S.Y.B.Tech Students' Conference on EDAI 1 in between 16th and 17th December 2022,
Vishwakarma Institute of Technology, Pune, INDIA.

Cyberbullying Analysis in social media

Puja Chavan, Mrunmayee Phadke, Siddhi Patil, Samved Patil, Vaishnavi Pethkar

Department of Multidisciplinary Engineering
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

Abstract* — *As people spend more time utilising technology that keeps them constantly connected to other people, cyberbullying is becoming increasingly common. Cyberbullies can communicate with their victims in a variety of methods, including text messaging, social networking websites and instant messaging through the internet. Cyberbullying is a significant issue that, like traditional bullying, may make the victim feel inadequate and unduly self-conscious and even lead to suicidal thoughts. Fake news and fake messages are a huge threat to the community since it can lead to theft and commotion thus calling for the need of cybersecurity. It is paramount to detect and ensure safety across social platforms. This paper proposes an approach to detect cyberbullying. Out of the three classifiers used – decision tree (77.54%), naïve bayes (73.32%), SVM (82.12%), SVM being more accurate than others was used.

Keywords *— Cyberbullying, Machine learning, Support Vector Machine, NLP, StreamLit.*

I. INTRODUCTION

Bullying that occurs online, such as on computers, tablets, and mobile phones, is referred to as cyberbullying. Cyberbullying can happen online through social media, forums, or gaming where users can read, interact with, or exchange content. It can also happen through SMS, Text, and applications. Sending, posting, or disseminating unfavorable, hurtful, or malicious content about someone else is considered cyberbullying. It can also involve disclosing sensitive or private information about another individual in a way that causes embarrassment or humiliation.

Examples from actual incidents might help people better comprehend the strategies that are frequently employed because there are numerous ways that cyberbullying can occur. Bullying can raise the likelihood of behaviors associated to suicide when combined with other risk factors. Cyberbullying can also be persistent, raising the risk of anxiety and depression. Some states have made the decision to prosecute young people who harass others, including urging someone to commit suicide, as criminal harassment. Some types of cyberbullying involve harassment that crosses the line into crime, while some strategies are used in romantic relationships and have the potential to escalate to physical violence. Cyberbullying occasionally veers into illegal or criminal action. The following places are where cyberbullying is most frequent: social media, Text messaging, Instant messaging, direct messaging, Emails, etc. There are two sources of federally collected data on youth bullying:

- The 2019 School Crime Supplement to the National Crime Victimization Survey (National Centre for Education Statistics and Bureau of Justice) indicates that, nationwide, about 16 percent of students in grades 9–12 experienced cyberbullying.
- The 2019 Youth Risk Behaviour Surveillance System (Centres for Disease Control and Prevention) indicates that an estimated 15.7% of high school students were electronically bullied in the 12 months prior to the survey.

II. LITERATURE REVIEW

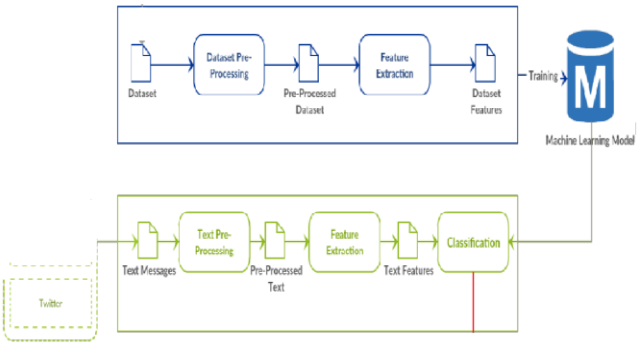
On earlier research, a binary classification task that separates cyberbullying from non-cyberbullying was used to identify cyberbullying in social media [3]. A SVM classifier technique was used in research by Divyashree et al. to identify cyberbullying posts on social media. The study trains the derived features from training phase input words using an SVM classification method. By analyzing user comments, it is possible to distinguish cyberbullying from testing phase input words. Lexical and syntactic data are retrieved from each user remark in order to classify if the comment is from a cyberbully or not [4]. In a study published in 2017, Ducharme et al. used a hybrid K-Nearest Neighbor/Support Vector Machine model to categorize remarks made in relation to cyberbullying. To ensure that the two classes were evenly distributed, 350 comments from the original data were sampled to create the training data for the study. Of these 350 comments, 175 were about bullying and 175 were not. According to the study, the KNNFilter method may cut training data volumes by over 50% while still producing a high-performing model. In comparison to the SVM model based on the complete training data, the majority of the hybrid models exhibited a cross validated accuracy between 70% and 80%. As a result, the hybrid model of the study seems to suggest that the strategy is effective even when using real-world data [5]. In order to find cyberbullying on the post social networking platform Instagram, Zhong et al. (2016) conducted research. The creation of Early Warning methods to recognize attack-vulnerable posted photographs was used in the study to detect cyberbullying. More than 3000 photographs from posts on the Instagram photo-sharing network, together with their corresponding comments, were used in the study's

methodology. In addition to conventional photos and text, the study makes use of novel features for topic classification derived from image description and a pretrained convolutional neural network of the pixel image. The primary conclusion drawn from the current findings is that adolescent internet users are unaware of the extent to which their online behaviors are engaged and how those behaviors affect the lives of others [9].

III. METHODOLOGY/EXPERIMENTAL

The project methodology is broken down into several main phases, some of which are as follows:
1. Data Collection
2. Data Pre-processing
3. Feature extraction
4. Training and Testing of Model
5. Model Deployment

Below is a flowchart of the proposed approach.



Two datasets were gathered throughout the data gathering process: one contained many tweets, while the other was divided into bullying and non-bullying categories. The two datasets were then combined to create a single dataset with an 80-20 train-test split.
Non-bullying, bullying, gender, ethnicity, religion, age, not-cyberbullying, and other cyberbullying are some of the parameters included in the data collection.

Techniques including lemmatization, stemming, RE functions, vectorization, and tokenization were used in the cleaning and data pre-processing phases.

- **Lemmatization:**

Lemmatization is the process of combining a word's several inflected forms into a single unit for analysis. Similar to stemming, lemmatization adds context to the words. As a result, it ties words with related meanings together.

- **Stemming:**

Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers. A stemming algorithm reduces the words "chocolates", "chocolatey", "Choco" to the root word, "chocolate" and "retrieval", "retrieved", "retrieves" reduce to the stem "retrieve". Stemming is an important part of the pipelining process in Natural language processing. The input to the stemmer is tokenized words.

- **Tokenization:**

Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

- **Vectorization:**

Without utilising loops, vectorization speeds up the Python code. Utilizing such a method can effectively reduce the amount of time that code needs to run. Different operations are carried out on vectors, such as the dot product of vectors, also known as the scalar product because it produces a single output, the outer product, which produces a square matrix with dimensions equal to the length X length of the vectors, and the element-wise multiplication, which produces elements with the same indexes while leaving the matrix's dimension unaltered.

Word clouds were employed for the data visualisation portion. By making the size of each word proportionate to its frequency, it is a visualisation technique that shows how often words occur in a given amount of text. The words are then placed in a group or word cloud.

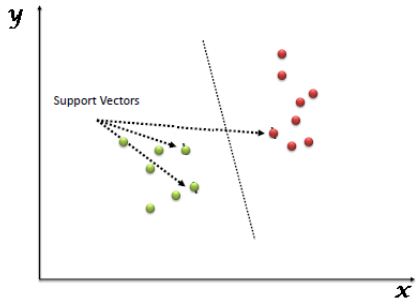
Decision trees, naive bayes, and SVM were the three models utilised for the classification (support vector machine). But out of the three, SVM was most accurate and gave an accuracy of (82.12). So, using SVM in final model deployment was the best choice out of three.

- **Support Vector Machine:**

Support vector machines are algorithms that find the best decision border between vectors that belong to a particular group (or category) and vectors that do not.

It is applicable to any type of vector that encodes any type of data. This means that in order to take advantage of the capabilities of SVM text categorization, texts must be converted into vectors.

With the SVM algorithm, each data point is represented as a point in n-dimensional space, with each feature's value being the value of a certain coordinate. Then, we carry out classification by identifying the hyper-plane that effectively distinguishes the two classes.



Support Vectors are simply the coordinates of individual observation. The SVM classifier is a frontier that best segregates the two classes (hyper-plane/ line).

For deployment of the model, app was made by using streamlit.

- **Decision Tree:**

A decision tree is a tree structure that looks like a flowchart, with an internal node representing a feature (or attribute), a branch representing a decision rule, and each leaf node representing the outcome. The root node is the node at the top of a decision tree. It learns to partition based on attribute value. It splits the tree recursively, which is known as recursive partitioning. This flowchart-like structure assists you in making decisions. It's a flowchart diagram-style depiction that easily replicates human level thinking. As a result, decision trees are simple to grasp and interpret.

Decision Tree is a form of white box ML method. It shares intrinsic decision-making logic that black box algorithms such as Neural Network do not have. Its training period is shorter than that of the neural network approach. The temporal complexity of decision trees is proportional to the amount of records and characteristics in the input data. The decision tree is a non-parametric or distribution-free strategy that does not rely on probability distribution assumptions. Decision trees are capable of handling high-dimensional data with excellent accuracy.

- **Naïve Bayes:**

A Naive Bayes algorithm is a probabilistic classification algorithm. It employs probability models based on strong independent assumptions. Because of separate assumptions, there is frequently no impact on reality. As a result, they are regarded as naive. Bayes' theorem can be used to generate probability models (credited to Thomas Bayes). Depending on the structure of the probability model, the Naive Bayes method can be trained in supervised learning.

A big cube with the following dimensions is used in Naive Bayes models:

1)The input field's name.
2)The value range might be continuous or discrete, depending on the type of input field. Continuous fields are split into discrete bins using the Naive Bayes technique.
3)The target field's value.

The Bayes theorem

Assume you defined a hypothesis based on your data.

The theorem will state the probability that the hypothesis will be true by multiplying the likely chances. In this manner, the theory will come true under specific conditions.

It then divides the product by the likelihood that the stated scenario will occur.
Pr(H|E) = Pr(H) * Pr(E|H) (E)

The hypothesis is that the document belongs to Categorical C because we are classifying papers. The presence of the word W is evidence.

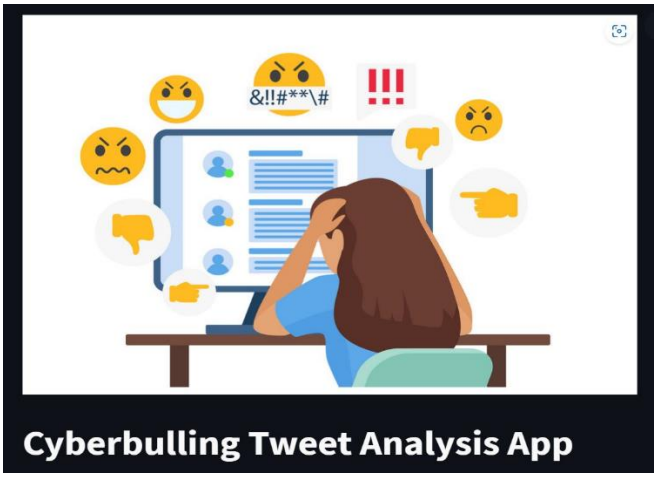
Because we are comparing two or more hypotheses, we may apply the ratio form of the Bayes theorem in classification tasks, which entails comparing the numerators within the formula (for Bayes aficionados: the prior times the probability) for each hypothesis:
Pr(C1|W) = Pr(C1) * Pr(W|C1) / Pr(C2) * Pr(W|C2)

Because of the enormous amount of words in a document, the formula is:

Pr(C1|W1, W2,...Wn) / Pr(C2|W1, W2,...Wn) =

Pr(C1) + (Pr(W1|C1) + Pr(W2|C1) +...Pr(Wn|C1)) /

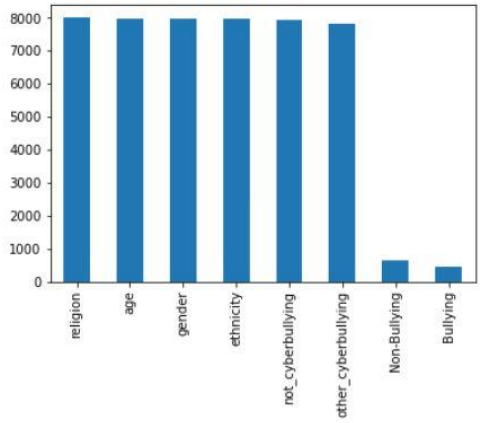
Pr(C2) + (Pr(W1|C2) + Pr(W2|C2) +...Pr(Wn|C2))



IV. Results and Discussions

On the cyberbullying dataset of tweets and other data set of bullying/non-bullying messages, the proposed approach was evaluated.

```
In [587]: #data visualization
          cb['text'].value_counts().plot(kind='bar')
Out[587]: <AxesSubplot:>
```



Following the processes outlined in the flowchart, extract the features after preprocessing the dataset. The dataset was then divided into ratios for train and test (0.8, 0.2). When evaluating the classifiers, accuracy, recall, and precision are taken into consideration, along with the f-score.

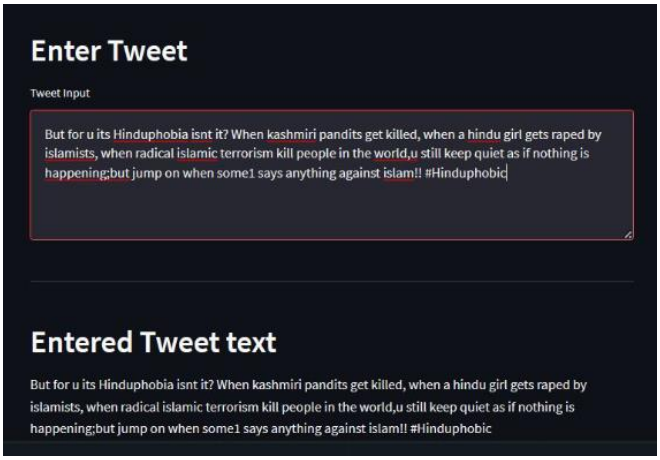
```
In [46]: # Model
         from sklearn.svm import SVC
         svm_model_linear = SVC(kernel= 'linear', C = 1).fit(X_train, y_train)
         svm_predictions  = svm_model_linear.predict(X_test)
         accuracy = svm_model_linear.score(X_test, y_test)
         print(accuracy)

         0.8212674323215751

In [49]: print(classification_report(y_test,svm_predictions))
         print(confusion_matrix(y_test,svm_predictions))
         print('Accuracy',accuracy_score(y_test,svm_predictions))
```

	precision	recall	f1-score	support
Bullying	0.53	0.37	0.43	90
Non-Bullying	0.60	0.22	0.32	123
age	0.92	0.99	0.95	1574
ethnicity	0.95	0.99	0.97	1639
gender	0.91	0.87	0.89	1631
not_cyberbullying	0.62	0.49	0.55	1588
other_cyberbullying	0.59	0.69	0.63	1521
religion	0.93	0.96	0.95	1586
accuracy			0.82	9752
macro avg	0.76	0.70	0.71	9752
weighted avg	0.82	0.82	0.82	9752

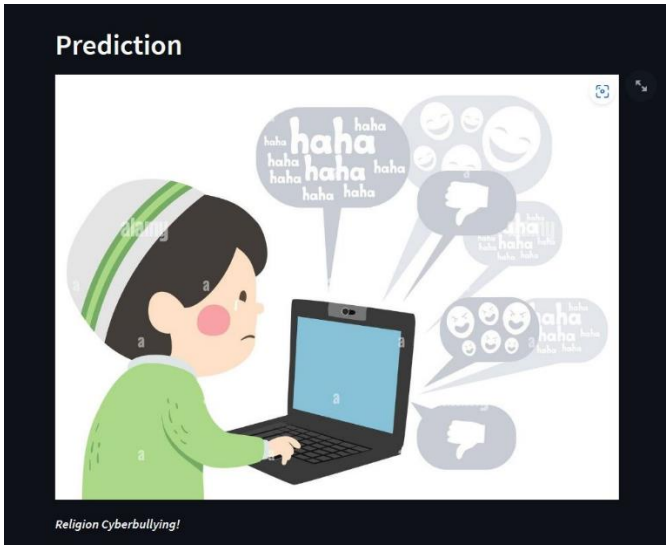
```
[[  33   11    1    7    8    4   23    3]
 [  15   27    2   11    8   12   46    2]
 [   0    0 1554    1    2    6   11    0]
 [   1    0    1 1621    4    4    8    0]
 [   5    1    8   14 1419   89   87    8]
 [   3    5   99   20   58  781  551   71]
 [   5    1   29   19   63  332 1047   25]
 [   0    0    3    9    5   32   10 1527]]
Accuracy 0.8212674323215751
```

The UI is an application through which the model was deployed. It predicts the type of the tweet into all of the stated categories as seen in the App, as was previously described. Here are a few samples of the interface, input, and text prediction. Since this text is a region-based tweet, the model correctly predicts that there would be discrimination based on religion, and the output displays this as such.

Input text:



Prediction:



V. CONCLUSION

We tried out three methods using machine learning to identify cyberbullying. But finally we used the SVM classifier to evaluate our model, as the accuracy was 82.12%.

Our model surpassed their classifiers in terms of accuracy and f-score when compared to another relevant piece of work, as well. Being able to recognize cyberbullying with this level of accuracy would undoubtedly help people use social media safely. The abundance of training data, however, restricts the ability to discover cyberbullying patterns. Therefore, more extensive data on cyberbullying is required to enhance the performance. Since deep learning techniques have been shown to perform better than machine learning methods on greater size data, they will therefore be appropriate with larger data.

ACKNOWLEDGMENT

We would like to thank our college Vishwakarma Instituteof Technology, Pune for giving us the opportunity to work on this project. We would like to thank our department. We would also like to thank our project guide Puja Cholke ma'am for constant support and guidance related to our project.

REFERENCES

[1] P. Galán-García, J.G. de la Puerta, C.L. Gómez, I. Santos, P.G. Bringas, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying," in Logic Journal of the IGPL.

[2] R.M. Kowalski, S. Limber, S.P. Limber, & P.W. Agatston, Cyberbullying: Bullying in the Digital Age, John Wiley & Sons, 2012. C.V. Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G.D. Pauw, W. Daelemans, V. Hoste, "Detection and Fine-Grained Classification of Cyberbullying Events," in RANLP (Recent Advances in Natural Language Processing)

[3] Divyashree, Vinutha H., Deepashree N. S., "An Effective Approach for Cyberbullying Detection and avoidance," in International Journal of Innovative Research in Computer and Communication Engineering.

[4] D. Ducharme, L. Costa, L. DiPippo, L. Hamel, "SVM Constraint Discovery using KNN applied to the Identification of Cyberbullying," The 13th International Conference on Data Mining.

[5] Michele Di Capua, Emanuel Di Nardo, Alfredo Petrosino, "Unsupervised Cyber Bullying Detection in Social Networks," in 23rd International Conference on Pattern Recognition (ICPR).

[6] V. Nahar, S. Al-Maskari, X. Li, C. Pang. "Semi-supervised Learning for Cyberbullying Detection in Social Networks," in ADC Databases Theory and Applications, pages 160-171, Springer International Publishing.

[7] AI-Enabled Cyberbullying-Free Online Social Networks in Smart Cities, Abdulsamad Al-Marghilani, 2022.

[8] Cyber Bullying Detection using NLP and Text Analytics, Yeo Khang Hsien, Zailan Arabee Abdul Salam, Vinothini Kasinathan,2022.

[9] An NLP-Assisted Bayesian Time Series Analysis for Prevalence of Twitter cyberbullying during COVID-19 pandemic, Christopher Perez,Sayar Karmarkar,2022.

[10] Analysing Cyberbullying using Natural Language Processing by Understanding Jargon in social media, Bhumika Bhatia, Anuj Verma, Anjum, Rahul Katarya,

[11] Crime Detection and Analysis from Social Media Messages Using Machine Learning and Natural Language Processing Technique, Xolani Lombo, Oyelade Olaide, Absalom El-Shamir Ezugw

[12] Linking textual and contextual features for intelligent cyberbullying detection in social media, Nabi Rezvani, Amin Beheshti, Alireza Tablordbar,2022.