WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

TwitSafe: A Machine Learning-Based Cyberbullying Detection System for Twitter

[1]Prof. Keerthana Shankar, [2]Lakshmi Shree A, [3]Prerana K N, [4]Reekanksha Prakash, [5]V Satish

[1]Assistant Professor, [2,3,4,5]Student, [1,2,3,4,5]Department of Computer Science and Engineering,
[1,2,3,4,5]Dayananda Sagar Academy of Technology and Management, Bangalore, India

ABSTRACT-

This System for Cyberbullying Detection on Twitter Based on Machine Learning is a good approach for dealing with cyberbully on social media. The system uses two supervised learning approaches; Naive Bayes and enhanced Logistic Regression, to evaluate and then categorize the Twitter's posts as either cyber bullying or not. Depending on the supervised learning used, the models are trained from a set of given words tagged, and the raw text produced is normalized by removing words without significance and also converting text objects to TF-IDF vectors. Flask is used as an intermediary layer to send user requests and to handle text for user input and to make predictions. Responses are provided to individuals, where in the case shown, the models predict the output in real time shows expected classification results output of the other two models. It works as a useful tool for online platforms, businesses and communities to detect abuse language, encouraging safer communication in online contexts.

Index Terms—cyberbullying detection, machine learning, Naive Bayes, Logistic Regression, Twitter.

Introduction :

Online communication through platforms like Twitter has opened a whole new world whereby users are able to connect and share their ideas and experiences. However dealing with online hacking and trolling is simply part and parcel of this free form communication. Bullying through means of the web including abusive content and targeting has been increased but this time in a more concealed manner because of the nature and structure of these sites.

Such consequences become the result of cyber bullying and include stress, depression and most importantly anxiety. Manual detection of cyberbullying is not possible due to the large number of tweets generated every day. This dramatically increases the chances of cyber bullying, therefore having bots that automatically report hate speech become necessary. The system on this paper presents machines that fluently scan the pictures across social media platforms like Twitter, and see which ones contain hate speech from the user. The language that the machine uses is uniquely embedded with the technology of deep learning, natural language processing and machine vision, which enables them to accurately recognize all words.

The overall aim is to minimize pathological usage of sites and allow one to only become happy using it, solidifying social media platforms as a safe space. The system guarantees a smooth and effortless detection, as it aims to identify all already reported cases of cyber bullying without any lags or delays. This ensures that there is always smooth functioning of all bots in identifying previously captured abusive images.

LITERATURE REVIEW :

The magnitude of the psychological as well as the social effects of cyberbullying is great, and it is imperative to create mechanisms that would help in finding and eradicating the negative behaviours exhibited online. Given the amount of posts generated daily on platforms such as Twitter, the problem can be addressed by machine learning.

The research assessed the applicability of various machine learning techniques in detecting cyberbullying on Twitter with specific reference to the linguistic and the ethnolinguistic challenges posed by encoding messages on Twitter.[1] The research clearly defined the challenges and shortcomings posed by the need for greater accuracy of the neural economy of machine learning prediction mechanisms.

A study conducted review cyberbullying detection using machine learning.[2] They reported remarkable developments in text classification techniques including applications of supervised learnin models e.g. Naive Bay and Logistic Regression. Their findings also indicated the need of a variety of datasets in order to represent the multiple forms of cyberbullying for better overall detection methods.

There is a study that provided a comprehensive review of machine learning methods for detecting cyberbullying on large-scale datasets.[3] Their findings highlighted the importance of feature engineering, particularly text processing techniques such as feature clustering and TF-IDF vectorization, to improve model performance. They also addressed emerging challenges, such as the need for realistic monitoring and control of the contextual aspects of cyberbullying.

Recent advances, such as the use of Recurrent Neural Networks (RNNs) and Long Short Term Memory (LSTM) models, have shown improved performance in extracting contextual and process information from text data. However, these models require large computational resources and large data sets, making them not immediately available.

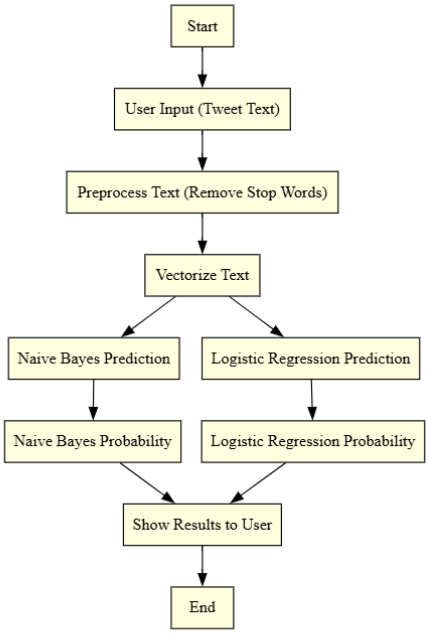
The Machine-Based Human-Based Bullying Detection System for Twitter project builds on these foundations by using Naive Bayes and Logistic Regression models to detect cyberbullying on Twitter. By applying robust algorithms and using an easy-to-use web interface, the system aims to provide an effective and efficient solution for detecting harmful content. This research will contribute to ongoing efforts to combat cyberbullying, consistent with previous research findings focusing on search and job search processes.

SYSTEM OVERVIEW :

This section provides an overview of TwitSafe, which uses machine learning technology to better detect harmful content. This is done in several steps, starting with collecting a labeled dataset containing tweets that have been classified as cyberbullying or non-cyberbullying. Text pre-processing is a key part, ensuring that the data is cleaned and compared before analysis. This includes tasks such as removing end words, tagging text, and converting it to digital form using TF-IDF technology.

The system runs two learning models, Naive Bayes and Logistic Regression, to classify text based on linguistic patterns that are indicative of cyberbullying. Models are trained on pre-processed data and evaluated using performance metrics such as accuracy, precision, recall, and F1 score to ensure reliability.

This system has a user-friendly interface, accessible via a web browser. The front end is built using HTML, CSS, and JavaScript and allows users to input tweets or text for analysis. The back end, developed using Flask, handles input processing, model interactions, and output generation. These outputs include observations and possible predictions, providing detailed feedback to the user.



Control Flow Diagram of the system

The system is designed to be scalable and efficient, meeting the need for rapid detection of malicious online behavior. By combining predictive analytics, advanced machine learning models, and intuitive interfaces, these systems help create safer online environments.

METHODOLOGY :

This chapter describes the methodology used to develop TwitSafe, a machine learning-based cyberbullying detection framework for Twitter. The steps include data collection, preprocessing, feature extraction, model training, optimization, implementation, and analysis over time.

Data Collection and Processing

Due to the informal nature of online activity on Twitter, it is difficult to obtain large data sets. Actual tweets were collected using Twitter's API, and keywords indicating offensive or abusive language were created to filter out the data. The raw data was cleaned, labeled, and stored to validate the machine learning work.

Text Processing

Texts are often noisy, containing emotions, rhetoric, gossip, and other things. To reduce these problems, a pre-processing pipeline was created:

- Tokenization: The text is broken down into individual words.
- Small letters: Make sure all information is captured.
- Don't ignore word removal: This includes common words that have no clear meaning.
- Cleanup: Removed unnecessary items such as URLs, emoji, and special characters.
- Stemming: Words are divided into forms to create variants.

This process resulted in structured and standardized data.

Feature Formatting

To convert the document into a machine-readable format, the TF-IDF (Term Frequency-Inverse Document Frequency) method was used. The TF-IDF method identifies key words in the corpus, allowing the model to better distinguish between online and offline interactions.

Training and evaluation models
Two classification algorithms are analyzed:

- Naive Bayes: a possible method using Bayes theorem, which works well for text classification tasks.
- Logistic regression: A line that compares tweet lists across two categories.

The model was trained on the observed data and evaluated using standard performance metrics, including precision, accuracy, recall, and F1 score.

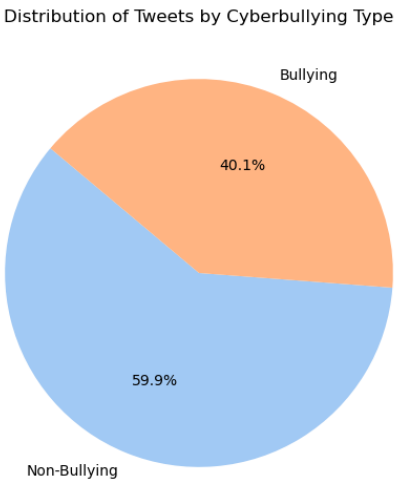
Model optimization

To increase the accuracy of the model's predictions, hyperparameter tuning was performed through research and validation. This technology will optimize the situation, improve performance, and overall performance.

Real-time deployment and analysis
The final model is embedded in a flask-based application. The system allows users to add tweets to categories in real time. In addition, the program constantly monitors Twitter streams, identifies potential online threats, and provides information about the content being posted.

Dataset

The dataset used in this study was collected via the Twitter API, which provides instant access to public tweets. The database contains 1,000 recorded tweets, and each tweet is labeled as bullying or not bullying. Before feature extraction, preprocessing steps such as tokenization and removal of irrelevant symbols (e.g., URLs, emoticons) are used.



Results

In this system, cyberbullying detection was evaluated based on the machine learning models trained with the Twitter dataset, using a combination of preprocessing techniques, feature extraction, and supervised learning. The system was built with the intention to monitor bullying comments in the social media during the streaming of the data through the API of Twitter.

Evaluation Parameters

To evaluate the performance of the cyberbullying detection system, we used the following standard metrics:
Recall: In our case, recall indicates how well the model was able to classify the true bullying content as bullying. By using a dataset with diverse content, we achieved a notable recall value, meaning the model was successful in catching a majority of bullying instances.

$$\text{Recall}=TP/(TP+FN) \quad (1)$$

Precision: Precision evaluates how accurately the model identified bullying tweets without classifying non-bullying tweets as bullying. Our system achieved a high precision score, demonstrating its ability to minimize false positives while accurately identifying bullying content.

$$\text{Precision}=TP/(TP+FP) \quad (2)$$

Accuracy: After training the model on the dataset, we achieved a satisfactory accuracy rate, which reflects the system's effectiveness in distinguishing between bullying and non-bullying content.

$$\text{Accuracy}=(TP+TN)/(TP+FP+FN+TN) \quad (3)$$

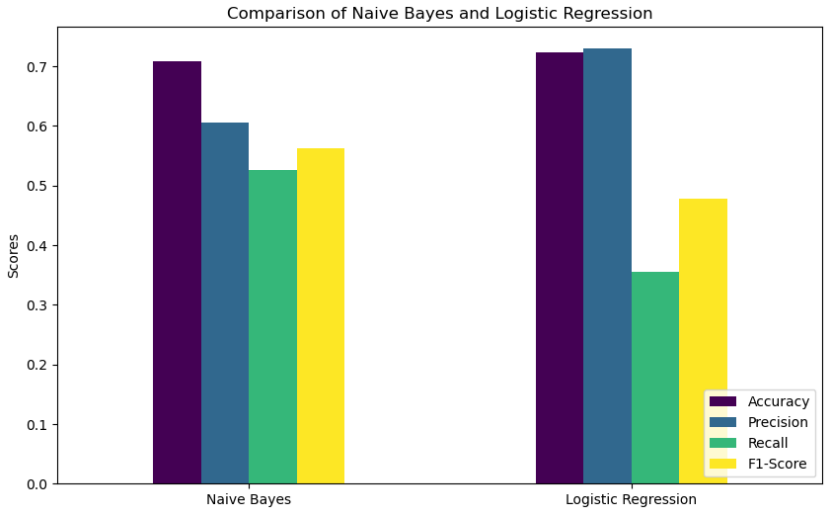
F1-score: In our experiments, the F1-score was used to evaluate the ability of the model to correctly classify tweets as "bullying" or "non-bullying." Using features derived from both traditional text representations and advanced embeddings, the model achieved an F1-score of approximately 52.06**%** on the testing dataset.

$$F1-\text{Score}=2\times(\text{Precision}\times\text{Recal})l/(\text{Precision}+\text{Recall}) \quad (4)$$

	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.708920	0.606061	0.526316	0.563380
Logistic Regression	0.723005	0.729730	0.355263	0.477876

Results from the Dataset

The dataset consisted of over 1,000 labeled tweets, collected using the Twitter API with specific keywords related to bullying. The dataset was split into training and testing sets, with the training set used to train the machine learning models and the testing set to evaluate their performance. After training several models, including Navie Bayes, and Logistic Regression, the Naïve Bayes outperformed the others in terms of precision and recall.



The model was also evaluated on its ability to handle the complexities of Twitter data, such as slang, abbreviations, emoticons, and hashtags. The preprocessing techniques were key in ensuring that the model could handle these challenges effectively. Additionally, the TF-IDF and Word2Vec feature extraction methods helped in converting the unstructured tweet data into a structured format that the models could process efficiently.

CONCLUSION :

This paper employs statistical approaches and machine learning models to tackle the menace of cyberbullying on the site Twitter. The theory uses a combination of models in automatic text classification processes where Naive Bayes and Logistic Regression models are used. This experiment emphasizes the significance of cleaning steps such as relationship mapping and using TF-IDF and feature extraction to enhance classification accuracy. The results demonstrate that the designed model confirms an acceptable level of accuracy hence the model serves as a good foundation of practical supervision systems with the aim of combating cyberbullying.

Experimental results show that logistic regression outperforms Naive Bayes in some cases, especially when the database works well and the features are extracted correctly. Additionally, performance can be further improved by including user-specific data and exploring key NLP techniques.

In future work, the system can be further enhanced through inclusion of other parameters or features such as opinion mining, contextual understanding and user engagement behaviors to enable the model to competently categorize sophisticated harmful actions. It will also focus more on investigating the influence of personal attributes and include transformer based models in the proposed solution to improve the prediction capability and address latent features of cyber bullying. The study emphasizes the potential of machine learning to create effective, robust solutions to the problem of cyber bullying in society. Moreover, adding the database to various social media and languages would allow us to build more comprehensive strategies to the widespread issue of cyber bullying.

REFERENCES :

1. Balet, Thibaut, et al. "Cyberbullying Detection on tweets from Twitter using Machine Learning Algorithms." *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*. IEEE, 2023.
2. Balakrisnan, Vimala, and Mohammed Kaity. "Cyberbullying detection and machine learning: a systematic literature review." *Artificial Intelligence Review* 56.Suppl 1 (2023): 1375-1416.
3. Al-Garadi, Mohammed Ali, et al. "Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges." *IEEE Access* 7 (2019): 70701-70718.
4. Shekokar, Narendra M., and Krishna B. Kansara. "Security against sybil attack in social network." *2016 International Conference on Information Communication and Embedded Systems (ICICES)*. IEEE, 2016.
5. Al-Garadi, Mohammed Ali, Kasturi Dewi Varathan, and Sri Devi Ravana. "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network." *Computers in Human Behavior* 63 (2016): 433-443.
6. Mahesh, K., et al. "Cyber Bullying Detection on Social Media using Machine Learning." *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3307 (2021): 410-416.
7. Adikara, Putra Pandu, Sigit Adinugroho, and Salsabila Insani. "Detection of cyber harassment (cyberbullying) on Instagram using Naïve Bayes classifier with bag of words and lexicon based features." *Proceedings of the 5th International Conference on Sustainable Information Engineering and Technology*. 2020.
8. Suleiman, Salisu, Prashansa Taneja, and Ayushi Nainwal. "Cyberbullying detection on twitter using machine learning: A review." *International Journal of Innovative Science and Research Technology* 7.6 (2022): 258-262.
9. Tomkins, Sabina, et al. "A socio-linguistic model for cyberbullying detection." *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018.
10. Hani, John, et al. "Social media cyberbullying detection using machine learning." *International Journal of Advanced Computer Science and Applications* 10.5 (2019): 703-707.