IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Machine Learning Based Cyber Bullying Detection

[1]Ejjigiri Siri Chandana, [2]Sudharshanam Bhargavi, [3]Mohammed Yaseen Nawaz, [4]Nemmani Swapna

[1234]Department of Electronics and Computer Engineering
Sreenidhi Institute of Science and Technology, Ghatkesar, India

Abstract: In the digital era, providing a safe and respectable online society is a top priority. Cyberbullying, an epidemic that presents a considerable danger to one's mental and social life, is impossible to avoid. This study addresses the issue of cyberbullying classification in Twitter data. We use the simple measures of sentiment analysis to do so in tandem with machine learning. In our project, we employ an SVM to classify tweets as cyberbullying or non-cyberbullying. The sentiment analysis pipeline produced by the sentiment analysis includes text preprocessing, transforming and weighting vectorization, and model training. Moreover, the system employs an interactive widget interface that allows the operators to infiltrate a tweet, and the system processes their input tweets to clean them up using usual expression extraction from URLs and non-alphanumeric characters before passing them to the pre-trained TF-IDF model for training and predicting if it's cyberbullied. Performance evaluation is performed through a classification report which provides precision, recall, F1-score, and support of each class metric. This system has the potential to enable users and administrators to intervene in cyberbullying cases in a preventative manner, making the online community a safer and more inclusive place.

Keywords - Cyberbullying detection, Sentiment analysis, Machine learning, Support Vector Machines (SVM), TF-IDF vectorization, Text preprocessing.

I. INTRODUCTION

Over the recent years, social media platforms have seamlessly become part of our daily lives as individuals created ways for communication and expression. However, the sunshine of the platforms has also been a seedbed for several vices that impact individuals adversely, and one of the factors is cyberbullying. Cyberbullying is defined as the use of digital communication to harass, intimidate, or cause harm to others. Often, the repercussions of this form of harassment are dire and overwhelming to the victims. Twitter as one of the widely used social media platforms I such an environment. Due to the brevity and ubiquity of the tweets, people have ventured in sharing harmful content making the platforms unhealthy to the users. Therefore, it is essential to detect and control cyberbullying on Twitter to make the platforms safe and healthy.

The primary goal of this project is the detection of cyberbullying in the Twitter social media network. With the help of sentiment analysis and certain machine learning techniques, the project extorts the information from tweets and determines if a message could be considered cyberbullying or not. This way, the data may support ideas about the prevalence of cyberbullying and serve as a tool for active users and platform administrators to deal with the problem from a preventative perspective. The project works based on the SVM – Support Vector Machines machine learning algorithm. This algorithm classifies the data well and may distinguish specific patterns for the cyberbullying situation on the platform.

The sentiment analysis pipeline for this research project comprises several critical processes. First, the text data is preprocessed to eliminate noise and normalize the state of the input. Subsequently, the text transformed into numerical features utilizing TF-IDF vectorization, a method that enables one to catch the word importance in a document as opposed to a collection of documents. Lastly, the SVM classifier is trained on the vectorized data to understand the pattern in the cyberbullying tweets. This study harbors an interactive UI,

enabling a user to input tweets and feed the network in real-time for instant analysis. By providing a platform for cyberbullying discernment, we allow most users to make their judgment with regard to the content they encounter in Twitter.

The efficacy of the cyberbullying detection system is determined by multiple stringent performance metrics, such as precision, recall, the F1-score, and support. The qualitative assessment of the model's performance enables us to determine the accuracy with which it can recognize cyberbullying content and how reliable the output is. In conclusion, this project presents an initiative to combat the widespread concern of cyberbullying on Twitter by adopting a preventative approach. By focusing on developing a system which integrates sentiment analysis with machine learning models and an interactive tool for users, we strive to make meaningful contributions to address one of the fundamental issues facing users, particularly young social media consumers.

II. LITERATURE SURVEY

In the light of this, cyberbullying has become an enormous problem and an important social problem in Twitter which in turn it has adverse implications for victims such as mental illnesses and psychosocial problems. A different study has to a great extent discussed the fact that cyberbullying is very common online and a number of issues involved with the detection and treatment of such behavior lie in online. What is often shared using Twitter in the way of space and the anonymity of the platform only aggravate the spread of harmful content. This means that appropriate approaches should be provided for the prevention and treatment of cyberbullying by effectively applying these strategies.

The role of sentiment analysis, machine learning and other techniques has been garnered by the growth in popularity among researchers in countering cyberbullying on the social media platforms. A text level sentiment analysis allows the development of algorithms for finding words or phrases that reveal the negative background of cyberbullying. The use of algorithm, coupled with support vector machine (SVM), has become a timely opportunity to differentiate the tweets into cyberbullying and non-cyberbullying categories based on linguistic markers and text patterns. At the same time these methods take advantage of labeled datasets to guide algorithms capable of distinguishing between harmful and benign content, therefore facilitating the identification of cyberbullying cases.

The art of making the key data and patterns of the social media in visual representation known and understood is referred to as data visualization techniques and is among most important tools for comprehending social media data. Methods such as histograms, word clouds and pie charts staples of this exploration are found handy in analyzing tweet chunks and taking frequencies of occurrence into account. Such kinds of visualizations present critical views into cyberbullying specific related stories, so that the same themes and key-words associated with this behavior can be identified.

Features, such as precision, recall, F1- score, and ROC curve analysis, are fundamental test diapers to be used to assess the performances of text classification models in cyberbullying detection. These metrics allow the researchers to assess the precision and trustworthiness of the classifiers when dealing with the bullying tweets and non-bullying tweets detection techniques. Therefore, the rates inform the success of the detection methods.

III. METHODOLOGY

3.1 Existing Methodology

This existing model adheres to the conventional infrastructure for sentiment analysis and cyberbullying detection which implies data collection, preprocessing, feature extraction, model training, evaluation and prediction. But it does not have the capability to facilitate real-time forecasts and user involvement which the proposed model with interactive widgets seeks to improve.

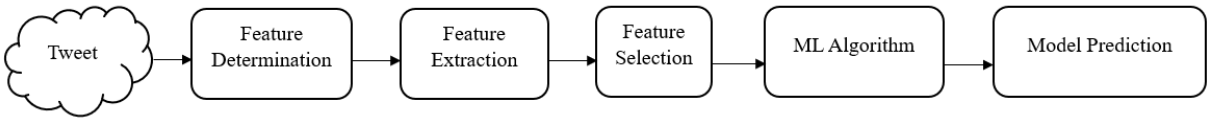

Figure 1: Block Diagram of Existing Model

Disadvantages:

3.1.1 Limited ability to capture temporal dependencies: Traditional machine learning methods, such as the Support Vector Machine (SVM) or the Random Forest, do not fit the temporal dependencies that are contributing in the analysis of the text sequences.

3.1.2 Limited ability to handle sequential data: The traditional machine learning methods do not have the ability to capture the sequential nature of data which is the limitation of cyberbullying tweet detection. Cyberbullying usually has a sequence of message exchanged and it is pertinent to closely scrutinize each text message to effectively identify or label bullying.

3.1.3 Limited ability to handle variable-length input*:* Traditional machine learning approaches are usually based on input data having a fixed-length. The Tweets can change as per the length of the phrases and, hence, being able to handle variable- length input is necessary for an accurate identification.

3.1.4 Limited ability to handle noisy data: A downside for cyberbullied tweet detection of the usual machine learning methods is the sensitivity to noisy data. It is very likely that tweets would be filled in with noisy data, including misspellings, inappropriate grammar, and informal language.

Such obstacles are overcome by employing the next-level machine learning principles, such as deep learning models, ensemble approaches, and natural language processing technologies, to optimize accuracy, scalability, fairness, and interpretability.

3.2 Proposed Methodology

The proposed methodology outlines a systematic approach for building a cyberbullying detection system tailored to Twitter. It involves several key steps:

3.2.1 Data Collection and Preprocessing: This cyberbullying detection system starts as an aggregation of Twitter data – both cyberbullying and non-cyberbullying tweets are included in one dataset. Utilizing a variety of treatment methods, conditioning the data and text normalization, guarantees the data cleanliness. Techniques such as tokenization and sentiment analysis tend to be also fine-tuned to help remove the noise, leaving only the harmful content Building on the accuracy and the precision of data collected and preprocessed this will lead the right way to more punctual analysis and classification. As a whole, the system's foundations include an intensive collection and filtration of Twitter data bringing precision to the cyberbullying detection megabytes, and thus making possible online security initiatives.

3.2.2 Data Visualization: Data visualization is the key component to explore the unique features of cyberbullying on Twitter. By count plot, histogram, word cloud and box plot, we assess the frequency of tweets, their lengths and also how word frequency varies. These visualizations help understand how the cyberbullying data looks like, thus aiding to discover important features for the classification algorithm. As an example, the controversial plot shows the distribution of cyberbullying labels in this manner, bringing to attention their occurrence. Histograms show the content of tweets, including word count differences, bullying language, and many other discourse characteristics. A word cloud demonstrates the most frequently occurring words that contribute to significant topics on cyberbullying. The tweets. Similarly, box plots present tweet length distributions based on labels. Such type of graphic, in turn, provides additional information about the variance in the length between cyberbullying tweets and other kinds of tweets. In general, data visualization helps us to better grasp the behaviors of cyberbullying in the case of Twitter, thereby taking care of identifying the classifications in efforts of making an online environment safer for all.

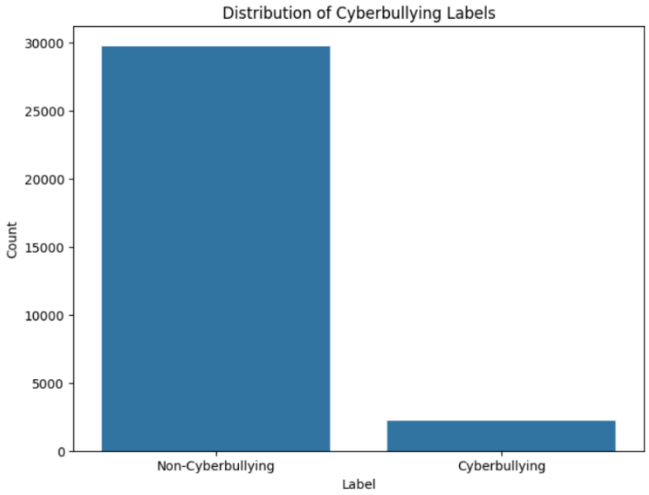

Figure 2: Distribution of Cyberbulling Labels

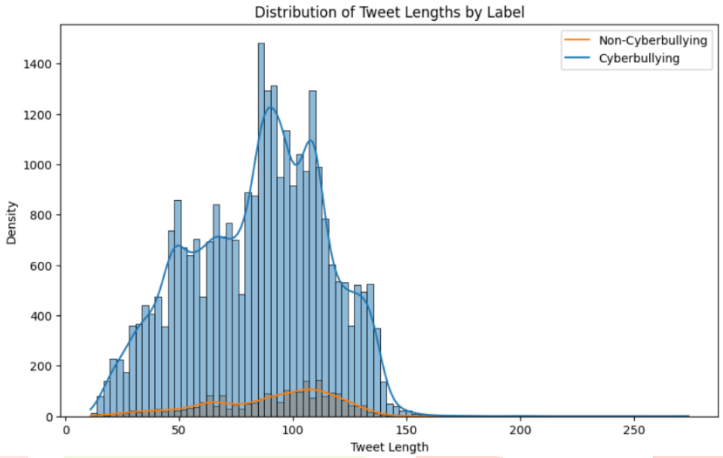

Figure 3: Distribution of Tweet Lengths by Label

Figure 4: Word Cloud for Non-Cyberbulling Tweets

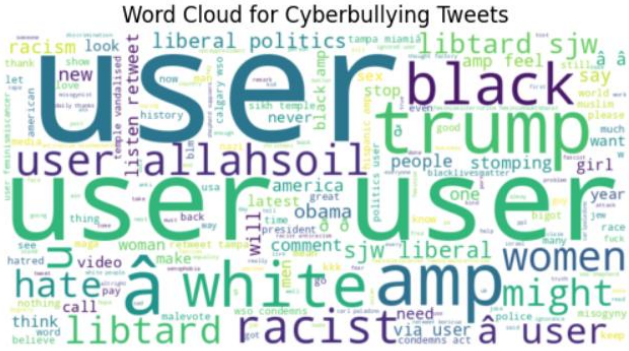

Figure 5: Word Cloud for Cyberbullying Tweets

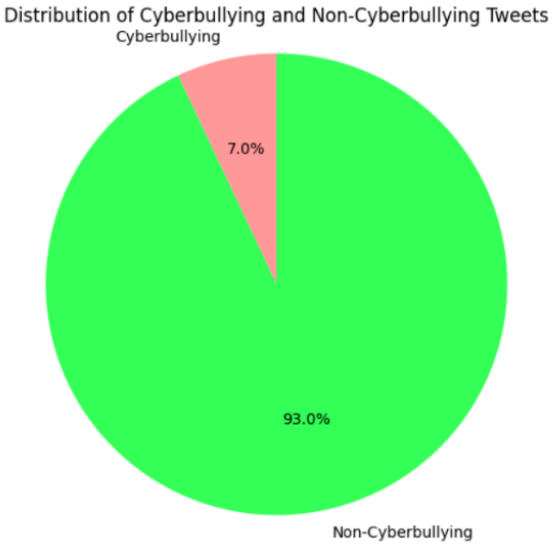

Figure 6: Distribution of Cyberbullying and Non-Cyberbullying Tweets as a Percentage

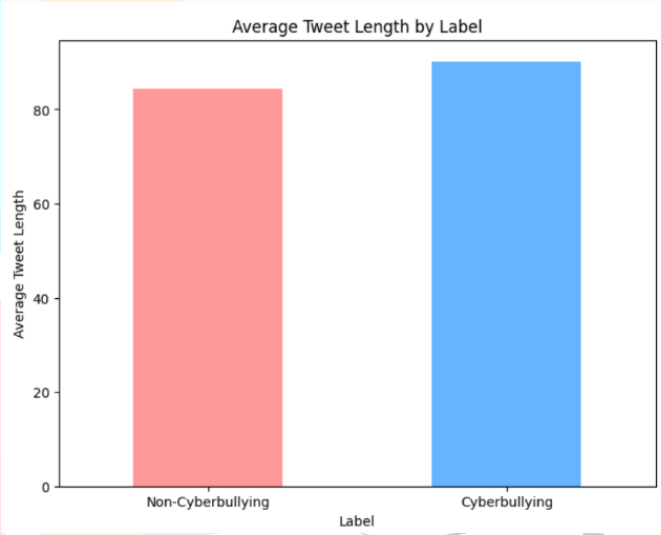

Figure 7: Average Tweet Length by Level

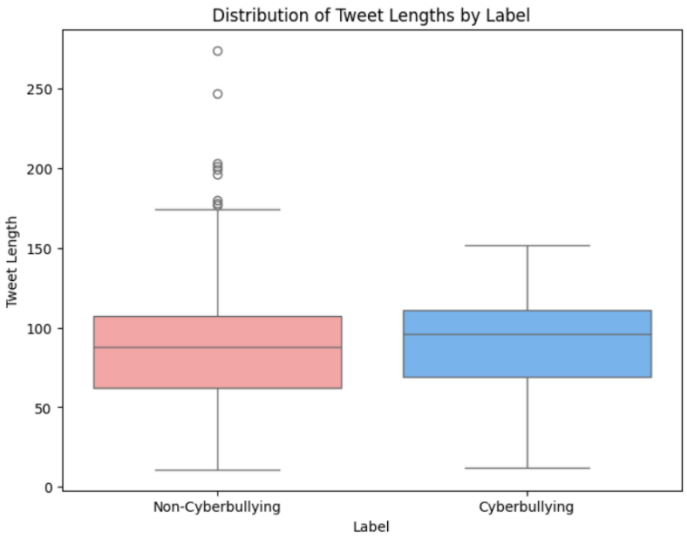

Figure 8: Distribution of Tweet Lengths by Label

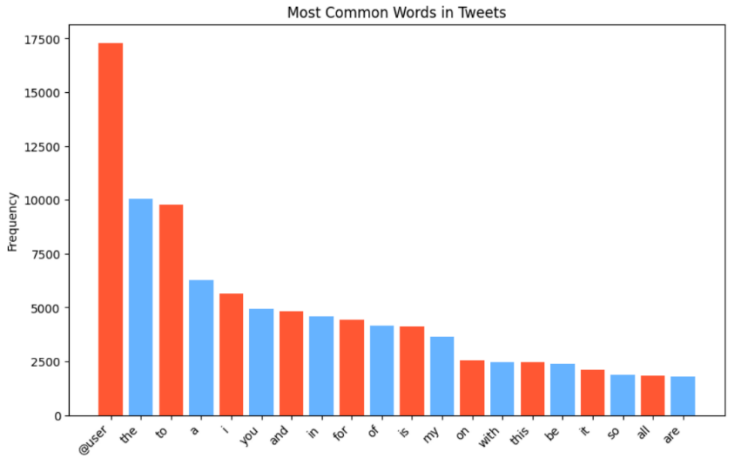

Figure 9: Most common words in Tweets

3.2.3 Text Classification: *Text* classification is the essence of our anti-cyberbullying tool on Twitter. With machine learning algorithms, mostly support vector machines (SVM), and the TF-IDF method, we convert textual data into features of the classifier that we can train using numerical data. The SVM model is trained to classify tweets that are transformed into cyberbullying and non-cyberbullying categories. Based on the strict evaluation measures of the accuracy, precision, recall, and F1-score plus visualization techniques using confusion matrices and ROC curves, we assess the model's performance. The actual example of a classified tweet may even prove, so to say, the utility of the cyberbullying detection system. In the nutshell, text classification helps us classify tweets with accuracy and thus contributes to the making of a safer online environment as it allows for the detection, and later the resolution of cyberbullying on Twitter.

```
SVC
SVC(kernel='linear', probability=True)
```

Figure 10: SVC Classifier

3.2.4 Model Evaluation: Evaluation of the model will make it possible to measure how efficient our cyberbullying detection system on Twitter. We use several metrics including accuracy, precision, recall, and F1 score to rate the model's efficacy in categorizing tweets as cyberbullying or non-cyberbullying. Moreover, the confusion matrices and ROC curves also show the model performance in graphical terms by depicting true positive and false positive rates. These metrics act as a measurement for the model's capacity to correctly identify cyberbullying activities and differentiate them from non-cyberbullying content. We rigorously evaluate the model's performance, so it can be relied upon and work towards the goal of secure online environment provision.

	precision	recall	f1-score	support
0.0	0.98	1.00	0.99	29720
1.0	0.99	0.74	0.84	2242
accuracy			0.98	31962
macro avg	0.98	0.87	0.92	31962
weighted avg	0.98	0.98	0.98	31962

Figure 11: Classification Report

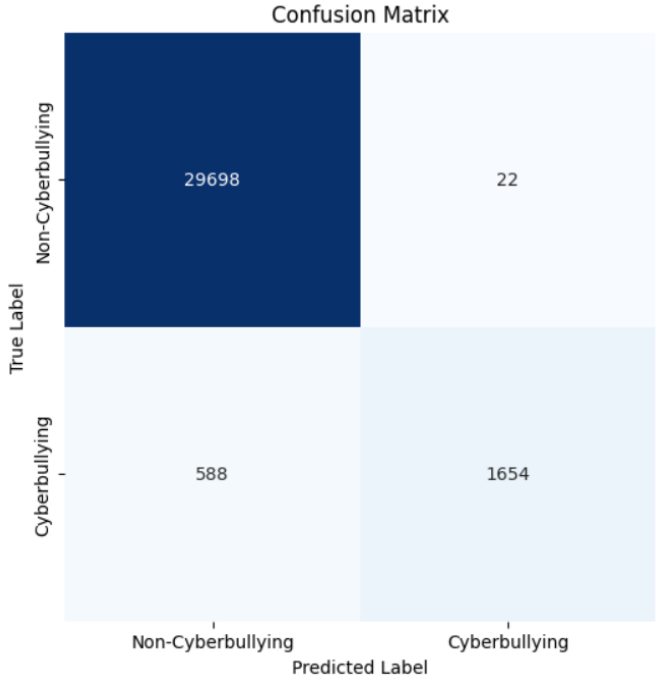

Figure 12: Confusion Matrix

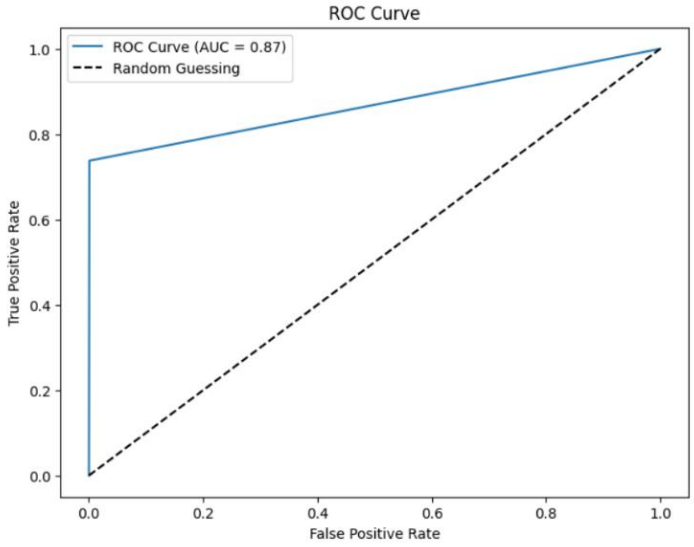

Figure 13: ROC Curve

The figure below shows the model we would like to implement in this project.

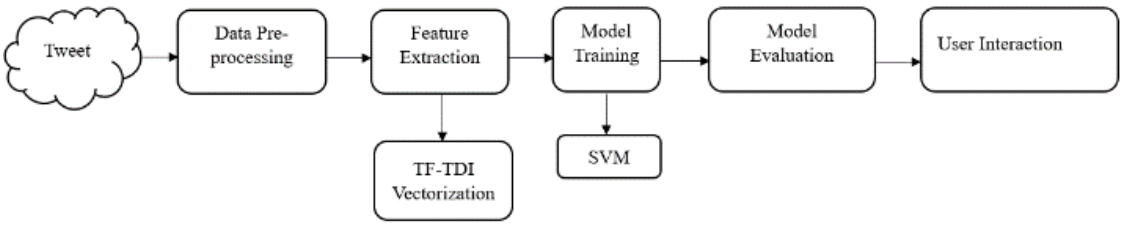

Figure 14: Block Diagram of Proposed Model

The block diagram below shows the sequence of operations followed by the proposed system beginning with data loading and preprocessing, then training and evaluating the model with an output of the real-time prediction functionality through an interactive widget interface.

We aim to carry out a project, which is based on machine learning, where we deploy not only an interactive but also cyberbullying detection system, increasing safety and inclusion online.

IV. REQUIREMENTS

4.1 Hardware Requirements

When it comes to hardware requirements, it is not that the code for sentiment analysis and cyberbullying detection powered by Twitter data is a recalcitrant demand. Here are the recommended hardware specifications:

4.1.1 CPU (Central Processing Unit):

• Use any contemporary multi-core CPU for handling the code execution process. A minimum of CPU having 2+ cores and more than 2.0 GHz clock speed is the bare minimum.

• Cases like Intel Core i3, AMD Ryzen 3, and their good equivalents can be provided.

4.1.2 RAM (Random Access Memory):

• It is recommended to have a minimum of 4GB of RAM to be able to provide efficient data loading, preprocessing and training tasks that require a lot of processing power.

• Rapid operations without stall, particularly when the data sets grow large are possible with 8GB of RAM and up.

4.1.3 Storage:

• Data storage for train.csv and test.csv as well as any other working files should be provided as Twitter split is done.

• All for free disk space must have 10 GB or more to ensure that the datasets and other files are kept safely.

The code provided here is runnable on the machine of a common user with a standard desktop or laptop computer, which lowers the hardware requirements it imposes. The code is constructed in such a manner and it shouldn't pose any issues with most present-day systems configuration.

4.2 Software Requirements

Here are the software requirements:

4.2.1 Operating System: Windows 10, macOS, or Linux

4.2.2 Integrated Development Environment (IDE): Jupyter Notebook or any Python-compatible IDE

4.2.3 Python Programming Language: Python is easy to understand and read in the code, and that's the reason why it offers complete libraries for fields such as web development, data analysis, and machine learning.

4.2.4 Data Processing Libraries:

4.2.4.1 Pandas: CSV file data processing and querying is used for this purpose too.

4.2.4.2 Regular Expressions (re): Depending on phase 1, this system utilizes its power for text preprocessing.

4.2.4.3 ipywidgets: Allows you to make use of enhancements such as the use of the interactive widgets. Use our AI to write for you about personal health management.

4.2.4.4 IPython.display: Occupies the functions, among others, of different output items.

4.2.4.5 scikit-learn (sklearn): Makes machine learning machine readability tools such as text vectorization, training, and evaluation available.

4.2.4.6 Seaborn and Matplotlib: It is broadly used for data visualization.

4.2.4.7 WordCloud: Text processing tool that segments word clouds from text data.

4.2.4.8 Collections.Counter: This function is responsible for counting the number of times an element occurs in the iterable.

These libraries do all the data/data loading and processing work, visualization, analysis, and classification making it easy to use and increases analysis effectiveness.

4.2.5 Data Requirements:

To efficiently and properly train and evaluate the cyberbullying detection system, Twitter datasets with labelled tweets including both cyberbullying and non-cyberbullying are required. Usually provided in CSV format (for example, train.csv and test.csv), these datasets allow data scientists to build models and test their performance. Every dataset should have the tweet text alongside labels that signifies whether a tweet is labeled as cyberbullying or not cyberbullying. Ensuring the existence of large datasets that are correctly labeled is very essential so as to build a machine learning model that can identify accurately cyberbullying behavior on social media platforms like Twitter.

5 RESULTS

The report details a cyberbullying detection via Twitter data elapsed utilizing data processing approaches and machine learning algorithms subsequently going through visualization methods. The report is started with a description of cyberbullying and the fact that nowadays it's mostly widespread through social media like Twitter. Highlights the role of developing the appropriate techniques of detection that are meant to reduce the degrading effects of cyberbullying on individuals and communities.

The research study depicts the methodology adopted for data collection, preprocessing, and analysis. Twitter data that covers both cyberbullying and non- cyberbullying tweets is collected and merged into one dataset through a combination of means. This dataset is preprocessed for handling of missing values as well as extraction of essential features for classification. Text preprocessing procedures, such as tokenization and cleansing, are implemented to guarantee that textual data can be trusted.

The process of data visualization is highly relevant for understanding the essence of the data and getting acquaintance with any manifested patterns. The use of visualization tools like the count plot, histogram, word cloud, and box plot are to show the distribution of cyberbullying labels, tweet lengths and words. These charts show that cyberbullying content on Twitter has its main features and help us understand what the most important ones are for classifying.

The SVM classification task belongs to machine learning algorithms, with SVM having a priority. The TF-IDF (Term Frequency-Inverse Document Frequency) method is utilized here for vectorization to convert textual data into numerical features fit for training the classifier. The model SVM is trained on the vectors data in order to classify tweets as cyberbullying and non-cyberbullying.

The measures of evaluation like accuracy, precision, recall and F1-score are the metrics to judge the success level of the trained model. We will use confusion matrices and ROC curves to better interpret how our model has performed in regard to true positive rate, false positive rate and area under the curve (AUC). The metrics that are provided in this context gives us a complete knowledge about how great is the accuracy of a model in the classification of cyberbullying Twitter posts.

Each sentence explains a certain tweet categorized as cyberbullying and not cyberbullying, demonstrating how the model arrives at its conclusion, providing context to this process. Snapshots of the particular tweets are also in the report for evidentiary purposes to show how the model did.

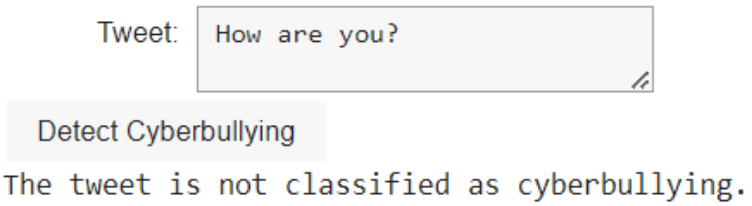

Figure 15: Text not classified as Cyberbullying

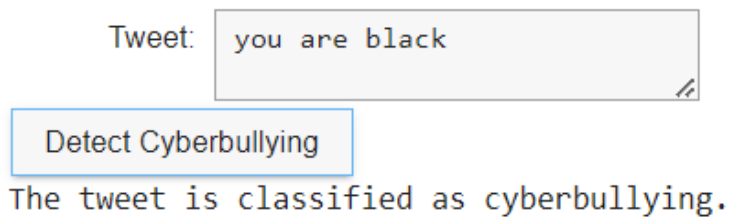

Figure 16: Text classified as Cyberbullying

The project results lay the groundwork for cyberbullying detection on Twitter, which is basically a complex data mining process of data preprocessing, visualization, machine learning modeling and performance evaluation. It generates useful results in terms of measuring the outcome of the classification method that has been created plus bulldog the need for applying machine learning algorithms for fighting against cyberbullying in online social networks.

VI. CONCLUSION AND FUTURE WORK

6.1 Conclusion

In brief, the Creation of a machine learning-based detection system for cyber bullying on Twitter has proved to be successful. Through our technique we could create a system, which is a type of interactive machine learning one capable of real-time tweet classification as cyberbullying versus non-cyberbullying action. Our model using Support Vector Machines (SVM) as a classifier with TF-IDF vectorization shows the ability to surprisingly high accuracy identifies cyberbullying behavior, which is important for creating a safe websites and applications.

We were incredibly fortunate, being happy with many different factors involved. Diverse data collection was the key in training the model. Our model has been trained on different data; the diverse data conveys subtleties of cyberbullying language. the utilization of sophisticated natural language processing functions, particularly those entailing TF-IDF vectorization, by the model helped to discriminate cyberbullying content from benign messages.

Our integrated platform creates an atmosphere freedom and accountability among online community members where cyberbullying victims can respond boldly to misbehavior.

Then, with the aim for the further enhancement of the model, rewriting the algorithm, applying advanced methods of sentiment analysis, and addressing ethical issues and biases is an area of focus ahead. In general, our work with security and the P-Club constitutes a progress towards a safe and considerate online environment.

6.2 Future Work

The completion of this project opens avenues for further research and development, including: The completion of this project opens avenues for further research and development, including:

6.2.1 Advanced Natural Language Processing Techniques: Discussing complex deep learning architectures and pre-trained language models can develop in system comprehension of difficult textual data. Such improvements will enable the system better cybercasting detection.

6.2.2 Multimodal Analysis: Integrating features from a swath of modalities alongside the text data may enrich the context of cyberbullying detection, and thus, lead to the high precision and robustness.

6.2.3 Contextual Understanding: One of the ways to improve the system's ability continue reading to detect subtle cases of cyberbullying may probably include contextual information from user profiles and social network dynamics.

6.2.4 Fairness and Bias Mitigation: Elimination of biases in cyberbullying algorithm is a considerable issue for normalization of internet platforms, namely fairness and equality.

6.2.5 User Engagement and Education: Placing educators and advocacy groups on the same team in the fight to end cyberbullying and promote digital citizenship will skillfully accomplish the goal of nurturing positive interactions in the digital world.

Under the umbrella of interdisciplinary working and inclusion, we can jointly ensure a more secure, healthier and more inclusive digital space for users.

REFERENCES

[1] Abutorab J S, Wagh R B, Gaikwad V S, Sonawane U D & Waghmare A I. (2022). Detection of Cyberbullying on Social Media using Machine Learning. International Research Journal of Modernization in Engineering Technology and Science, 04(05), 4526-4534.

[2] Desai A, Kalaskar S, Kumbhar O, & Dhumal R. (2021). Cyber Bullying Detection on Social Media using Machine Learning. ITM Web of Conferences, 40, 03038. International Conference on Automation, Computing and Communication 2021 (ICACC-2021).

[3] Nektaria Potha and Manolis Maragoudakis. Cyberbullying detection using time series modeling. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, pages 373–382. IEEE, 2014.

[4] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp.759–760.

[5] Kelly Reynolds, April Kontostathis, Lynne Edwards, "Using Machine Learning to Detect Cyberbullying", 2011 10th International Conference on Machine Learning and Applications volume 2, pages 241 –244. IEEE, 2011

[6] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbul- lying detection," InternationalJournal of Advanced Computer Science and Applications, vol. 9, no. 9, 2018.

[7] N. Majumder, A. Gelbukh, I. P. Nacional, and E. Cambria, "Deep learning-based document modeling for personality detection from text," IEEE Intell. Syst., vol. 32, no. 2, pp. 74–79, 2017.

[8] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. "Detecting offensive language in social media to protect adolescent online safety". In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71–80. IEEE, 2012

[9] Walisa Romsaiyud, Kodchakorn na Nakornphanom, Pimpaka Prasert- silp, Piyaporn Nurarak, and Pirom Konglerd, "Automated cyberbullying detection using clustering appearance patterns", In Knowledge and Smart Technology (KST), 2017 9th International Conference on, pages 242–247. IEEE, 2017

[10] Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of researchon cyberbullying victimization. Computers in Human Behavior 26, 277–287.

[11] Mason, K. L. (2008). Cyberbullying: A preliminary assessment for school personnel. Psychology in the Schools, 45(4), 323-348.

[12] Hinduja, S., & Patchin, J. W. (2008). Cyberbullying: An exploratory analysis of factors related to offending and victimization. Deviant Behavior, 29(2), 129-156.

[13] Hinduja, S., & Patchin, J. W. (2010). Cyberbullying: A review of the legal Issues facing educators. Part of a special issue: Cyberbullying: Preventing School Failure, 55(2), 71-78.

[14] Hinduja, S., & Patchin, J. W. (2011). High-tech cruelty. Educational Leadership, 68(5), 48-52.

[15] Hoff, D. L., & Mitchell, S. N. (2009). Cyberbullying: Causes, effects, and remedies. Journal of Educational Administration, 47(5), 652-665.