DETECTION OF CYBERBULLYING USINGMACHINE LEARNING

[1] Dr. Marry Prabhakar, [2] Yenumula Koteswar, [3] Katakam Alekhya, [4]Bollepally Akshitha, [5]Athaluri Alekhya

[1] Associate Professor, Department of Information Technology, Vignan Institute of Technology and Science,
Hyderabad, Telangana, India.
[1]marryprabhakar@gmail.com

[2,3,4,5]Students, Department of Information Technology, Vignan Institute of Technology and Science,Hyderabad,
Telangana, India.
[2]koteswar1447@gmail.com, [3]alekhyakatakam01@gmail.com, [4]akshitha807@gmail.com,[5]alekhyaathaluri@gmail.com

Abstract

The advent of the digital age has enabled people to a new form of bullying which often results in social stigma. This new form of bullying is Cyber bullying which is a crime in which a perpetrator targets a person with online harassment and hate. Social networks provide a rich environment for bullies to find and harass vulnerable victims. Messages or comments concerning sensitive topics that are personal to an individual are more likely to be internalised by a victim, often ending in tragic outcomes. This phenomenon is creating a demand for automated, data-driventechniques for analysing and detecting such behaviour on the internet. In this project, a machine learning-based approach is proposed to detect cyber bullying activities from social network data. Naïve Bayes classifier is used to classify the type of message i.e., cyber bullying and non-cyber bullying message. Messages and take necessary actions. Our evaluation of performance results reveals that the accuracy of the proposed approach increases with more classification data.

Key Words: Cyber bullying, machine learning, feature extraction, text classification, Naïve Baye

I.INTRODUCTION

Millions of young people spend their time on social networking, and the sharing of information is online. Social networks have the ability to communicate and to share information with anyone, at any time, and in the number of people at the same time. There are over 3 billion social media users around the world. According to the National Crime Security Council (NCPC), cyber bullying is available online where mobile phones, video game apps, or any other way to send or send text, photos, or videos deliberately injure or embarrass another person. Cyber bullying can happen at any time all day, week and you can reach anyone anywhere via the internet. Text, photos, or videos of cyber bullying may be posted in an undisclosed manner. It can be difficult, and sometimes impossible, to track down the source of this post. It was also impossible to get rid of these messages later.

Several social media platforms such as Twitter, Instagram, Facebook, YouTube, Snapchat, Skype, and Wikipedia are the most common bullying sites on the internet. Some of the social networking sites, such as Facebook, and the provision of guidance on the prevention of bullying. It has a special section that explains how to report cyber-bullying and to prevent any blocking of the user. On Instagram, when someone shares photos and videos made by the user to be uncomfortable, so the user can monitor or block them. Users can also report a violation of our Community and make Recommendations to the app. As the social lifestyle exceeds the physical barrier of human interaction and contains unregulated contact with strangers, it is necessary to analyse and study the context of cyberbullying. Cyberbullying makes the victim feel that he is being attacked everywhere as the internet is just a click away. It can have mental, physical, and emotional effects on the victim. Cyberbullying mainly takes place in the form of text or images on social media. If bullying text can be distinguished from non-bullying text, then a system can act accordingly. An efficient cyberbullying detection system can be useful for social media websites and other messaging applications to counter such attacks and reduce the number of cyberbullying cases. The objective of the cyberbullying detection system is to identify the cyberbullying text and also take its meaning into consideration. One first analyzes the various aspects of a particular text and then applies the previous information or visuals to find the context of the text. There is a need to create a personalized system that can access such a text effectively and efficiently.

II. LITERATURE SURVEY

M. Di Capua, et al. [1] proposes an unsupervised approach to develop a cyberbullying model based on an amalgam of features, based on traditional textual features as well as some “social features”. The features were separated into 4 categories as Syntactic features, Semantic features, Sentiment features, and Social features. The results of this hybrid unsupervised methodology surpassed the previous results. The author then tested the YouTube dataset with 3 different Machine Learning Models: a Naive Bayes Classifier, Decision Tree Classifier (C4.5), and a Support Vector Machine (SVM) with a Linear Kernel. It was observed that clustering results for the hate posts turned out to have a lower precision in the youtube dataset when compared to the Form Spring tests, as textual analysis and syntactical features perform differently on both sides. When this hybrid approach was applied to the Twitter dataset, it resulted in a weak recall and F1 Score. The model proposed by the authors can be improved and used in building constructive applications to mitigate cyberbullying issues.

J. Yadav, et al.[2] proposes a new approach to cyberbullying detection in social media platforms by using the BERT model with a single linear neural network layer on top as a classifier. The model is trained and evaluated on the Form spring forum and Wikipedia dataset. The proposed model gave performance accuracy of 98% for the Form spring dataset and of 96% for the Wikipedia dataset which is relatively high from the previously used models. The proposed model gave better results for the Wikipedia dataset due to its large size g without the need for oversampling whereas the Form spring dataset needed oversampling R. R.

Dalvi, et al.[3] suggests a method to detect and prevent Internet exploitation on Twitter using Supervised classification Machine Learning algorithms. In this research, the live Twitter API is used to collect tweets and form datasets. The proposed model tests both Support Vector Machineand Naive Bayes on the collected datasets. To extract the feature, they have used the TFIDF vectorizer. The results show that the accuracy of the cyberbullying model based on the Support Vector Machine is almost 71.25% that is better than the Naive Bayes which was almost 52.75%.

Trana R.E., et al. [4] goal was to design a machine learning model to minimize special events involving text extracted from image memes. The author has compiled a database containing approximately 19,000 text views published on YouTube. This study discusses the effectiveness of the three machine learning machines, the Uninformed Bayes, the Support Vector Machine, and the convolutional neural network used on the YouTube database, and compares the results with the existing Form databases. The authors further investigated algorithms for Internet cyberbullying

following four categories: race, ethnicity, politics, and generalism. SVM has passed well with the inexperienced Naïve Bayes and CNN in the same gender group, and all three algorithms have shown equal performance with central body group accuracy. The results of this study provided data that can be used to distinguish between incidents of abuse and non-violence. Future work could focus on the creation of a two-part segregation scheme used to test the text extracted from images to see if the YouTube database provides a better context for aggression-related clusters.

N. Tsapatsoulis, et al. [5] a detailed review of cyberbullying on Twitter is presented. The importance of identifying different abusers on Twitter is given. In the paper, various practical steps required for the development of an effective and efficient application for cyberbullying detection are described thoroughly. The trends involved in the categorization and labeling of data platforms, machine learning models and feature types, and case studies that made use of such tools are explained. This paper will serve as an initial step for the project in Cyberbullying Detection using Machine learning.

G. A. León-Paredes et al.[6] have explained the development of a cyberbullying detection model using Natural Language Processing (NLP) and Machine Learning (ML). A Spanish cyberbullying Prevention System (SPC) was developed by applying machine learning techniques Naïve Bayes, Support Vector Machine, and Logistic Regression. The dataset used for this research was extracted from Twitter. The maximum accuracy of 93% was achieved with the help of three techniques used. The cases of cyberbullying detected with the help of this system presented an accuracy of 80% to 91% on average. Stemming and lemmatization techniques in NLP can be implemented to further increase the accuracy of the system. Such a model can also be implemented for detection in English and local languages if possible.

P. K. Roy, et al. [7] detail about creating a request for the discovery of hate speech on Twitter with the help of a deep convolutional neural network. Machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Gradient Boosting (GB), and K-nearby Neighbors (KNN) has been used to identify tweets related to hate speech on Twitter and features have been removed using the tf-idf process. The best ML model was SVM but it managed to predict 53% hate speech tweets in a 3: 1 dataset to test the train. The reason behind the low prediction scale was the unequal data. The model is based on the prediction of hate speech tweets. Advanced forms of learning based on the Convolutional Neural Network (CNN), Long-Term Memory (LSTM), and their Contextual LSTM (CLSTM) combinations have the same effects as a separate distributed database. 10-fold cross- validation was used along with the proposed DCNN model and obtained a very good recall rate. It was 0.88 for hate speech and 0.99 for non-hate speech. Test results confirmed that the k-fold cross- validation process is a better decision with unequal data. In the future, the current database can be expanded to achieve better accuracy.

S. M. Kargutkar, et al. [8] had proposed a system to give a double characterization for cyberbullying. The system uses Convolutional Neural Network (CNN) and Keras for content examination as the relevant strategies at that time provided a guideless view with less precision. This research involved data from Twitter and YouTube. CNN accuracy was 87%. In-depth learning-based models have found their way to identify digital harassment episodes, they can overcome the imprisonment of traditional models, and improve adoption.

III. EXISTING SYSTEM

Existing system uses traditional machine learning technique like SVM which includes the data pre-processing, regular expressions which has less accuracy to predict the cyber bullying in social media messages.

Disadvantages of existing system

- Feature extraction

- Traditional Model
- Social networking and chatting sites require automated detecting and processing methods.

IV. PROBLEM STATEMENT

Social networking and online chatting application provide a platform for any user to share knowledge and talent but few users take this platform to threaten users with cyberbullying attacks which cause issues in using these platforms. To provide a better platform for users to share knowledge on social networking sites there isa need for an effective detection system that can automate the process of cyberbullying detections and take decisions. Therefore, in this project, we you a dataset which would allow the users to build a predictive model to detect the message most accurately.

V. PROPOSED SYSTEM

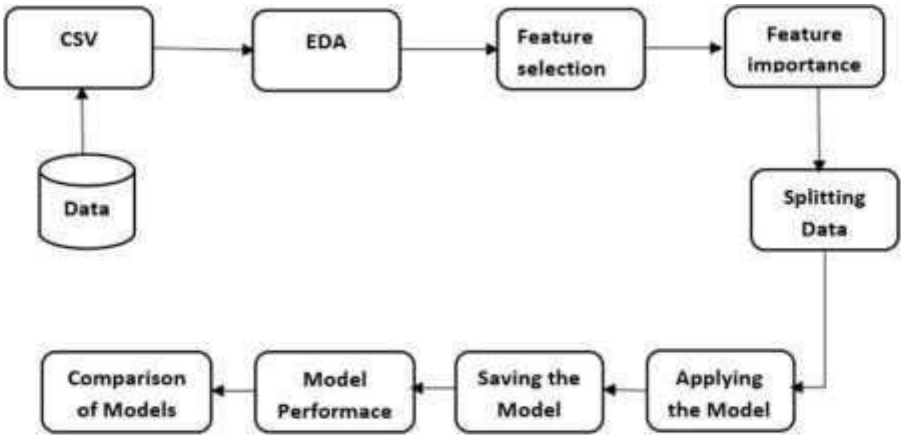
Our system is machine learning based and trained on a dataset from Kaggle. The model will classify the messages whether it belong to bullying or non-bullying. Machine learning approaches like Naïve Bayes which used conditional probability and Support vector classifier.

A.Features

- The latest machine learning models are used for training models that are accurate.
- Cyberbullying detection process is automatic and time taken for detection is less and it workson the live environment.

VI. SYSTEM ARCHITECTURE

System Architecture is a conceptual model that describes the structure and behaviour of multiplecomponents and subsystems.



System Architecture

VII.IMPLEMENTATION

Implementation is the stage where the theoretical design is turned into working system.The most crucial stage in achieving a new successful system with best accuracy and classifying the new images given by the user. The system can be implemented once after testing is done and found working according to the specification. It involves careful planning, investigation of the both current and existing system with implementation design of methods to achieve the change over an evaluation of change in method. Two major tasks are collecting the relevant data and training the model based on the data with model evaluation.

A. Process Data information

As social media usage becomes increasingly prevalent in every age group, a vast majority of citizens rely on this essential medium for day-to-day communication. Social media's ubiquity means that cyberbullying can effectively impact anyone at any time or anywhere, and the relative anonymity of the internet makes such personal attacks more difficult to stop than traditional bullying.

On April 15th, 2020, UNICEF issued a warning in response to the increased risk of cyberbullying during the COVID-19 pandemic due to widespread school closures, increased screen time, and decreased face-to-face social interaction. The statistics of cyberbullying are outright alarming: 36.5% of middle and high school students have felt cyber bullied and 87% have observed cyberbullying, with effects ranging from decreased academic performance to depression to suicidal thoughts.

- Data Source: Open Source
- Data Collected From: Kaggle
- Data Source Link: https://www.kaggle.com/datasets/andrewmvd/cyberbullying- classification
- Features: In light of all of this, this dataset contains more than 47000 tweets labelled according to the class of cyberbullying:
- Age
- Ethnicity
- Gender
- Religion
- Other type of cyberbullying
- Not cyberbullying
- The data has been balanced in order to contain ~8000 of each class.
- Trigger Warning These tweets either describe a bullying event or are the offense themselves, therefore explore it to the point where you feel comfortable.
- Create a multiclassification model to predict cyberbullying type;
- Create a binary classification model to flag potentially harmful tweets;
- Explore words and patterns associated with each type of cyberbullying.

B. Data Pre-Processing

Data pre-processing is the one of the important concepts to perform on each before trainingthe model.

C. Data cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabelled.

D. Data Normalization

Data normalization is a set of techniques and rules used to improve the consistency of data, standardize it, and maintain its integrity. The process removes any redundancies or duplicates in the database and reduces the storage space requirements.

E. Data augmentation

Data augmentation is a process of artificially increasing the amount of data by generating new data points from existing data. This includes adding minor alterations to data or using machine learning models to generate new data points in the latent space of original data to amplify the dataset.

F. Splitting the Data

Splitting the data into train and test based on the random on all images. Train data user sending to the model and test data is to use for the model performance. Always train data must be high because user training the model. Best way of splitting is 80-20 percentage or 70-30 percentage.

G. Model Applying

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0(Cyber bullying) or 1(Non- Cyber bullying), etc. Classes can be called as targets/labels or categories.

H. Saving the Model

User going to save the trained model in h5 files to save the weights and other parameters for the predictions.

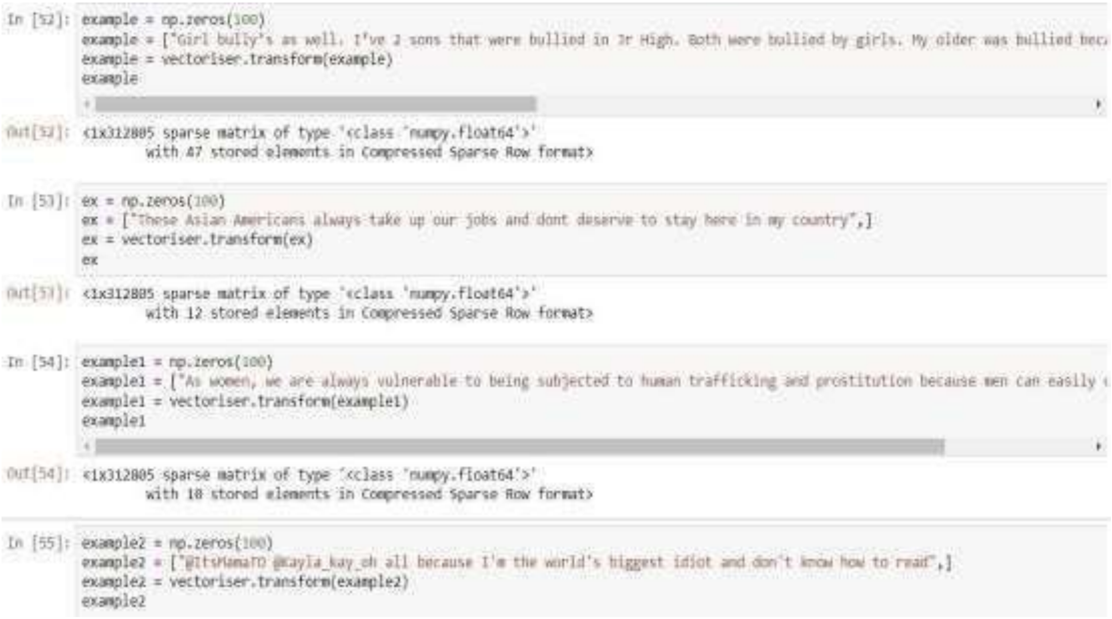
I. Model Performance

To calculate the model performance, user checking the type 1 and type 2 errors with TP, FP, FN, TN to create confusion matrix and calculating the accuracy, precision, recall and sensitivity with F beta score.

JPredictions

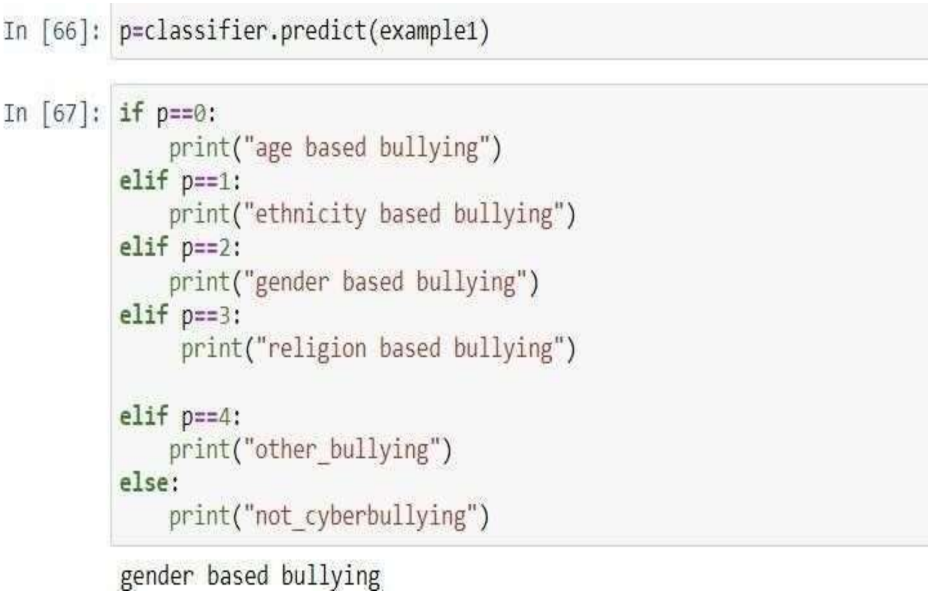
When user passes the new text into the saved model, model must extract the features of the new text and classify the predictions.

VIII.RESULTS



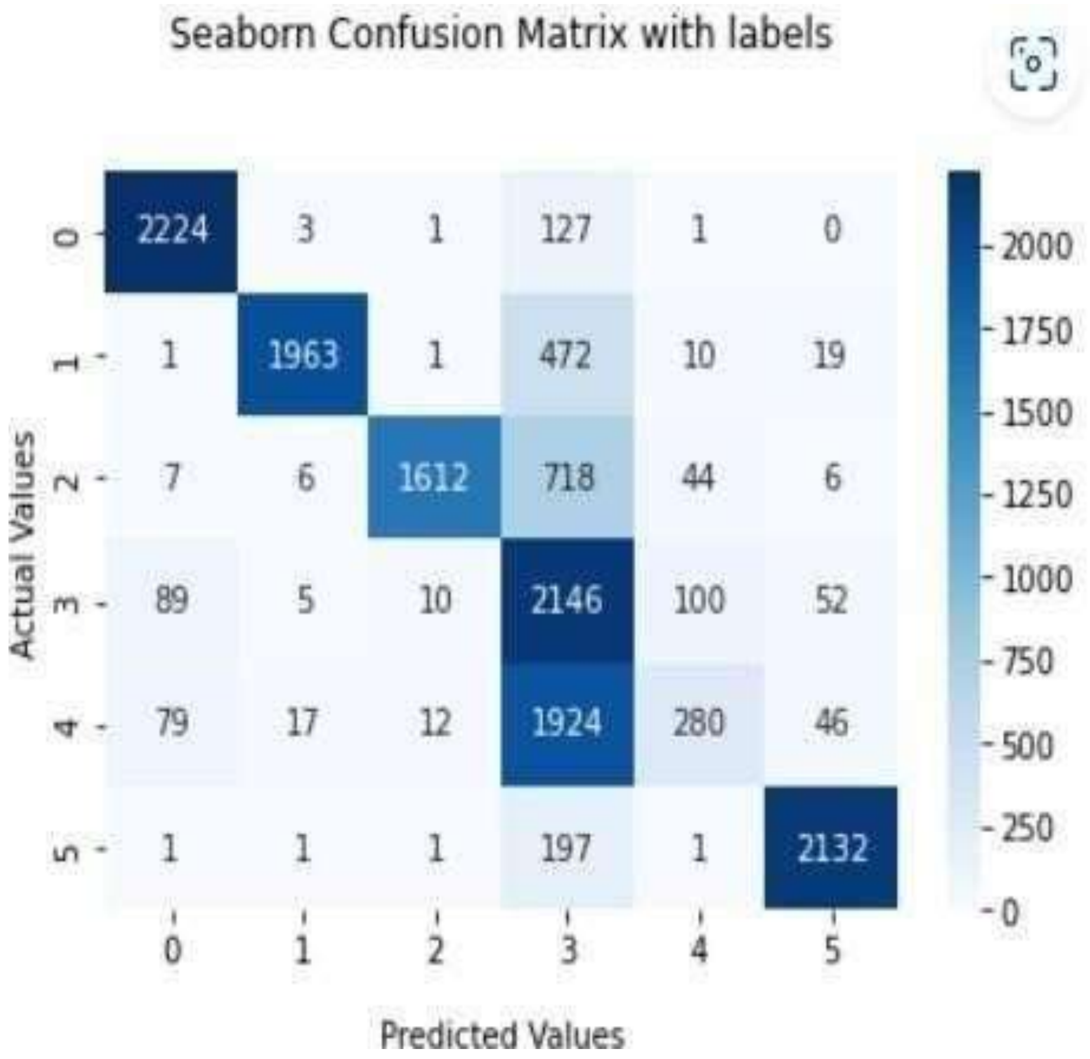
```
In [52]: example = np.zeros(100)
         example = ["Girl bully's as well. I've 2 sons that were bullied in Jr High. Both were bullied by girls. My older was bullied bec
         example = vectoriser.transform(example)
         example

Out[52]: <1x312805 sparse matrix of type '<class 'numpy.float64'>'
                 with 47 stored elements in Compressed Sparse Row format>

In [53]: ex = np.zeros(100)
         ex = ["These Asian Americans always take up our jobs and dont deserve to stay here in my country",]
         ex = vectoriser.transform(ex)
         ex

Out[53]: <1x312805 sparse matrix of type '<class 'numpy.float64'>'
                 with 12 stored elements in Compressed Sparse Row format>

In [54]: example1 = np.zeros(100)
         example1 = ["As women, we are always vulnerable to being subjected to human trafficking and prostitution because men can easily
         example1 = vectoriser.transform(example1)
         example1

Out[54]: <1x312805 sparse matrix of type '<class 'numpy.float64'>'
                 with 18 stored elements in Compressed Sparse Row format>

In [55]: example2 = np.zeros(100)
         example2 = ["@ItsMamaTO @Kayla_kay_oh all because I'm the world's biggest idiot and don't know how to read",]
         example2 = vectoriser.transform(example2)
         example2
```

Sample Output

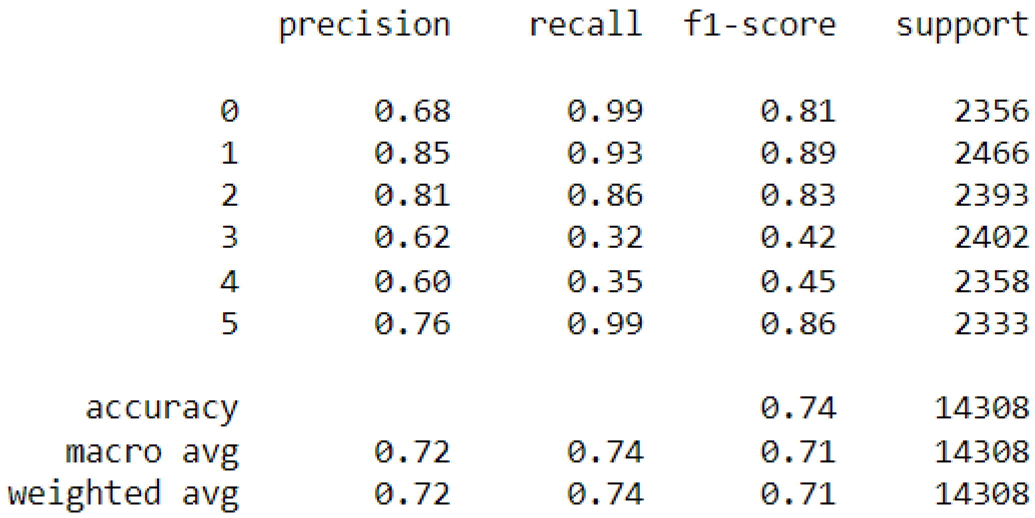


```
In [66]: p=classifier.predict(example1)

In [67]: if p==0:
             print("age based bullying")
         elif p==1:
             print("ethnicity based bullying")
         elif p==2:
             print("gender based bullying")
         elif p==3:
              print("religion based bullying")

         elif p==4:
             print("other_bullying")
         else:
             print("not_cyberbullying")

         gender based bullying
```

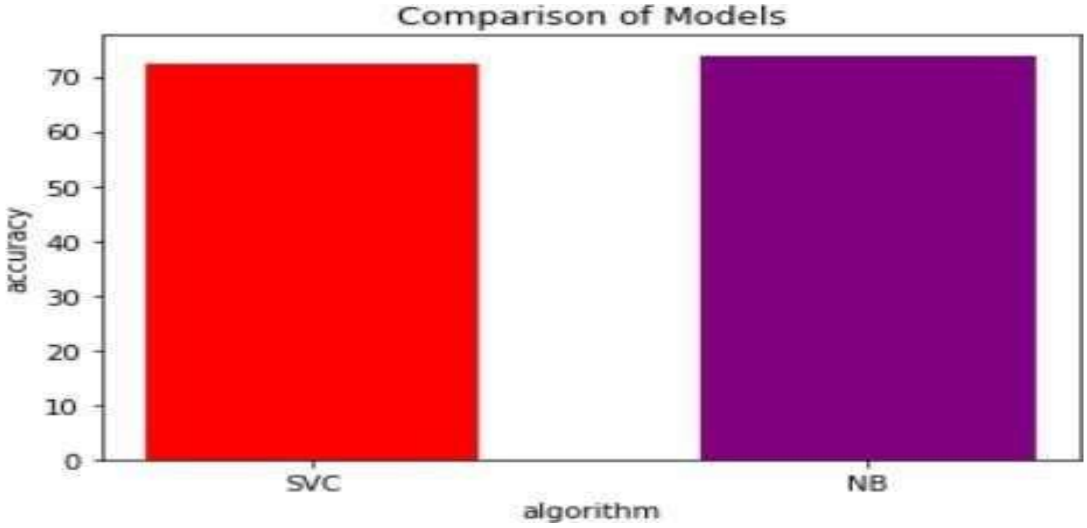
Sample Output



Confusion Matrix-SVC


```
              precision    recall  f1-score   support

           0       0.68      0.99      0.81      2356
           1       0.85      0.93      0.89      2466
           2       0.81      0.86      0.83      2393
           3       0.62      0.32      0.42      2402
           4       0.60      0.35      0.45      2358
           5       0.76      0.99      0.86      2333

    accuracy                           0.74     14308
   macro avg       0.72      0.74      0.71     14308
weighted avg       0.72      0.74      0.71     14308
```

Classification Report for Multinomial NB Classifier



Comparison of classification algorithms for Cyber bullying

IX CONCLUSION

We proposed a Machine Learning supervised approach in detecting cyberbullying based on the five features that can be used to define a cyberbullying or non-Cyberbullying message using theClassification model. While considering the features the classification models achieved 72-74% accuracy when trained the traditional machine learning models. The classification model can achieve more accurate results if provided with a large dataset. We can try to achieve even better results in the cyberbullying detection process if we consider more features in the project. Based on all the features an application can be created to detect the bullying traces and thus help in detecting and reporting such posts. Feature extraction and attribute selection may also increase the accuracy of the model and by giving exact data set to train the model will be giving the higher range of accuracy for the model. In this model we focused on detecting if we try to train the model so that it can identity text when given input then we can apply to real time applications in the future.

REFERENCES

[1]M. Di Capua, E. Di Nardo and A. Petrosino, Unsupervised cyberbullying detection in socialnetworks, ICPR, pp. 432-437, doi: 10.1109/ICPR.2016.7899672. (2016)

[2] J. Yadav, D. Kumar and D. Chauhan, Cyberbullying Detection using Pre-Trained BERT Model,ICESC, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700. (2020)

[3] R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893. (2020)

[4] Trana R.E., Gomez C.E., Adler R.F. (2021) Fighting Cyberbullying: An Analysis of AlgorithmsUsed to Detect Harassing Text Found on YouTube. In: Ahram T. (eds) Advances in Artificial Intelligence, Software and Systems Engineering. AHFE 2020. Advances in Intelligent Systems andComputing, vol 1213. Springer, Cham. https://doi.org/10.1007/978-3-030-51328-3_2. (2020)

[5] N. Tsapatsoulis and V. Anastasopoulou, Cyberbullies in Twitter: A focused review, SMAP, pp.1-6, doi:

10.1109/SMAP.2019.8864918. (2019)

[6]G. A. León-Paredes et al., Presumptive Detection of Cyberbullying on Twitter through NaturalLanguage Processing and Machine Learning in the Spanish Language, CHILECON pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684. (2019)

[7] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, A Framework for Hate Speech DetectionUsing Deep Convolutional Neural Network, in IEEE Access, vol. 8, pp. 204951-204962,, doi: 10.1109/ACCESS.2020.3037073. (2020)

[8] S. M. Kargutkar and V. Chitre, A Study of Cyberbullying Detection Using Machine LearningTechniques, ICCMC, pp. 734-739, doi:10.1109/ICCMC48092.2020.ICCMC-00013 7. (2020) [9]Jamil, H. and R. Breckenridge. Greenship: a social networking system for combating cyber- bullying and defending personal reputation., ACM : n. pag. (2018)