1st International Conference on Women Development in Engineering, Science & Technology (WD-EST'23)

03-05 October 2023

Machine Learning Approach for Cyber Bullying Detection on Social Media Platform

Saba Yousha[1,*], Sajjad Ali Memon[2] and Shahnawaz Talpur[3]

[1] *Department of Computer Information & Engineering, Mehran University of Engineering and Technology Jamshoro Sindh, Pakistan*

[2] *Department of Telecommunication Engineering, Mehran University of Engineering and Technology Jamshoro Sindh, Pakistan*

[3]*Department of Computer System Engineering, Mehran University of Engineering and Technology Jamshoro, Sindh, Pakistan*

[*]Corresponding author email: sabayusha@gmail.com

ABSTRACT

The Internet and Electronic Media have taken over the world and Social media platforms are now among the most popular way to communicate. However some users harm these platforms and "Cyber bullying" is the major problem in this regard. Cyber Bullying is the type of bullying that uses technology to disrespect and hurt others. To combat this issue numerous researchers have prevented ways and methods but still detection is needed to overcome this menace. This study attempts to highlight an approach to detect Cyber bullying on Social Media platform. The results reveal that SVM classifier is better than other classifiers.

Keywords:

Cyber bullying, Support Vector Machine (SVM), Machine Learning, Social Media Platforms

1. Introduction

Social Media is a Platform that allows individuals to share their images, videos and daily updates as well as to connect and interact with their friends online [1]. Nowadays, Social Media is used by almost everyone. Social Networking has advantages, but it also has drawbacks. One of those aspects that need to be solved is "Cyber bullying". Cyber bullying is becoming a topic of Concern, particularly among teenagers. Cyber Bullying is defined as the "intentional, persistent and hostile usage of Information Technology to affect or damaged other individual" [2]. Cyber Bullying is a sort of harassment or bullying that occur online. It include disclosing private information about someone to make them feel ashamed, angry and destructive internet remarks or posts. One serious issue that is pervasive over the world is "Cyber Bullying". Bullied individuals experience depression, engaged in self-harm and in the worst case they commit suicide. Cyber Bullying is a serious issue that has to be addressed. Despite the fact that many scholars and researchers have proposed machine learning algorithms, they typically did not take Social Media into an account of detecting it [2]. It is

very challenging to defend community from the alarming surge in cyber Crime. In this era of internet where people actually live on online and digital platforms. Youngsters are the main target of Cyber Bullying according to the survey. Sending or uploading harsh or insulting comments with the purpose to harm someone's character or posting an in appropriate photograph or video to disrespect other [9]. Cyber Criminals utilize Social media as a platform to perpetrate various type of data breaches, Such as Cyber Bullying by assessing the information. Online Communication is just mistreatment carried out via phone or internet. It is global issue that is rapidly expanding. One of the best places of the people of all background and cultures to publically declare themselves is on Twitter. The latest content may be shared and thoughts and ideas exchanged through Social Media [13]. Our method for identifying bullying on Social media platform for detecting cyber bullying uses a sample of tweets. According to the results hash tags are useful in recognizing in Cyber Bullying and by (Tweets) written and behavioral features improves detection performance. This is crucial area of research that needs to be concentrated on because bullies frequently uses Twitter to insult other. Thus, Bullying on Social Media must therefore be handled carefully and take into consideration.

Several studies have proposed methods to identify Cyber bullying conduct a poll to examine effort made to detect and stop Cyber Bullying. Carried out Research to determine Cyber Bullying and Cyber aggression on Social Media. According to a study, approximately 20% of students admitted to being Cyber Bullied. Suicidal behavior and a risk in a depressive symptoms have been both associated to Cyber Bullying [1]. Do not expose your home address and Cell Phone number, move fast this advises a number of authors [3]. Utilize the Twitter data set to evaluate a proposed Algorithm. The algorithm aims to find the bully and the target of the bullying .Outlined a strategy for locating and classifying abusive language on Social Media. Implemented in-depth understanding to discover Cyber bullying on Twitter and developed an approach for preventing disrespectful and insulting conduct on Twitter. According to a survey, the victims of Cyber Bullying rarely tells their elders when it happens. The experts suggests keeping your privacy setting on Social Network at a top standard and avoiding engaging with strangers [1].

The Literature review on the most current studies and the development of the various strategies for cybercrime detection and prevention is being given. Presented their research in order to analyze how users behave while interacting or publishing for everyone to see [4].Proposed a method that combines transfer learning and message categorization to track

Cyber Bullying. A method for locating and classifying objectionable words on Social Media was proposed. Described a multilingual approach for the recognition of Cyber Bullying [1].There are numerous studies that focus on detecting Cyber Bullying using Machine Learning .A bag-of-words strategy was used to developed a Supervised Machine Learning System that can assess a statement's emotion and Contextual factors [7].Had to use a detection graph model and assessment algorithm to separate the multiple criminals and victims in order to find and rate the victim and attackers. They plan to go further in future consider hidden bullying and advanced trends. [9]. Today users may communicate with each other and keep up relationships even without ever meeting them Social Media has ingrained itself firmly into our daily lives. This technology is used as a mean of communicating and as a source of current data by more than a billion people. The Research primarily concentrate on mentally addressing the problem with the support of statistical information and preventative suggestion [13]. In this study we employed a different method to detect Cyber Bullying. We developed a technique for tracking Twitter Cyber Bullying. The proposed methodology is an area model that makes use of Tweets features including arrangement, actions, users and content to generate an AI classifier for labeling tweets as Cyber Bullying or Non Cyber Bullying [16]. This research aims to identify Cyber Bullying in posts from Twitter. Provided a away to categorize tweets as "bullying" using SVM supervised Machine Learning method .Analyzed people's behaviors using Machine Learning Algorithm.

2. Materials and Method

Information on the adopted methodology is provided in this Section. Getting to the results achieved that is displayed in the Fig. 1. It illustrates how Cyber Bullying was found in the collected data set.

Search terms like "Cyber Bullying", "Datasets", "Social Media" etc were employed to type these data. Data was first gathered, and then some features that needed the data to be in its basic context were extracted. Similarity measure profanity and other indicators were collected from Twitter in order to identify Cyber Bullying. By the use of a Twint Scraping tool, the data may be retrieved using the Twitter API and data is extracted from a Twitter account. The Primary objective of tweet extraction is to extract structured data from unstructured.

2.1. Data Acquisition

It is also known as DAQ or DAS is the process of taking signals that measure actual physical occurrences in the real world and digitalizing them so that a computer and software may alter them. In this we gathered data from output response in physical form then electronically

converting it into an equivalent form with no information lost and providing it to the collectors. The Transformation of analogue Signals obtained from various sensor into digital data that a computer can analyze is a crucial stage in the data collecting process. Incoming analogue material should be measured at discrete time intervals and reduced to one of a set of predefined values because it need to be saved in the computer's memory as discrete point of information represented by binary integers. Most often, a Digital-to-Analogue (D/A) conversion component on a DAQ card within the PC or attach to it via a Universal Serial Bus (USB) port is used to do this.

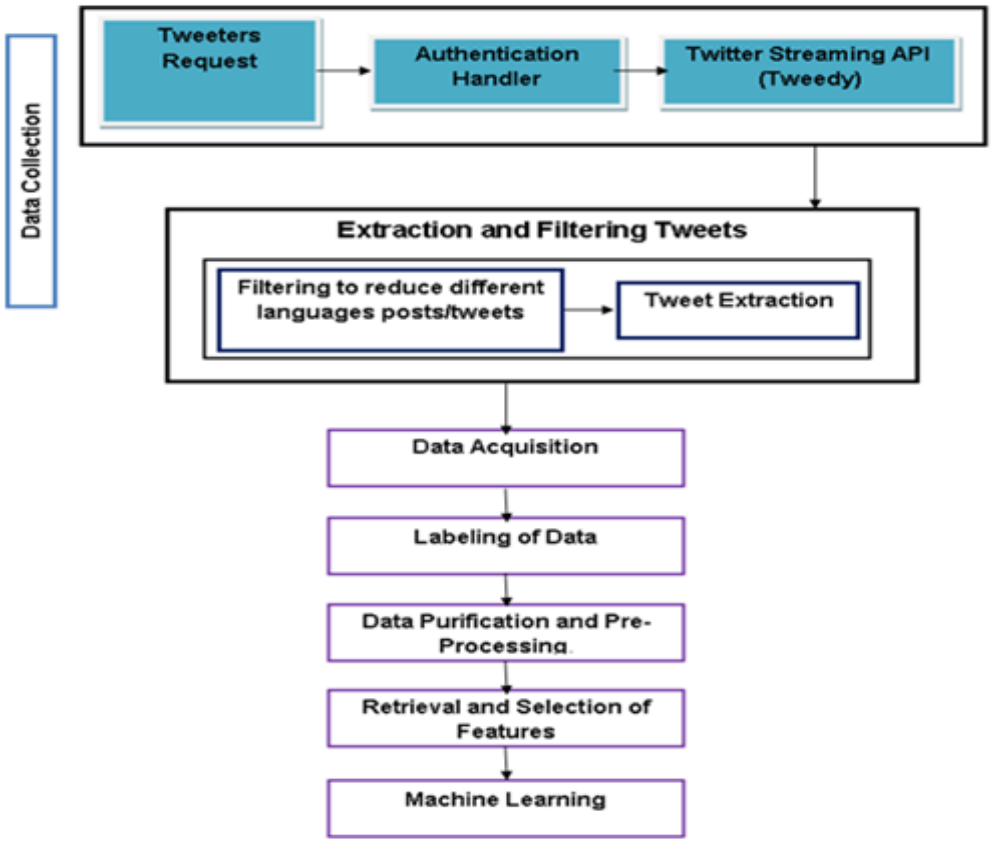

Fig.1: Flow chart of Methodology

2.2. Labeling of Data

A Tweet is a brief text message that can only be 140 characters long and it is posted on Twitter, the most popular mini platform. Hash tags Such as, "#win", "#contest", "#Giveaway" commonly indicated the contents of a Tweets. User names Such as, "@office", "@tech", "@Miss" are represented by words beginning with "@". Labeling of Data means a Collection of sample that have been marked with one or more labels is referred to as labeled data. A set of un label data is often supplemented with descriptive tags by labeling. For

instance, a data label might describe the topic of news story, the general tone of a tweets, the word use in group discussion, whether a photo featured a horse or a cow and what type of action is being performed. Data Labeling is an essential component of data pretreatment for Machine Learning especially for supervised learning, where input and output data were both classified and labeled to serve as a learning foundation for subsequent data processing. To lay the ground work for dependable learning patterns in Machine Learning, enormous volume of data are frequently needed. They must label or mark the data they utilize to guide learning based on the data attributes that support the model in arranging the data into structures that yield the intended result. In this step we label the data that we receive from the output.

2.3. Data Purification and Pre Processing

The translation of the Raw material into a comprehensible format is known as data preparation. Data preprocessing is an important step in data mining to increase data effectiveness. Any Analytical algorithm's results are directly impacted by the data preprocessing techniques used. In this we purify the data by following these steps first we do the quality assessment, then cleaning the data, after it we transform the data and at last reduction of data. Putting in missing quantities, normalizing or eliminating noisy data and outliers and intend to find are all part of this risk. The data cleaning procedures uncover and removes any flaws and irregularities in the data, hence enhancing its qualities. Data entry errors, Incorrect numbers and other type of inaccurate data can cause problem with the data quality. The process basically turns "Dirty" data into clean data. The results of Dirty data are not reliable or exact. Data that is rubbish produces garbage. Therefore, how this data is handled become significant. This approach is recommended for records when the greatest quantity of missing data makes the information invalid. We proposed that it is a time to expand the idea of data cleaning to meet the demands of contemporary Machine Learning. We detect relationship between the dataset treatment techniques and we suggest Machine Learning Clean, a comprehensive data purification models that combines the method and facilitates the creation of accurate and impartial systems.

2.4. Data Analysis

Class count chart in Fig. 2 shows the shows count of each class in the dataset the dataset is extremely balanced with almost same amount of instances for each class. Utilizing the metrics recall, precision, F-measure, accuracy, and specificity, the assessment of online harassment Detection Using SVM The classification algorithm is carried out on dataset collected from Twitter in this part. The technique part provides a description of the dataset

being used and the data annotation techniques for cyber bullying based on machine learning, namely SVM. This model was chosen from the most advanced social media platforms for detecting cyber bullying. The original articles' considered standard models' initial settings for parameters are utilized. But Python 3.7.4 was employed for the tests. Import, NLTK, Pandas, Tweepy, SK-Learn, and other necessary libraries were used in implementation and research settings. On a personal computer with an Intel Core-*i*5 CPU, Windows 10, and 8 Gigabytes of RAM, the experimental assessments are conducted. Using the NLTK Python library, the Training and testing datasets are created from the input dataset. Additionally, it is divided into three separate situations for the evaluation: accuracy (83.26%), precision (83.66%), and recall (83.5%). The metrics used for evaluation are designed to show each approach's tweet categorization efficiency at its peak. In addition to using cross-validation, the employed method is performed to get the standard deviation of this assessment measure.

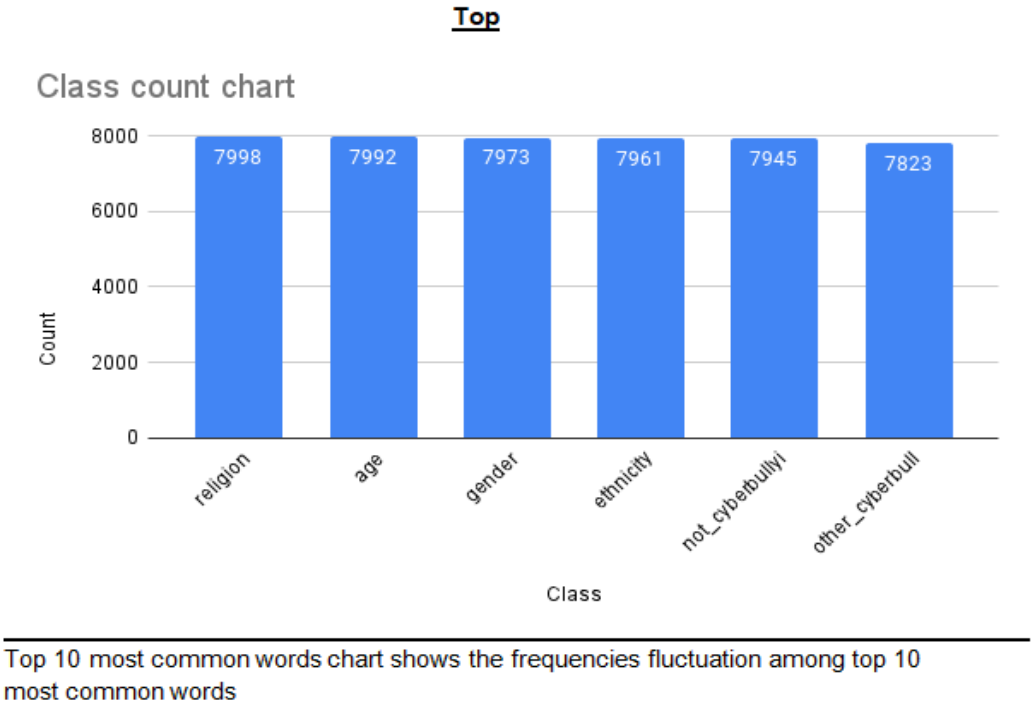

Fig. 2: Cyber Bullying according to Classes

Word count chart (Fig. 4) shows the count of the most common words in the dataset.

Figs. 5(a-c) shows the unigram, bigram and trigram frequency fluctuation respectively based on the count between different top 10 grams.

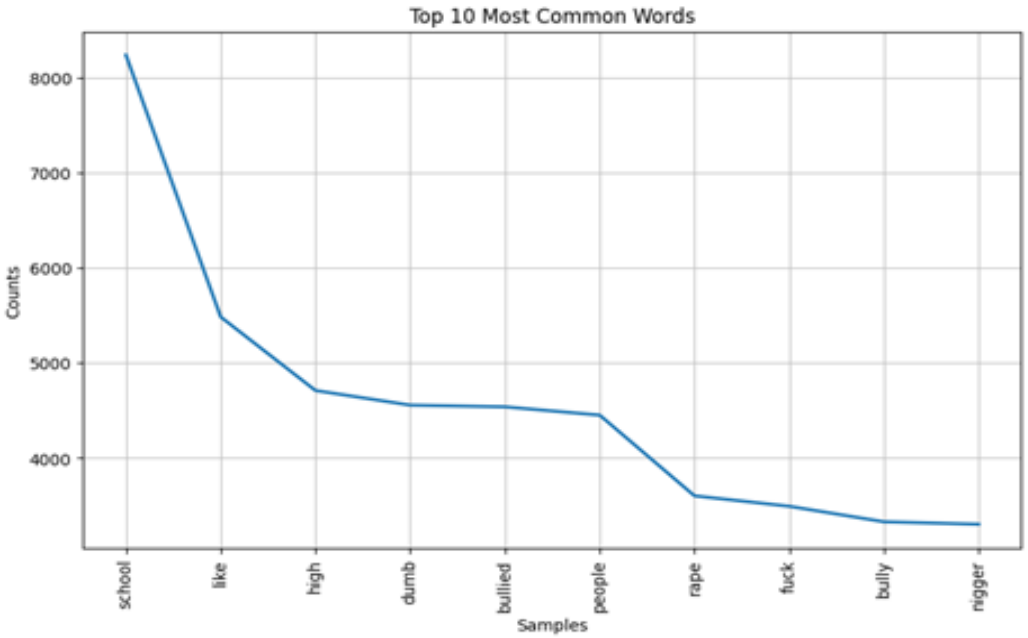

Fig. 3 : Word frequency Line Chart

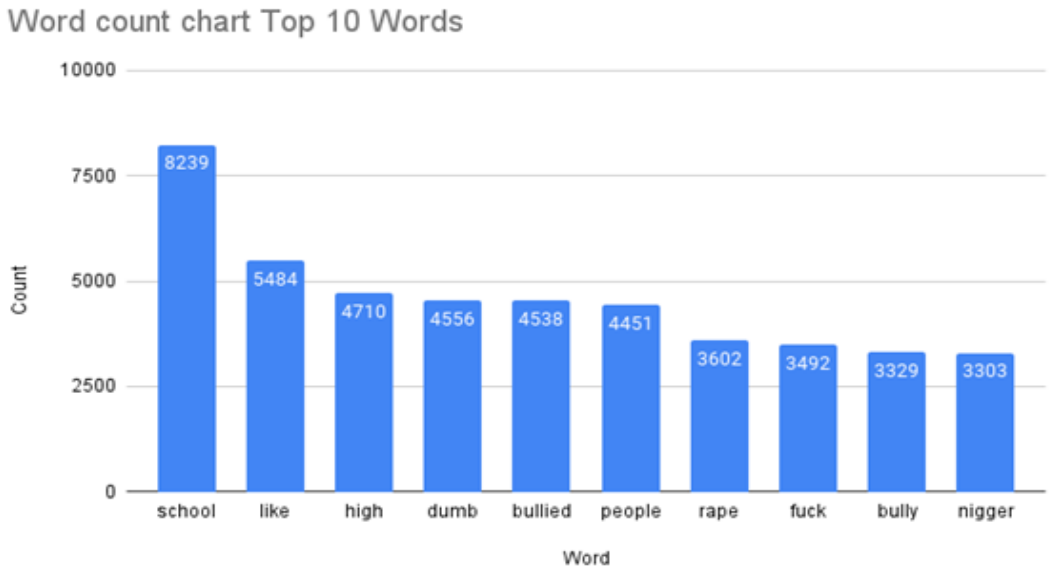

Fig. 4: Most common Words in the Data set

2.5. Retrieval and Selection of Features

One of the Key elements of the features testing process is the features selection process. A predictive model is created in this way by lowering the number of input variables. Through the removal of unnecessary or redundant aspects, features selection techniques are used to decrease the amount of input Variables. The list of characteristic is then reduced to the ones that are most important to the Machine Learning Algorithm. In supervised Learning, a

features selection aims to determines the most beneficial collection patterns that may be developed to generate effective Frame work. We employ Feature Selection in this Machine Learning technique to increase the process' accuracy. By concentrating on the most important factors and removing the excessive and irrelevant ones. It also improves the Algorithms' ability to predict outcomes. The Following three advantages of Features Selection.

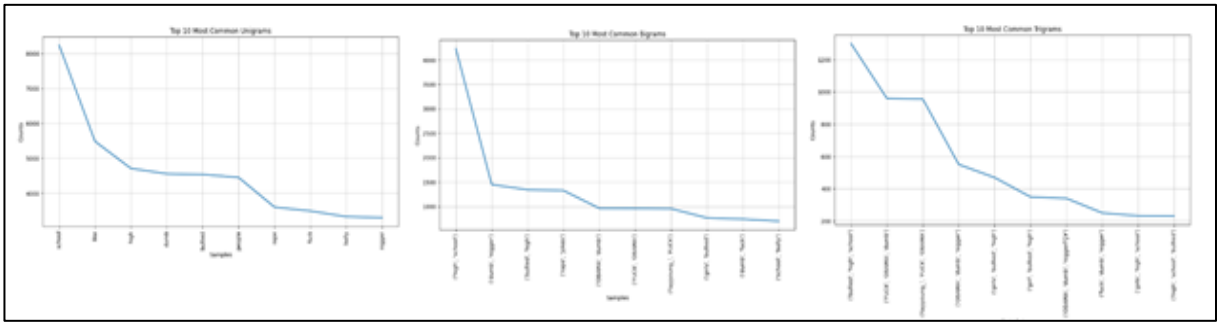

Fig. 5: a) Unigram, b) Bigram and c) Trigram

- Lowest Class Fitting: In this less duplicate data reduces the chance that judgments will be based on Noise.
- Enhance Accuracy: Better modeling evidence to support from fewer misleading data.
- Minimizes the Training Time: Algorithm run more quickly with less data.

2.6. Machine Learning

Artificial Intelligence has a branch known as Machine Learning, which gives computer a capacity to automatically learn from data and previous experiences to spot patterns and make predictions with a minimum human involvement. In this we used Machine Learning to extract the tweets from twitter because Machine Learning uses methods to find trends and learn in an ongoing procedure, deriving valuable knowledge from massive amount of data. Instead of depending on any preconceived equation that may serve as a model, Machine Learning Algorithm apply computing techniques to learn directly from data. We have used Supervised Machine Learning Algorithm this kind of Machine Learning uses supervision, where computers are trained on labeled datasets and allowed to make predictions based on the training data. According to the labeled data sets, certain input and output parameters have already been mapped. Consequently, the input and related output are used to train the machine. The analysis is also examined for accuracy. The Machine Learning Algorithm is either deployed or continually trained with an increased testing set until the necessary accuracy is obtained, depending on its accuracy. We have applied Support Vector Machine

Algorithm because the SVM method's objective is to establish the optimal line or decision boundary that can divide break n-dimensional area into groups, allowing us to quickly classify fresh data points in the future. A Hyper plane is the name given to this optimal decision boundary (Fig. 6).

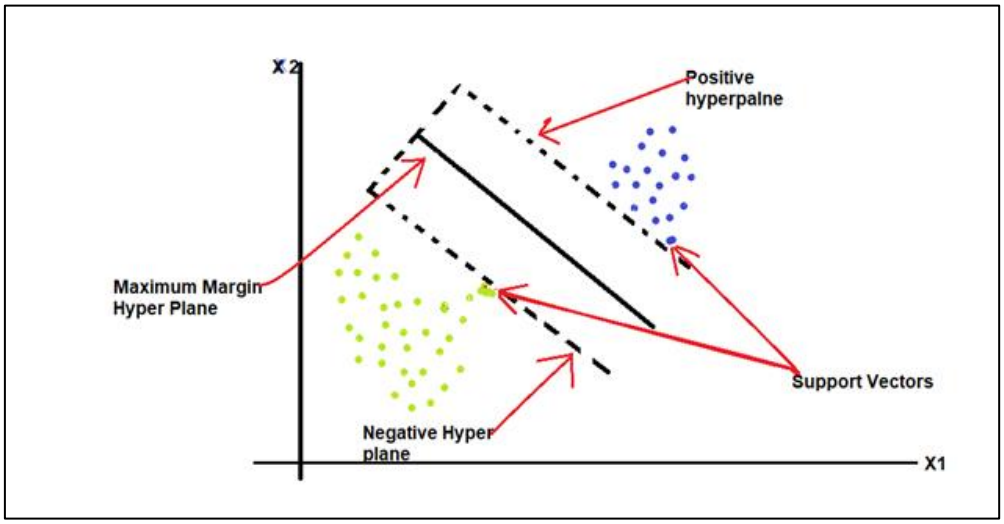

Fig. 6: Decision Boundary

3. Results and discussion

The text was initially converted into numerical form and then passed through a label encoder since the training data Algorithm was built on numeric input. Subsequently, the dataset was divided into 80% for training and 20% for testing purposes, followed by the application of the SVM Technique in a Machine Learning Algorithm. It demonstrates the precision achieved after diagnosing the Twitter dataset, which contains tweets categorized as instances of bullying.

The subsequent section compares the testing outcomes of the classifier with several machine learning models considered to be industry standards. Based on various input dataset scenarios, the prediction results for cyberbullying were validated, yielding accuracy scores of 83.26%, precision scores of 83.66%, and recall scores of 83.05% (Fig. 7). Confusion matrices were employed in the performance evaluation process, as depicted in Fig. 8.

SVM classifier trials were conducted for each data set input scenario (Fig. 9). Subsequently, the Machine Learning SVM model acquired the following: class count, word count, data collection, labeling, purification, retrieval, and an SVM classifier, all while being taken into consideration.

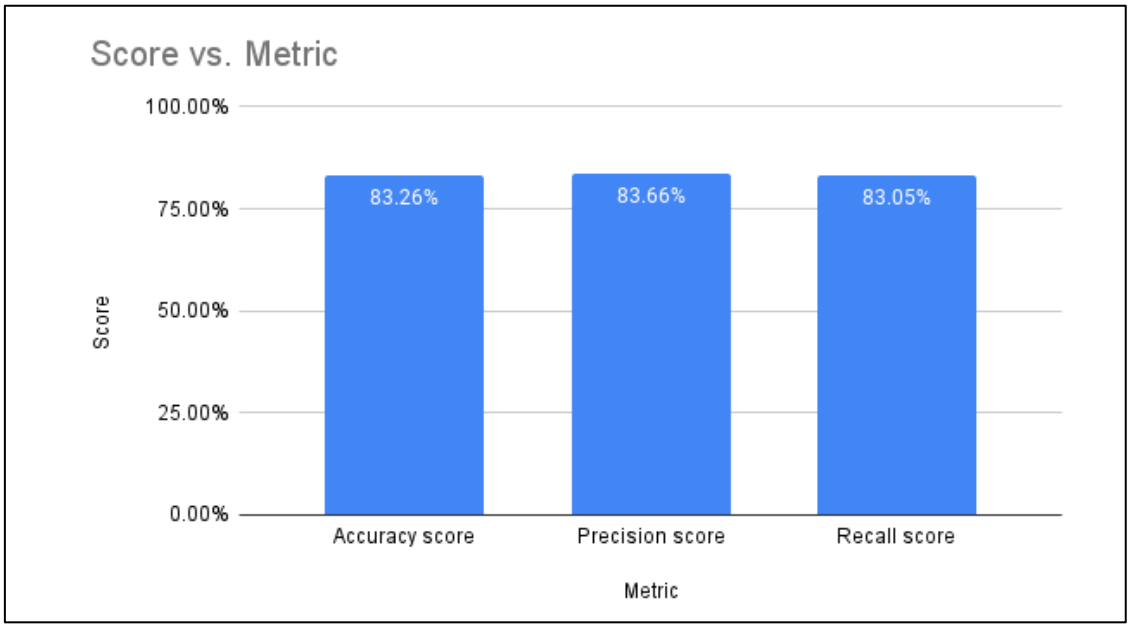

Fig. 7: Model performance on the basis of accuracy, precision and recall

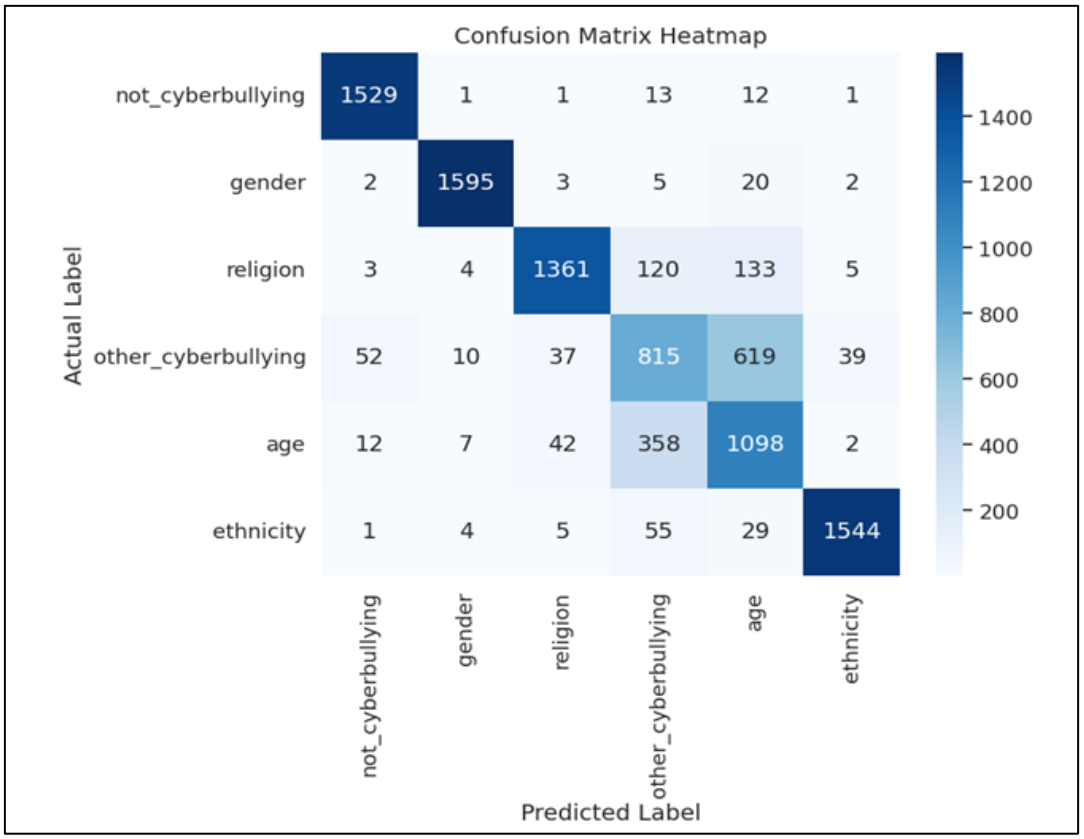

Fig. 8: Confusion matrix of model

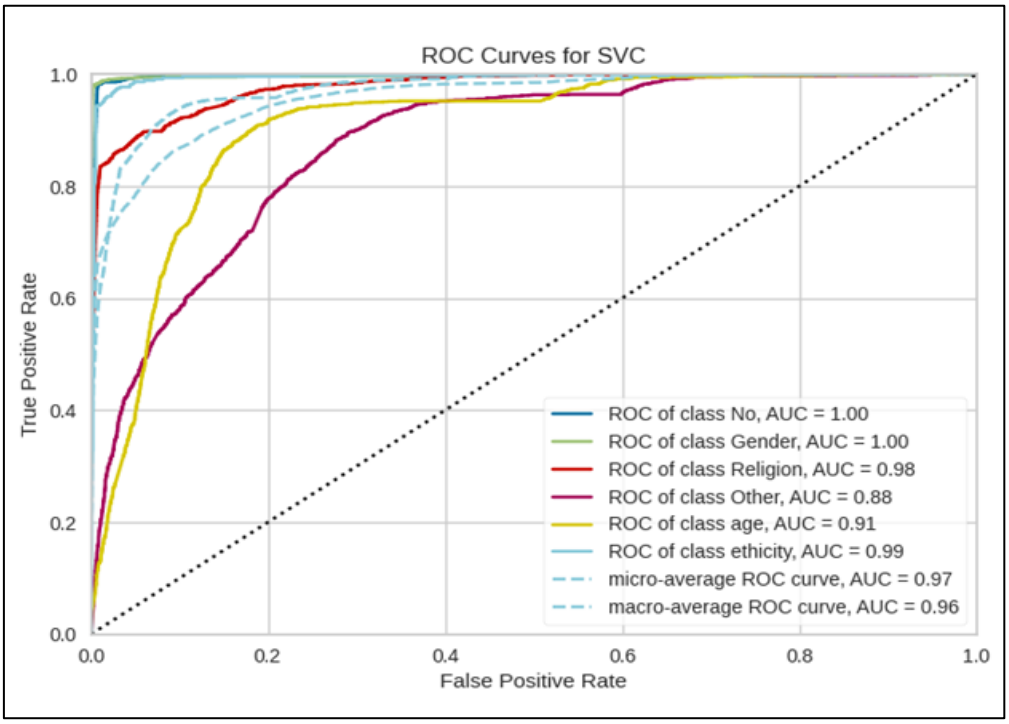

Fig. 9: ROC curve of SVM classifier

4. Conclusion

In this study machine learning was implemented to detect cyber bullying. The phrase "Cyber Bullying" is broad and has many diverse meanings. Social Media is one of those elements that is essential for recognizing bullying. On Twitter criticizing someone might have negative impact on the victim. According to our observation, earlier studies did not take into account this Twitter on Bullying. So such component was also included in this research. Our findings indicate that Social traits might reveal potential online bullying. In essence, this implies that comprehension of the Social environment in which a Communication is shared is as significant to that of the communication itself.

Acknowledgment

I would want to express my gratitude to Allah Tallah, the most Gracious and most Merciful for bestowing upon me the capacity and chance to study as well as for His unending gifts.

References

[1] Ali, A., & Syed, A. M. (2020). Cyberbullying detection using machine learning. *Pakistan Journal of Engineering and Technology*, *3*(2), 45-50.

[2] Huang, Q., Singh, V. K., & Atrey, P. K. (2014, November) .Cyber bullying detection using social and textual analysis .In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia* (pp. 3-6).

[3] Murshed, B. A. H., Abawajy, J., Mallappa, S., Saif, M. A. N., & Al-Ariki, H. D. E. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. *IEEE Access*, *10*, 25857-25871.

[4] Jain, V., Saxena, A. K., Senthil, A., Jain, A., & Jain, A. (2021, December). Cyber-Bullying Detection in Social Media Platform using Machine Learning .In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)* (pp. 401-405). IEEE.

[5] Sharma, C., Ramakrishnan, R., Pendse, A., Chimurkar, P., & Talele, K. T. (2021, July).Cyber-Bullying Detection Via Text Mining and Machine Learning .In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[6] Wu, C. S., & Bhandary, U. (2020, December).Detection of hate speech in videos using machine learning.In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 585-590) .IEEE.

[7] Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020, December). Cyberbullying detection on social networks using machine learning approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-6). IEEE.

[8] Behzadi, M., Harris, I. G., & Derakhshan, A. (2021, January).Rapid Cyber-bullying detection method using Compact BERT Models.In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)* (pp. 199-202). IEEE.

[9] Singh, N., & Sharma, S. K. (2021, March). Review of Machine Learning methods for Identification of Cyberbullying in Social Media .In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (pp. 284-288). IEEE.

[10] Nikhila, M. S., Bhalla, A., & Singh, P. (2020, July).Text imbalance handling and classification for cross-platform cyber-crime detection using deep learning .In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

[11] Alotaibi, M., Alotaibi, B., & Razaque, A. (2021) .A multichannel deep learning framework for cyberbullying detection on social media.*Electronics*, *10*(21), 2664.

[12] Shah, R., Aparajit, S., Chopdekar, R., & Patil, R. (2020). Machine Learning based Approach for Detection of Cyberbullying Tweets. *Int. J. Comput. Appl*, *175*(37), 51-56.

[13] Chandrasekaran, S., Singh Pundir, A. K., & Lingaiah, T. B. (2022). Deep learning approaches for cyberbullying detection and classification on social media.*Computational Intelligence and Neuroscience, 2022*.

[14] Muneer, A., & Fati, S. M. (2020).A comparative analysis of machine learning techniques for cyberbullying detection on Twitter.*Future Internet, 12*(11), 187.

[15] Islam, M. M., Uddin, M. A., Islam, L., Akter, A., Sharmin, S., & Acharjee, U. K. (2020, December). Cyberbullying detection on social networks using machine learning approaches. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-6). IEEE.

[16] Kargutkar, S. M., & Chitre, V. (2020, March).A study of cyberbullying detection using machine learning techniques .In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 734-739). IEEE.

[17] Kumar, R., & Bhat, A. (2022).A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. *International Journal of Information Security, 21*(6), 1409-1431.

[18] Wang, K., Xiong, Q., Wu, C., Gao, M., & Yu, Y. (2020, July).Multi-modal cyberbullying detection on social networks .In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[19] Dalvi, R. R., Chavan, S. B., & Halbe, A. (2020, May). Detecting a Twitter cyberbullying using machine learning .In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp .297-301).IEEE.

[20] Pericherla, S., & Ilavarasan, E. (2021, February) .Performance analysis of word embeddings for cyberbullying detection. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1085, No. 1, p. 012008).IOP Publishing.