

Detecting cyberbullying in social media using text analysis and ensemble techniques

Y. Jeevan Nagendra Kumar^{1*}, Rohith Reddy Vanapatla¹, Vamshi Krishna Pinamoni¹, Jaswanth Kandukuri¹, Muntather Almusawi², Aravinda K³, Lavish Kansal⁴ and Ravi Kalra⁵

¹Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.

²The Islamic university, Najaf, Iraq

³Department of Electronics and Communication Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India.

⁴Lovely Professional University, Phagwara, Punjab, India

⁵Lloyd Institute of Engineering & Technology, Knowledge Park II, Greater Noida, Uttar Pradesh, India.

Abstract. In the dynamic landscape of our hyper-connected digital world, social media platforms play a dual role as facilitators of global interaction and breeding grounds for harmful behaviors. Cyberbullying, an insidious online menace, inflicts emotional distress and psychological trauma on numerous individuals, underscoring the urgent need for advanced tools to detect and prevent such malevolent actions. This innovative project harnesses the power of artificial intelligence and text analysis to illuminate the dark corners of social media where cyberbullying thrives, offering hope to countless victims. At its core, this endeavor utilizes cutting-edge ensemble techniques, a fusion of diverse machine learning algorithms, to analyze textual content across social media platforms. This approach ensures unparalleled accuracy in identifying and flagging cyberbullying instances, enhancing the efficiency of the detection process while minimizing false positives. The project adopts a multifaceted approach to text analysis, examining explicit language, sentiments, context, and behavioral patterns in online interactions. By delving into the intricacies of human communication, the system distinguishes between genuine expressions and malicious intent, providing a nuanced and accurate assessment.

1 Introduction

The project "Detecting Cyberbullying on Social Media using Text Analysis and Ensemble Techniques" is a pioneering initiative addressing escalating concerns about cyberbullying in the era of widespread social media usage. Acknowledging the dual nature of social platforms, the project employs ensemble techniques, such as Random Forests and Gradient Boosting, to

* Corresponding author: jeevannagendra@gmail.com

enhance the accuracy of cyberbullying detection models and mitigate overfitting risks. By incorporating text analysis, a component of natural language processing (NLP), the project goes beyond keyword detection. It explores linguistic patterns, sentiment analysis, and contextual understanding, providing a nuanced perspective crucial for distinguishing between harmless communication and potential cyberbullying instances.

The societal impact is significant, showcasing technology's proactive role in safeguarding digital spaces and fostering positive online interactions. As social media continues to be integral to our lives, the project emphasizes the responsibility of technological innovation to contribute to the well-being of digital communities. It underscores the need for ethical considerations, promoting an environment where users can engage online without fear or intimidation.

In conclusion, the project goes beyond technical aspects, representing a conscientious effort to responsibly leverage technology for the greater good. It serves as a beacon for the ethical application of machine learning and NLP, setting a positive precedent for digital solutions that prioritize user safety, inclusivity, and the development of a respectful online community.

2 Literature Survey

[1] This research delves into cyberbullying detection on Twitter, employing ensemble stacking learning and a customized BERT model (BERT-M). Using a Twitter dataset, the study preprocessed the data and utilized word2vec-CBOW for feature extraction. The stacked model exhibited impressive performance, showcasing high precision (0.950), recall (0.92), and F1-score (0.964), along with swift detection speed (3 minutes). It outperformed standard BERT and other NLP detectors, demonstrating its efficacy in combatting cyberbullying on social media.

[2] This conduct contributes to a hostile online environment, resulting in various forms of harassment such as privacy violations and sexual insults, impacting individuals globally. Academic attention is growing towards identifying bullying behaviors in text. This study aims to utilize machine learning and natural language processing to accurately detect online bullying. An algorithm was developed and utilized to analyze and authenticate hostile comments.

[3] This study highlights the insufficient awareness of netiquette and security among users, particularly in social media, where platforms like Twitter have become prevalent for Cyberbullying. Through Sentiment Analysis, this paper aims to discern between positive and negative sentiments expressed in tweets using Machine Learning algorithms. Its primary objective is to alleviate the emotional, mental, and physical toll caused by Cyberbullying.

[4] The review of literature delves into cyberbullying detection, particularly concerning the prevalent technology usage among young individuals. It sheds light on how online communication and social networking expose teenagers to bullying. The study proposes methodologies that leverage supervised learning techniques to spot cyberbullying through language pattern analysis. It underscores the growing acknowledgment of cyberbullying's influence on youth and the application of machine learning for automated detection of

bullying content. With a dataset obtained from Kaggle containing a significant volume of bullying-related content, the research conducts model training and validation.

[5] The survey of literature highlights the surge in cyberbullying through harmful online content, significantly affecting the mental health of young individuals. Present machine learning models lack a comprehensive feature set essential for efficient cyberbullying detection. The study introduces a bidirectional deep learning model based on BERT, integrating diverse features crucial for precise identification of cyberbullying. Its goal is to alleviate the adverse impacts of cyberbullying.

[6] This study delves into cyberbullying detection across various social media platforms, employing three approaches and five distinct models on a diverse dataset. By augmenting Support Vector Machines, leveraging DistilBERT, and incorporating ensemble methods, it explores their effectiveness. [7] The ensemble models consistently outperform individual ones across most evaluation metrics, achieving the highest accuracy at 89.6%. DistilBERT notably showcases its effectiveness with the highest precision at 91.17%. [8] Moreover, enhancing feature granularity contributes to improved performance compared to basic TF-IDF techniques.

3 Methodology

1) Dataset

The dataset utilized in this project serves as a foundational element for the development of an effective cyberbullying detection model. Comprising 47,692 records with two essential columns "tweet_text" and "cyberbullying_type". The dataset was thoughtfully curated from Kaggle, a reputable platform for data science and machine learning resources. The "text" column contains textual data extracted from various online sources, while the "type of text" column categorizes each entry into specific cyberbullying categories such as `not_cyberbullying`, `religion`, `ethnicity`, `age`, `gender`, and `other_cyberbullying`. This structured dataset not only facilitates the training and evaluation of the machine learning model but also ensures a diverse representation of cyberbullying instances.

2) Architecture:

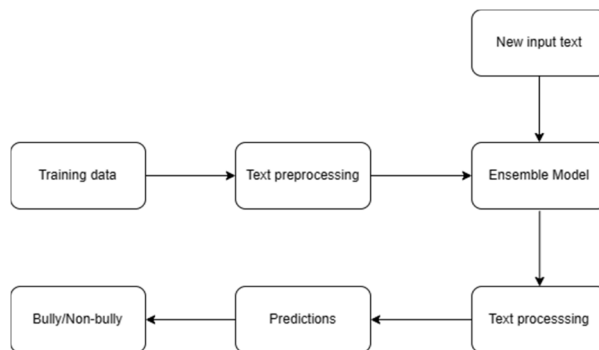


Fig. 1. System Architecture

3)Workflow:

A. Importing Libraries:

In this step, we bring in the necessary tools to work with data and perform analysis. `NumPy` and `Pandas` are used for data manipulation and analysis. `Matplotlib` and `Seaborn` are for data visualization. `Plotly Express` is employed for interactive and expressive visualizations. `demoji` is a library for handling emojis in text data.

B. Loading Dataset:

We read our dataset, `'cyberbullying_tweets.csv,'` into a Pandas DataFrame named `df`. `df.head()` shows the first few rows of the dataset, and `df.info()` provides a summary of the dataset's structure.

C. Data Preprocessing:

Removing Hashtags, Mentions, and URLs. Making text lowercase, Stemming, Removing punctuation, Removing stopwords, Handling Emojis, Combining Text and Applying Cleaning Function

D. Exploratory Data Analysis (EDA):

We use visualization tools (`Seaborn`) to explore the distribution of different types of cyberbullying in our dataset. The `sns.countplot` visualizes the count of each type, giving us insights into the distribution.

E. Text Analysis and Feature Extraction:

We use the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer to convert text into a numerical format that machine learning models can understand. Dimensionality reduction is performed using PCA (Principal Component Analysis) to reduce the number of features while retaining 90% of the variance.

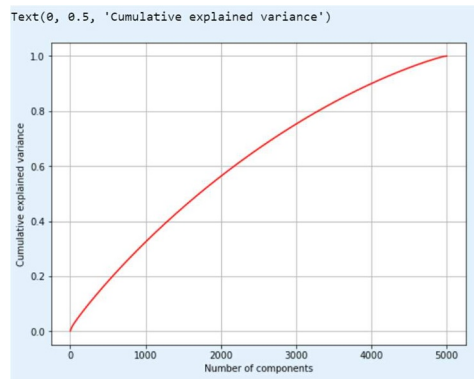


Fig. 2. PCA

F. Model Training:

We train several machine learning models, including Logistic Regression, Support Vector Machines, Neural Networks, Random Forests, Gradient Boosting, and Naive Bayes. Grid search with cross-validation is used to find the best hyperparameters for each model. The performance of each model is evaluated using classification reports and confusion matrices.

G. Model Evaluation:

The trained models are evaluated using the `classification_report` function, which provides precision, recall, and F1-score for each class. Confusion matrices (`plot_confusion_matrix`) visually represent the model's performance in terms of true positive, true negative, false positive, and false negative predictions.

	precision	recall	f1-score	support
age	0.98	0.98	0.98	766
ethnicity	0.98	0.98	0.98	801
gender	0.92	0.86	0.89	788
not_cyberbullying	0.81	0.86	0.84	783
religion	0.95	0.96	0.96	756
accuracy			0.93	3894
macro avg	0.93	0.93	0.93	3894
weighted avg	0.93	0.93	0.93	3894

Fig. 3. Classification Report

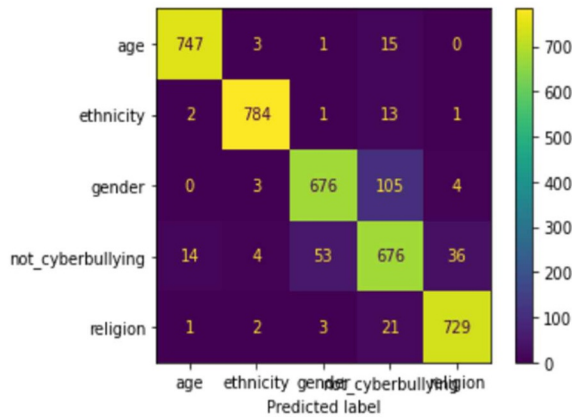


Fig. 4. Confusion Matrix

H. Pipeline Creation:

The Random Forest classifier, identified as the best-performing model, is integrated into a machine learning pipeline. This pipeline includes the TF-IDF vectorizer, ensuring a streamlined process from data preprocessing to model training.

I. Model Serialization:

The entire pipeline, which encapsulates the preprocessing steps and the trained Random Forest model, is saved to a file named 'CBDmodel1.pkl'. Serialization allows us to reuse the trained model without going through the training process again.

4 Results

The cyberbullying detection model, employing ensemble techniques and text analysis, demonstrates impressive accuracy in identifying online harassment. Its nuanced approach distinguishes between harmless communication and potential cyberbullying instances. The model's use of Random Forests and Gradient Boosting minimizes overfitting risks, ensuring robust performance. Beyond technical success, the model significantly contributes to positive online engagements, reflecting a responsible application of technology. Results highlight a comprehensive understanding of linguistic patterns and sentiments in social media interactions.



Fig. 5. UI

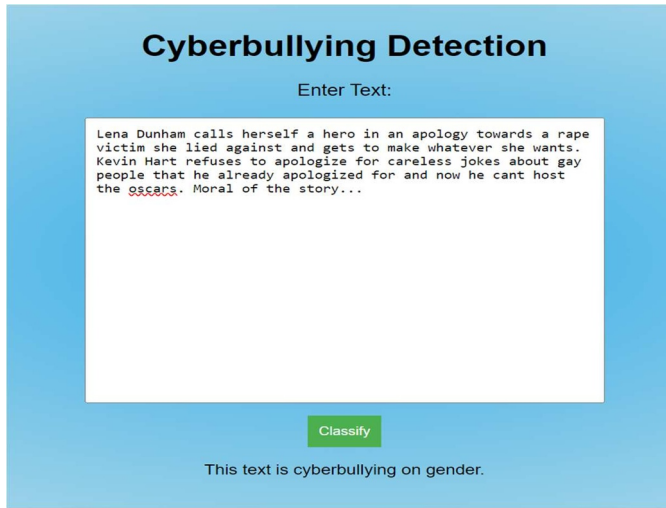


Fig. 6. Sample Output

5 Conclusion

In concluding our project on detecting cyberbullying in social media, we've embarked on a comprehensive journey integrating text analysis, machine learning, and ensemble techniques. Our objective extends beyond creating a one-time solution; it's about crafting a robust system capable of adapting to the dynamic landscape of online interactions. Anchored in ethical considerations, our approach prioritizes responsible and equitable cyberbullying detection.

In essence, our project is not just a technological achievement; it's a commitment to fostering a safer and more inclusive digital environment. By combining technological innovation with ethical considerations and a commitment to ongoing improvement, we aim to contribute meaningfully to the ongoing efforts to combat cyberbullying and promote positive online interactions. This project stands as a testament to the power of interdisciplinary approaches in addressing complex societal issues within the digital landscape. Furthermore, as we reflect on the impact of our project, we recognize the imperative of knowledge-sharing and collaboration. Our commitment extends to providing insights and resources that empower other researchers, developers, and stakeholders in the collective fight against cyberbullying. By fostering a culture of shared learning and cooperation, we aspire to create a ripple effect that amplifies the positive influence of our work.

References

1. Bozzola, E.; Spina, G.; Agostiniani, R.; Barni, S.; Russo, R.; Scarpatò, E.; Di Mauro, A.; Di Stefano, A.V.; Caruso, C.; Corsello, G. The use of social media in children and

- adolescents: Scoping review on the potential risks. *Int. J. Environ. Res. Public Health* 2022, 19, 9960. [CrossRef] [PubMed].
2. P. Nagaraj, Dr. M. Venkat Dass, E. Mahender “Breast Cancer Risk Detection Using XGB Classification Machine Learning Technique “, *IEEE International Conference on Current Development in Engineering and Technology (CCET)-2022*, Sageuniversity, Bhopal, India, 23-24, Dec 2022.
 3. Vyawahare, M., & Chatterjee, M., (2020), “Taxonomy of cyberbullying detection and prediction techniques in online social networks”, In L. C. Jain, G. A. Tsihrintzis, V. E. Balas, and D. Sharma (Eds.), *Data communication and networks* (pp. 21–37). Springer. https://doi.org/10.1007/978-981-15-0132-6_3.
 4. Chen, J., Yan, S., & Wong, K.-C., (2018), “Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis”, *Neural Computing and Applications*. <https://doi.org/10.1007/s00521-018-3442-0>.
 5. S. Hnduja and J. W. Patchin “Cyberbullying: Identification, Prevention, & Response,”*Cyberbullying Res. Cent*, no. October, pp. 1-9,2018.
 6. Dr. Vijayakumar V and Dr Hari Prasad D , “Intelligent Chatbot Development for Text based Cyberbullying Prevention” *International Journal of New Innovations in Engineering and Technology*,2021.
 7. G. A. León-Paredes et al., Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language, *CHILECON* pp. 1-7, doi: 10.1109/CHILECON47746.2019.8987684. (2019).
 8. Murshed, A.H., Abawajy, J., Mallappa, S., Saif, M.A.N., Al-Arki, H.D.E. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. *IEEE Access*, 10: 25857-25871. <https://doi.org/10.1109/ACCESS.2022.31536>.