve tweets, retweets, direct messages, mentions, and detect posts in real-time. The proposed system is designed to Detect cyberbullying messages, retrieve information about them, and do proper actions right on the spot. The tweets are streamed in real-time using Twitter API. Tweets are first processed to remove noise and repetitive letters. Using the modules from the cyberbullying filter. Offensive messages are detected using the keywords previously collected. The detected messages are classified according to their assigned weights. The detected abusive posts are sent to a host server. That saves the post information (user id, post id, post, post type, weight) as JSON objects. The user application retrieves the abusive posts from the host server. Shows the post information (e.G., Username, post type, post, weight), where the text is colored based on the weight of the offensive post.

D. Detecting A Twitter Cyberbullying Using Machine Learning

R. R. Dalvi, S. Baliram Chavan and A. Halbe [4] "Detecting A Twitter Cyberbullying Using Machine Learning," Using Supervised Binary classification Machine Learning algorithms, the suggested methodology is used to detect and prevent cyberbullying on Twitter. For data preprocessing, the Natural Language Toolkit (NLTK) is used. The model is evaluated on both Support Vector Machine and Naive Bayes, also for feature extraction, using the TFIDF vectorizer. On a dataset gathered from multiple sources such as Kaggle, and Github, the SVM, and Naive Bayes are compared. The dataset is separated into training and testing once it has been preprocessed and feature extracted. As the results show, Support Vector Machine has great accuracy for detecting cyberbullying content of around 71.25%, which is better than Naive Bayes. The accuracy, recall, f-score, and precision of both SVM and Naive Bayes are calculated. SVM, surprisingly, outscored Naive Bayes in every aspect.

III. PROPOSED SYSTEM

The proposed system can be used to detect cyberbullying in social media. The main difference with previous research is that it uses a Semi-Supervised learning model. A Semi-supervised model is a bridge between Supervised Learning and Unsupervised Analysis. It is a machine learning methodology in which a little amount of labeled data is combined with a large amount of unlabeled data to train a model. It's about learning with both labeled and unlabeled data. We are implementing this method as Supervised learning is less practical in real-world scenarios and to make use of the cheap and abundantly available unlabeled data.

We collected the dataset from Twitter comments. Then it undergoes Exploratory data analysis. In data pre-processing, the categorical values are converted to numerical values for fast computation. Then the dataset is divided into training and testing data. Training data using SVM and Support vector machines (SVMs) are a set of semi-supervised learning methods used for classification, regression, and outlier detection.

IV. METHODOLOGY

We have collected a dataset from Twitter comments where the dataset is labeled as offensive and non-offensive. The collected dataset consists of 24783 and it is partitioned into 20620 offensive and 4163 Non-Offensive. It contains the ID of a person and their full comment, categorizing them as offensive or non-offensive.

A mixture of supervised and unsupervised analysis is used in semi-supervised learning. It is the method of learning labeled and unlabelled data. A support vector machine is a model used to analyze data and discover patterns in classification and regression analysis. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class.SVM is a mathematical function-based model that is used to simulate complicated, real-world issues.

A. Implementation

1. Load the dataset

2. EDA: Exploratory data analysis:

- Data analysis
- Checking for missing values
- Statistical data analysis
- Data visualization

3. Data preprocessing: Convert categorical values to numerical values for fast computation.

4. Divide data into feature and labels

5. Divide the dataset into training and testing data
Semi-supervised model

6. Apply the model SVM:
- Train data using SVM
- Test untrained data
- Classify the values as offensive or non-offensive

7. Calculate the accuracy of the model using a confusion matrix

Phase 1: EDA

With the use of summary statistics and graphical representations, exploratory data analysis refers to the crucial process of doing first investigations on data in order to uncover patterns, uncover anomalies, test hypotheses, and check assumptions.

Steps in EDA:
- Load the dataset
- Peek the dataset values
- Check the information of the dataset.
- Check for null values and replace the null value with 1.
- Check the information of the dataset
- Function to plot a bar chart

Phase 2: Machine learning algorithms: SVM

The second phase is machine learning, we use SVM (Support Vector Machine) to train the model.

SVM

A support vector machine is a model used to analyze data and find patterns in classification and regression analysis. When the data has exactly two classes, the support vector machine (SVM) is utilized.

The purpose of the SVM algorithm is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. A hyperplane is a name for the finest decision boundary.

The extreme points or vectors that assist in creating the hyperplane are chosen by SVM. Support vectors represent extreme examples, due to which the method is called a Support Vector Machine. Consider the diagram below, which shows how a decision boundary or hyperplane is used to classify two different groups.
SVM can be of two types:

Linear SVM: It is a classifier that is used for linearly separable data, which implies that if a dataset can be classified into two classes using a single straight line, it is called linearly separable data, and the classifier is named.

Non-linear SVM: It is used for non-linearly separated data, which implies that if a dataset can't be classed using a straight line, it's non-linear data, and the classifier employed is a Non-linear SVM classifier.

Label propagation

Label propagation is a variation of Semi-supervised graph inference algorithms. It is used for classification tasks and Kernel methods to project data into alternate dimensional spaces.

Label Propagation makes no changes to the raw similarity matrix created from the data. The algorithm iterates across a modified version of the original graph, computing the normalized graph Laplacian matrix to normalize the edge weights. Spectral clustering also uses this approach.

rbf (expf_0(−γ|x−y|2),γ>0)

Where γ is specified by keyword gamma.

The dataset is split into training and testing data. The input of the dataset is the message and the output is the label.The data is split into 4 variables namely x_train, x_test, y_train, y_test using train_test_split function. The input (message) is stored in x and output (label) is stored in y. It is again divided into two; 70% for training and 30% for testing.

In semi-supervised learning, both labeled and unlabeled data are given for training. Since our dataset is labeled, we remove some of the labels, and therefore unlabeled data is created.

During the divide and conquer method, the system splits the dataset into training and testing data. Some of the labeled data which is used for training is converted as unlabeled and the training data is stored as x_train_lab, x_train_unlab, y_train_lab, y_train_unlab. The system is trained using both labeled and unlabeled data, then it is tested using the remaining 30% test data and the result is obtained.

Label propagation is used for training and testing. TF IDF analyses and memorizes the text data and SVC (Support vector classifier) is used for classification, they are connected using a pipeline.

x_test is used for testing the trained model. Based on the output of SVC, the label (offensive or non_offensive) is stored in y_predict. An accuracy score and confusion matrix is generated by comparing obtained y_predict and original label (y_test).

Accuracy is calculated using the following equation;

$$\text{Acc} = (tp + tn)/(tp * tn)$$

where tp = True positive numbers
tn = True negative numbers

The block diagram of the proposed system is shown below

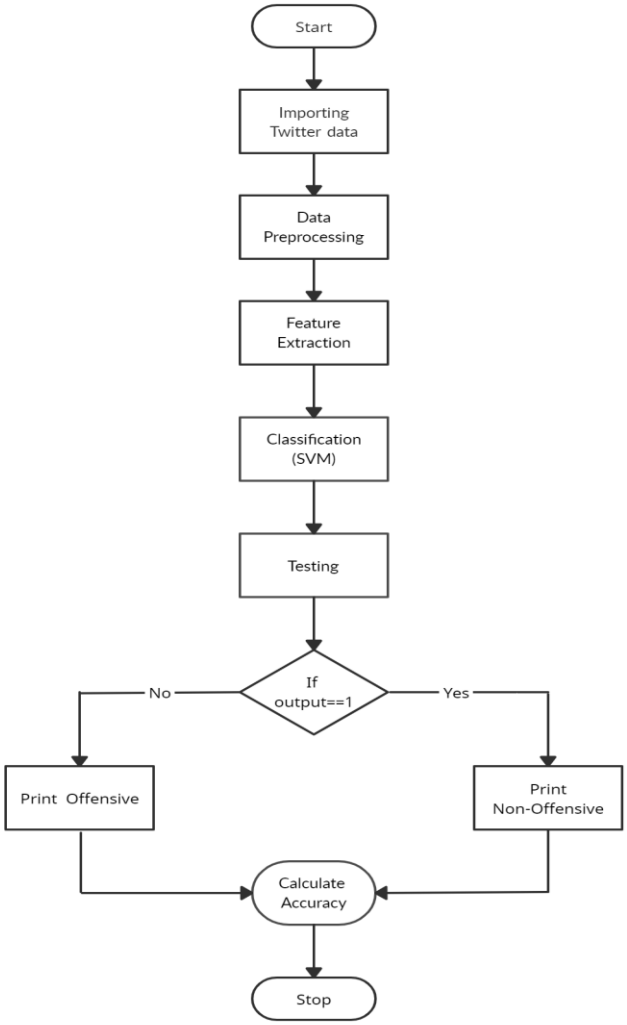

Fig. 1 Block Diagram

V. RESULT

Phase 1:

Visualization

- Length of the character in the Text

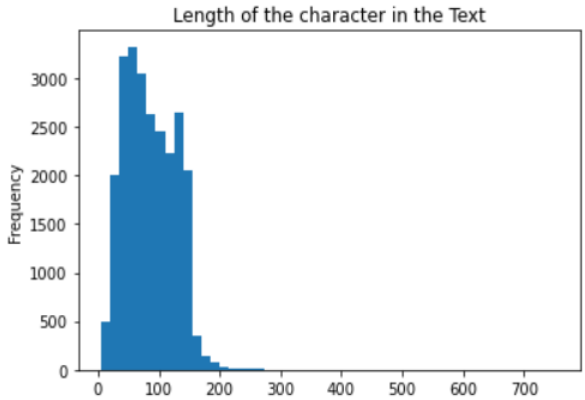

Fig.2 Length of the character in the Text

- Count of offensive and Non_offensive

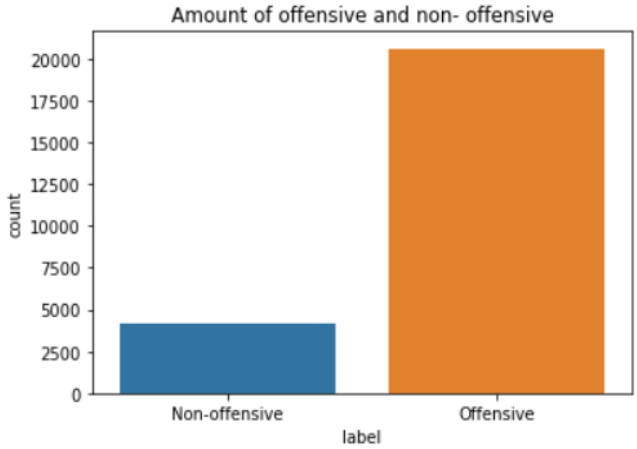

Fig. 3 Count of offensive and Non_offensive

Phase 2:

Semi-supervised learning using label propagation

- After training the system using label propagation we got an accuracy of 84%.
- The plot of the confusion matrix

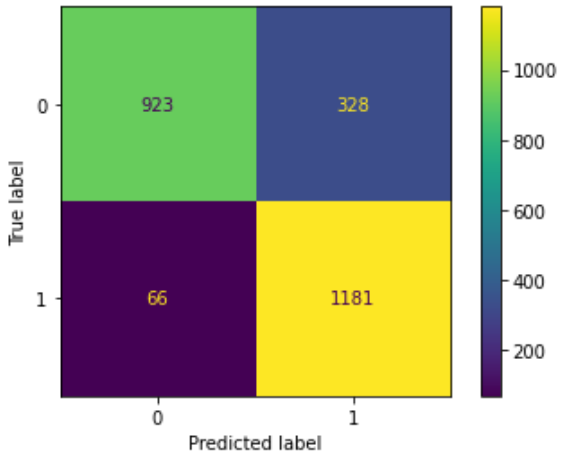

Fig.4 Plot of Confusion matrix

- Sample output is given below

```
[42] predict([
            "Hi How are you?",
            "Go to hell!",
            "You are useless!",
            "Thanks for the help"
     ], label_prop_model)

     array(['Non-offensive', 'Offensive', 'Offensive', 'Non-offensive'],
           dtype='<U13')
```

Fig. 5 Sample Output

VI. CONCLUSION

The tremendous advancement of technology is having an impact on how we communicate on social media platforms, leading to cyberbullying and personal harassment. Although many types of research addressed cyberbullying in social media platforms, the techniques used for the detection

prove to be inefficient in classification. Cyberbullying has become a severe problem in recent times. In the proposed system, we represent a new approach for the detection of offensive comments; a Semi-supervised learning model. It is implemented using a label propagation algorithm. The experiments were conducted on a dataset collected from Twitter comments. The TF IDF is used for analyzing and memorizing text data.SVM is used for training the system. Cyberbullying is a growing issue and it must end. And with our project, we are working towards this goal.

FUTURE WORK

Our project aims at the future scope that has the ability to train a system, which includes detection and removing the cyberbullying in each and every conversation and chat; which includes (English and Hindi) with many mixtures of codes in each language. In our studies, we had noticed that we could extend our system to many social media platforms in the future.

REFERENCES

[1]. I. Nazar, D. Zois and M. Yao, "A Hierarchical Approach for Timely Cyberbullying Detection," *2019 IEEE Data Science Workshop (DSW)*, Minneapolis, MN, USA, 2019, pp. 190-195, doi: 10.1109/DSW.2019.8755598.

[2]. S. Andleeb, R. Ahmed, Z. Ahmed and M. Kanwal, "Identification and Classification of Cybercrimes using Text Mining Technique," *2019 International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, 2019, pp. 227-2275, doi: 10.1109/FIT47737.2019.00050.

[3]. D. Mouheb, M. H. Abushamleh, M. H. Abushamleh, Z. A. Aghbari and I. Kamel, "Real-Time Detection of Cyberbullying in Arabic Twitter Streams," *2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, CANARY ISLANDS, Spain, 2019, pp. 1-5, doi: 10.1109/NTMS.2019.8763808.

[4]. R. R. Dalvi, S. Baliram Chavan and A. Halbe, "Detecting A Twitter Cyberbullying Using Machine Learning," *2020 4th International Conference on Intelligent Computing and Control Systems* (ICICCS), 2020, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893.