MACHINE LEARNING-BASED CYBERBULLY DETECTION

Mandadi Jashwanth Reddy[1], Balapuram Pruthvi Reddy[1], Chityala Charan[1] and Kasi Bandla[2]
Department of Electronics and Computers Engineering[1,2]
Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, 501301
Email – kasibandla@sreenidhi.edu.in[2]

Abstract: The digitization of relationships and technological improvements had a significant impact on the centennials, who were forced to keep social media accounts under strict maintenance requirements. Despite the amusement Cyberbullying has been acknowledged as a significant problem in Malaysia and many people in their 100s are its victims. This is due to the accessibility of social media. A few studies, however, have been published in identifying attempts at cyberbullying on social media. Therefore, a solution that can identify attempts at cyberbullying on social media and uses appropriate data science techniques would be excellent. This study developed three supervised learning predictive models—Naive Bayes, SVM, and LSTM—using the suspicious tweets dataset from Kaggle. The models were tweaked using Random Grid Search and Keras tuner to provide a workable solution. In conclusion, the accuracy and area under the curve (AUC) values for the Naive Bayes model were 88.4% and 0.81, respectively, the best. Whilethe LSTM model came in second place with an AUC value of 0.58 and an accuracy of 90.6%. As a result, the accuracy and AUC values of the LSTM model can both be enhanced with more records.

Keywords: Cyberbully, threats, machine learning models, cyberattacks, victims.

1 Introduction

Cyberbullying is a severe societal issue that has acquired substantial attention in recent years due to the extensive use of the internet and social media. It describes the use of electronic tools and communication technology to annoy, threaten, or intimidate others. The internet's anonymity and accessibility have made it simpler for cyberbullies to target their victims, which frequently results in serious effects including sadness, anxiety, and even suicide. According to Mahlamgu Owolawi in 2018, cyberbullying is the willful and persistent abuse or pestering of a person through tools like social media platforms. This behaviour is meant to intimidate and degrade the target person. According to a study, these are just a few examples of cyberbullying that may occur on social media. Other instances of this behaviour include disseminating false information or humiliating photographs of the victim, sending direct messages with abusive language, or pretending to be the victim and sending unsolicited messages on their behalf. Bark team analysis of student suicides via the Internet in 2017 tells Cyberbullying has increased due to abuse, which has led to some student suicides.

2 Presentation of the idea

The purpose of this project is to develop an appropriate text classification model for determining the intent of cyberbullying, perform proper data pre-processing into an appropriate format for dataanalytics processing, and evaluate the performance of the predictive model using the proper evaluation measures.

3 Review of Related Literature

As per our research, multiple journals have been published on cyberbullying, but various challenges are faced in their research due to their low latency in the methodologies used. Some ofthem are evaluated based on our analysis as follows.

Cyberbullying is a complex issue that has been the subject of numerous studies, some of which have encountered challenges due to the limitations of their research methodologies. In one study, data science was employed to identify cyberbullying attacks, but the resulting accuracy was limiteddue to the unpredictability of big data. The precision rate and recall were found to be in the averagerange, indicating that the model could only be used to a limited extent, particularly in industrial settings [1]. Another study addressed these limitations by developing an initial model using Support Vector Machines (SVM), which resulted in improved accuracy rates and average precisionand recall. Subsequently, a Convolutional Neural Network (CNN) model was developed to addressa wider range of challenges and demonstrated further improvements in accuracy and precision compared to the previous research [2].

To further compare the efficacy of deep learning methodologies with traditional machine learningmethods, yet another study utilized Long Short-Term Memory (LSTM) models, resulting in a morenuanced understanding of the strengths and limitations of different approaches to cyberbullying detection [3]. These findings collectively demonstrate the ongoing efforts to develop more effectiveand accurate methods for detecting cyberbullying but also highlight the challenges that remain in this field. Overall, these findings highlight the ongoing efforts to develop more effective and accurate methods for detecting cyberbullying. However, the challenges posed by the variability ofbig data and the complexity of this issue continue to present significant hurdles for researchers. Further research is needed to address these challenges and to develop more robust methods for identifying and preventing cyberbullying.

4 Implementation of the prototype

4.1 Proposed Method

This project aims to identify cyberbullying intent inside tweets from Twitter to support parents, teenagers, and authorities with cyberbullying difficulties, in addition to reducing events that may happen as a result of cyberbullying action. The finished product won't simply be a predictive model; it will also give decision-makers output in the form of a word cloud and bag of words.

4.2 Methodology

For organizing, planning, and carrying out data mining activities, organizations typically use the CRISP-DM (cross-industry standard process for data mining) approach. It offers a planned and rigorous process that is divided into 6 parts. The business challenge is defined and the project goals for the cyberbullying tweet detection are established in the first stage of the project, known as business understanding. For example, the goal might be to quickly identify tweets that involve cyberbullying and notify the appropriate authorities. Data exploration and collection, also known as data understanding, is the next phase. To complete this phase, it may be necessary to collect a sample of tweets from social media sites, understand how the tweets are organized, and identify the various types of cyberbullying that are frequently observed in the data. The next step is data preparation, which involves employing pre-processing techniques to eliminate or manipulatenoise, duplicate data, incomplete data, and the generation of new variables. The modelling phaseis wheremodelling methods are selected following the research's goal, which is to forecast cyberbullying. Thus, in this project, support vector machines and naive Bayes machine learning methods are employed.

4.3 Formulas

The confusion matrix is used to evaluate the model, allowing for critical evaluations of the model'saccuracy, precision, recall, and F1 score. Based on the formula shown below, accuracy, precision, F1 score, and recall are calculated.

$$\text{Accuracy} = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + True\ Negative + False\ Positive + False\ Negative)}$$

$$\text{Precision} = \frac{True\ Positive}{(True\ Positive + False\ positive)}$$

$$\text{Recall} = \frac{True\ Positive}{(True\ Positive + True\ Negative)}$$

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$= \frac{2(True\ Positive)}{2(True\ positive) + False\ Positive + False\ negative}$$

To do this, it may be essential to develop a web application or API that can evaluate tweets and classify them as cyberbullying or non-cyberbullying after the models have been put into use in thereal world to assist business decisions.

4.4 Overview of the Process Model

By employing data preprocessing techniques like data cleaning, the collected dataset will be pre-processed from data that are dirty and have noise, duplication, and missing entries. Since the data only consists of two columns—the tweets and a Boolean variable indicating the intention to engagein cyberbullying—variable selection will not be carried out. Data transformation and data cleaningare carried out during the pre-processing phase. The dataset will be split into training and testing data sets with an 80:20 ratio after pre-processing is complete. The model will next be trained and tested using three algorithms—support vector machine, Naive Bayes, and long-term short memory—on the preprocessed dataset. After the model training is finished, the model is tested to determine which model is best based on the test results [5]. So, following the identification of the best model, the best model will be implemented in a web application.

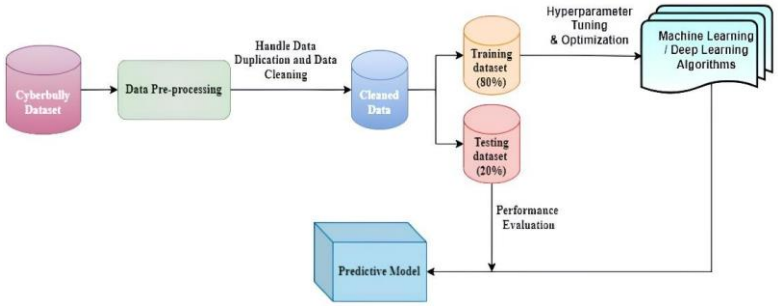

Fig. 1. Process flow diagram

4.4.1 Data Pre-processing

Every data analytics project begins with exploratory data analysis (EDA), where the dataset is frequently examined to gain insight into characteristics like data type and uncover trends to further investigate the Kaggle dataset used for the cyberbully detection model containing suspicious tweetswith connections to terrorism threats and cyberbullying. The dataset, which includes more than 60000 tweets, was compiled from Twitter.

4.4.2 Data Exploration (Pre-Cleaning)

The variables that make up the label column in Fig. 2 show that the proportion of suspicious tweets far outweighs the proportion of non-suspect tweets. Over 90% of the dataset's tweets are suspicious, with the remaining 10% being suspicious tweets. The label column includes 6k records of doubtful records and 50k records of non-suspicious recordings. Since the dataset is not balanced since suspicious records make up the remaining records and given non-suspect material fills up about 90 of the dataset, a synthetic minority over-sampling method (SMOT) was employed to balance the data.

4.4.3 Data Cleaning

The pre-processing steps involve using the "drop_duplicates" function to eliminate duplication. After removing duplicates, there are 6133 suspicious and 53,574 non-suspicious data points. Converting the text to lowercase standardizes it. Emojis are transformed into text using the "demojize" function from the emoji library. NLTK's word tokenize function tokenizes the messages, removing punctuation and stop words. Lemmatization using wordnet and POS tagging enhances accuracy.

4.4.4 Data Transformation

Following the completion of data cleaning, data is split into training and test datasets with an 80:20 ratio, and then a count vectorizer is used to tokenize a set of text documents and create a vocabulary of known words as well as encoding new documents with that vocabulary. The result is an encoded vector with the full vocabulary's length and an integer count of how many times each word appears in the input data. In 2020, Brownlee Once the count vectorizer has been run, itis preserved because if it were to be run again, the model would not match because the vectorizer would not be unique.

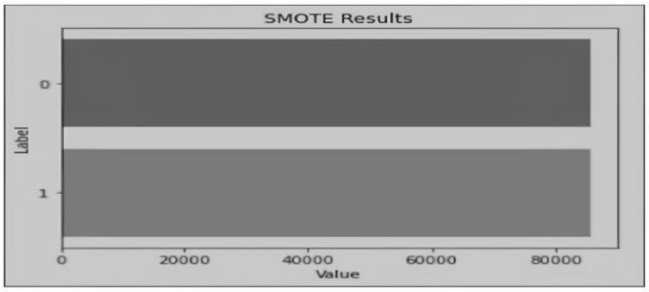

Fig. 2. SMOTE Result

When producing data points, Smote uses the KNN algorithm, choosing the nearest adjacent point from the minor class. Once minor and major classes are equal, this process is repeated. To prepare the vectorized training dataset for ML models, the smote mechanism is used. This mechanism balances suspicious data 0 with non-suspicious data 1, resulting in a balanced data split that minimises overfitting (see Figure 3).

Data Exploration (Post-Cleaning): A word cloud is a visual representation of words that providesa fundamental understanding of a dataset; in this case, the word cloud represents Twitter messages.

5 UML Diagrams

Class diagrams: Class diagrams are helpful for illustrating a system's static structure and for clarifying the connections between different classes, their properties, and functions. They help in defining the characteristics and functions of the classes in greater depth, as well as in determining the interfaces and dependencies between the classes.

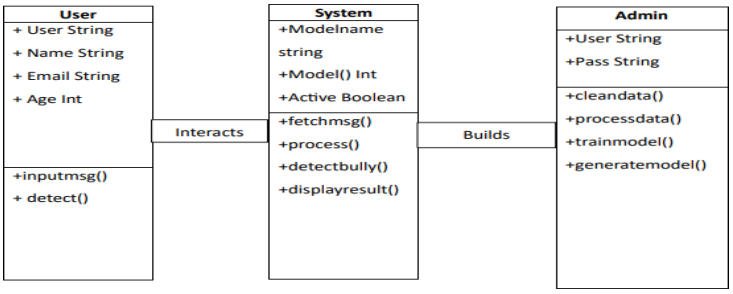

Fig. 3. Class Diagram

Use case diagram: In this illustration, the actor tweets at the system to start the process of detecting cyberbullies. After that, the administrator performs several activities on the data, including preprocessing, machine learning model training, testing, analysis, and perhaps a model improvement in the context of a more general detection use case.

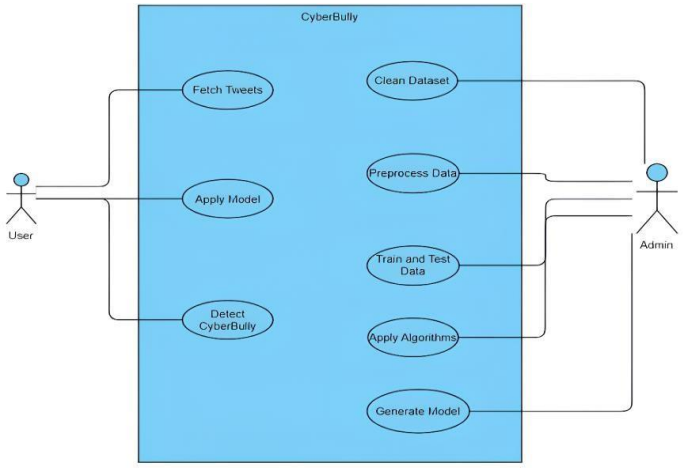

Fig. 4. Use Case Diagram

6. Result

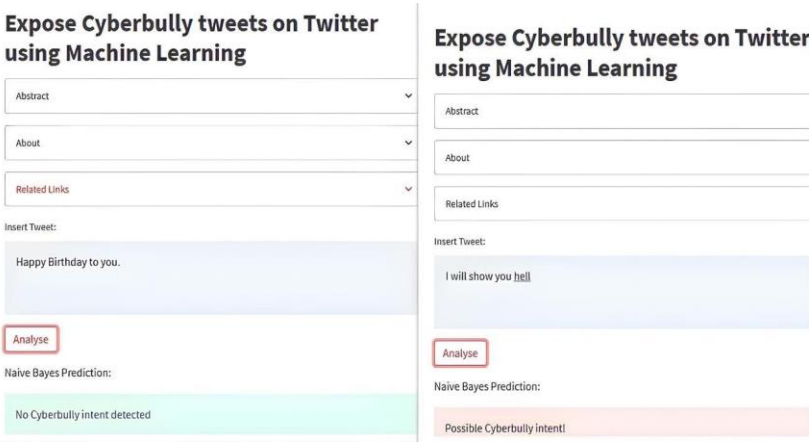

Fig. 5. Positive Cyberbully Intent and no Cyberbully Intent

7. Conclusion

Cyberbullying is a problem that is getting worse online and needs to be stopped in order to protect the safety and well-being of users on social media platforms. With excellent accuracy rates, the application of neural networks to recognize various types of cyberbullying in tweets has yielded encouraging results. The neural network was trained to recognize patterns and characteristics of cyberbullying using a large dataset of tweets, enabling quick identification and prevention. Neural networks have the potential to be a powerful tool in the fight against cyberbullying, and more research in this area might contribute to the creation of a more welcoming and secure online community. Giving a solution using modern technology, such as a predictive model, is advantageous for the social media industry. The Naive Bayes model, which had the best results out of the four, had an accuracy rate of 88%. To create a model that is more accurate and effective than previous models, proper data pre-treatment and optimization techniques are needed. There are numerous ways to improve the LSTM algorithm's layers because there are so many different possibilities accessible. Additionally, classic algorithms like the random forest might also benefit from more inventive options and combinations.

However, there are also limitations to the effectiveness of machine learning algorithms in detecting cyberbullying. For example, machine learning algorithms may struggle with detecting sarcasm, irony, or other forms of nuanced language that may be used in cyberbullying. Additionally, machine learning algorithms require large amounts of data to train effectively, and there may be privacy concerns associated with collecting and analyzing personal data. Overall, cyberbullying detection using machine learning algorithms is a promising approach to addressing the problem of cyberbullying. With continued research and development, these algorithms have the potential to become even more accurate and effective at detecting cyberbullying and helping to create a safer and more respectful online environment.

References

1. Dalvi, R. R., Chavan, S. B., & Halbe, A. (2020). Detecting A Twitter Cyberbullying Using Machine Learning. IEEE. Mumbai: IEEE. doi:10.1109/ICICCS48265.2020.9120893
2. Al-Ajlan, M. A., & Ykhlef, M. (2018). Deep Learning Algorithm for CyberbullyingDetection. International Journal of Advanced Computer Science and Applications, 9, 199-205. doi:10.14569/IJACSA.2018.090927
3. Ghosh, S., Chaki, A., & Kudeshia, A. (2021).Cyberbully Detection Using 1D-CNN and LSTM. Proceedings of International Conference on Communication, Circuits, and Systems. Bhubaneswar: KIIT University. doi:10.1007/978-981-33-4866-0_37
4. Neelakandan S, Sridevi M, Saravanan Chandrasekaran, Murugeswari K, Aditya Kumar Singh Pundir, Sridevi R, T.Bheema Lingaiah, "Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2163458, 13 pages, 2022. https://doi.org/10.1155/2022/2163458
5. R. Jeevitha, K. Chaitanya, N. Mathesh, B. Nithyanarayanan and P. Darshan, "Using Machine Learning to Identify Instances of Cyberbullying on Social Media," *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India, 2023, pp. 207-212, doi: 10.1109/ICSCDS56580.2023.10104720.
6. M. Nisha and J. Jebathangam, "Deep KNN Based Text Classification for Cyberbullying Tweet Detection," *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2022, pp. 1550-1554, doi: 10.1109/SMART55829.2022.10047054.
7. A. A. Khan and A. Bhat, "A Study on Automatic Detection of Cyberbullying using Machine Learning," *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2022, pp. 1167-1174, doi: 10.1109/ICICCS53718.2022.9788299.