Proceedings of the Third International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2023)
IEEE Xplore Part Number: CFP23K58-ART; ISBN: 979-8-3503-4363-2

2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) | 979-8-3503-4363-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIMIA60377.2023.10425920

Using NLP Techniques for Cyberbullying Tweet Recognition

[1]Arunkarthick A K*, [2]Beebi Naseeba, [3*] Nagendra Panini Challa, [4] Kommineni Ajay
[1, 2, 3, 4] School of Computer Science and Engineering (SCOPE), VIT-AP University, Andhra Pradesh, India.
[1]arunkarthick348@gmail.com, [2]beebi.naseeba@vitap.ac.in, [3*] nagendra.challa@vitap.ac.in,
[4]ajay.21bce7004@vitapstudent.ac.in
[3*]Corresponding Author: nagendra.challa@vitap.ac.in

***Abstract* – With rise in usage of internet, cyberbullying is becoming more and more widespread. Cyberbullying has emerged as a pervasive online threat, leading to serious psychological and social consequences for victims. Recognizing and mitigating cyberbullying is a critical societal concern. Cyberbullying has led to significant and devastating consequences, underscoring the urgent necessity for effective detection methods. This study aims to address this need through the application of sentiment analysis and employs a novel approach to leverage Natural Language Processing (NLP) and Machine Learning techniques and detect the instances of cyberbullying within Tweets. The proposed methodology involves processing individual tweets and identifying potential cyberbullying threats for further scrutiny and action.**

Index Terms – Cyberbullying, Sentiment Analysis, NLP (Natural Language Processing), Machine Learning.

I. INTRODUCTION

The rapid proliferation of social media platforms and online communication has fundamentally altered the manner in which we establish connections and engage with each other. While this digital age has ushered in countless opportunities for information sharing, networking, and community building, it has also brought to the forefront a dark and troubling phenomenon: cyberbullying. Cyberbullying, a form of harassment that occurs in digital spaces, has emerged as a pervasive and destructive issue, leaving victims emotionally scarred and socially isolated. As our lives become increasingly intertwined with the digital realm, the need for effective cyberbullying detection and intervention mechanisms becomes more urgent than ever before.

The effects of cyberbullying are significant and extend extensively. People subjected to cyberbullying experience not only psychological distress but also significant social and academic repercussions. The relentless nature of online harassment, often occurring in the form of hurtful messages, threats, or public shaming on social media platforms like Twitter, can lead to long-lasting emotional trauma. Moreover, the anonymity and accessibility to online platforms make cyberbullying a pervasive and insidious problem, impacting individuals of various age and backgrounds.

Recognizing the gravity of this issue, researchers and technologists have turned to artificial intelligence (AI) and machine learning as a potential solution. This research paper delves into the domain of cyberbullying tweet recognition, a critical aspect of identifying and mitigating online harassment. By leveraging the capabilities of Artificial Intelligence, particularly Natural Language Processing (NLP) and machine learning methodologies, we embark on a journey to develop an advanced cyberbullying detection system.

Natural Language Processing (NLP) plays a pivotal role in our research by enabling the analysis of the textual content of tweets. NLP techniques are employed to process and decipher the language used in online communications, allowing us to extract meaningful insights from the text. These insights include sentiment analysis to understand the emotional tone of the content, as well as the identification of linguistic markers associated with different types of cyberbullying. NLP techniques, such as lemmatization and acronym expansion, help us refine the text for analysis.

In this research, we present a comprehensive study that outlines our methodology, findings, and implications. Our goal is to shed light on the innovative techniques and strategies employed to recognize instances of cyberbullying within tweets, ultimately contributing to the creation of a safer and more inclusive online environment. We explore the intricacies of sentiment analysis, feature engineering, model training, and evaluation, all of which constitute the foundation of our cyberbullying detection system.

As we explore the nuances of this research, it becomes clear that the development of an effective cyberbullying tweet recognition system is not only a technological endeavor but also a moral imperative. Our work carries the potential to not only protect vulnerable individuals from the perils of online harassment but also to foster a culture of respect, empathy, and digital responsibility in the virtual spaces we inhabit.

In the pages that follow, we detail our methods, present our findings, and discuss the broader societal implications of our research. In the course of this exploration, our objective is to offer valuable insights and make a meaningful contribution

to the continuous endeavors aimed at countering cyberbullying in our progressively interconnected world.

II. LITERATURE REVIEW

Cyberbullying, a distressing manifestation of online aggression, has garnered increasing attention from researchers, educators, policymakers, and society at large. As online platforms continue to evolve and proliferate, the incidence of cyberbullying has risen, demanding robust solutions for its detection and prevention.

Prevalence and Impact of Cyberbullying

A foundational understanding of the prevalence and impact of cyberbullying is essential for framing the urgency of research in this domain. Multiple studies have investigated the prevalence rates of cyberbullying across various demographics and platforms.

- **Prevalence Rates:** Smith et al. (2008) highlighted that nearly one-third of adolescents in the United States have encountered cyberbullying in one form or another. Similarly, a study by Kowalski et al. (2014) emphasized the global nature of this issue, revealing that cyberbullying affects youth across diverse cultural contexts.
- **Psychological and Social Consequences:** Research consistently underscores the profound psychological and social consequences of cyberbullying. Victims often endure elevated levels of anxiety, stress, and depression (Patchin & Hinduja, 2010). What sets cyberbullying apart is its ability to extend beyond physical boundaries, subjecting victims to prolonged distress and feelings of vulnerability (Slonje & Smith, 2008).
-

Technological Approaches to Cyberbullying Detection

The advent of artificial intelligence (AI) and machine learning offers promising avenues for addressing the challenge of cyberbullying. Various technological approaches have been explored.

- **Sentiment Analysis and Natural Language Processing (NLP):** Sentiment analysis, a subset of NLP, has been a focal point of research. Scholars like Dinakar et al. (2012) have explored sentiment analysis as a means to detect offensive language and hurtful content within tweets, showcasing the potential of NLP techniques to discern the emotional tone and intent behind online messages.
- **Machine Learning Techniques:** Machine learning techniques, including deep neural networks and support vector machines, have gained traction in cyberbullying detection. For instance, Ribeiro et al. (2018) leveraged deep learning models to identify cyberbullying in online conversations, yielding promising results. Additionally, ensemble methods and transfer learning have been explored to enhance model accuracy and generalization (Yin et al., 2019).

III. PROPOSED ARCHITECTURE

The goal of our research is to develop an advanced system for recognizing cyberbullying tweets, contributing to the ongoing efforts to combat cyberbullying. Our approach builds upon the existing body of research in this domain, combining cutting-edge techniques in Machine learning and Natural Language Processing to create a robust and effective tool for identifying instances of cyberbullying within tweets.

1.Data Collection and Preprocessing: The research work begins with comprehensive data collection effort. The data obtained contains diverse and extensive tweets that contains many instances of cyberbullying. The dataset is also curated to ensure relevance, diversity and representativeness across various social media platforms. During data preprocessing, the data with noise, irrelevant information and personally identifiable information is removed in order to protect user privacy. A standard tweet consists of 140 characters or less, including text, emotions, symbols, and URLs. To optimize performance, we implement data cleaning and preprocessing steps. These measures entail the removal of symbols, URLs, and stop words, in addition to acronym expansion and lemmatization. Emoticons play a crucial role in sentiment analysis as they widely express opinions and emotions, and they should not be overlooked. Emoticons usually consist of symbols such as ')', ':', and '='. These symbols serve to strengthen the emotional tone of the text or convey elements of irony.

2. Feature Engineering: To empower our model to recognize cyberbullying effectively, we will employ advanced feature engineering techniques. We will extract a wide range of features from the tweet text, including linguistic patterns, sentiment analysis, user mentions, hashtags, and more. These features will provide a holistic understanding of the tweet content and context, facilitating accurate detection.

3. Exploratory Data Analysis for Multiclass Classification: In addition to cyberbullying detection, our research extends to categorizing cyberbullying into various types. We recognize that cyberbullying is not a one-size-fits-all phenomenon, and its manifestations can vary widely. Our exploratory data analysis aims to predict and classify tweets into five distinct types of cyberbullying: Gender-Based Cyberbullying: Tweets targeting individuals based on their gender or gender identity. Religion-Based Cyberbullying: Identifying instances where religion becomes a target of harassment. Age-Based Cyberbullying: Recognizing cyberbullying that focuses on a person's age or life stage. Ethnicity-Based Cyberbullying: Detecting tweets that engage in harassment based on a person's ethnicity or racial background. Other Types of Cyberbullying: This category encompasses cyberbullying that may not fit into the above categories but still constitutes online harassment.

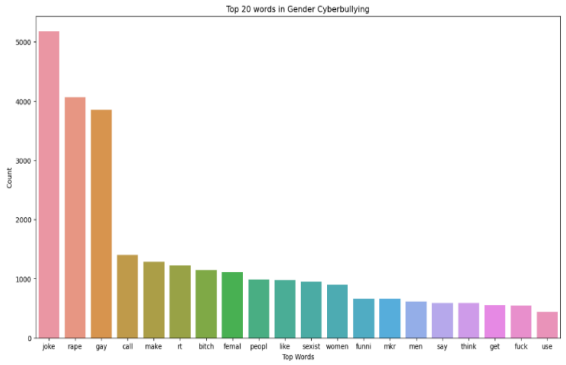

Fig1. Top words count of Gender based cyberbullying

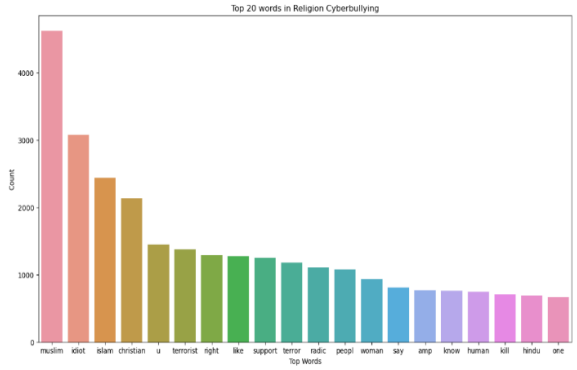

Fig2. Top words count of Religion based cyberbullying

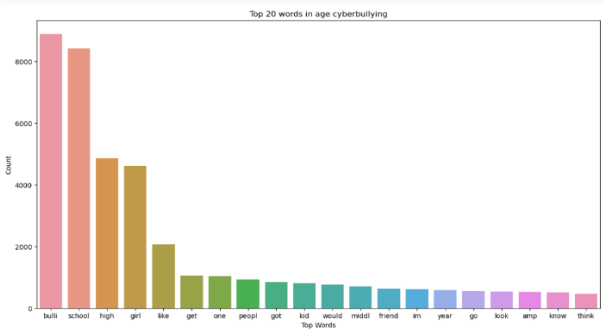

Fig3. Top words count of age-based cyberbullying

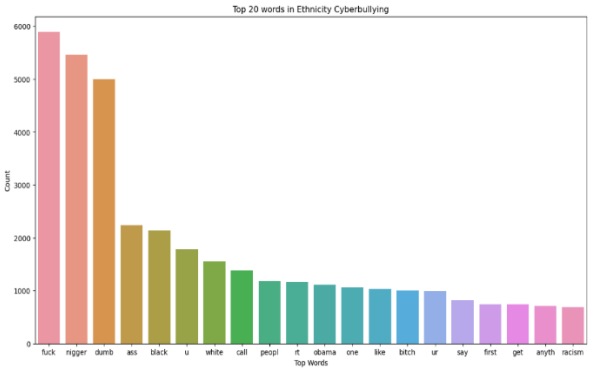

Fig4. Top words count of ethnicity based cyberbullying

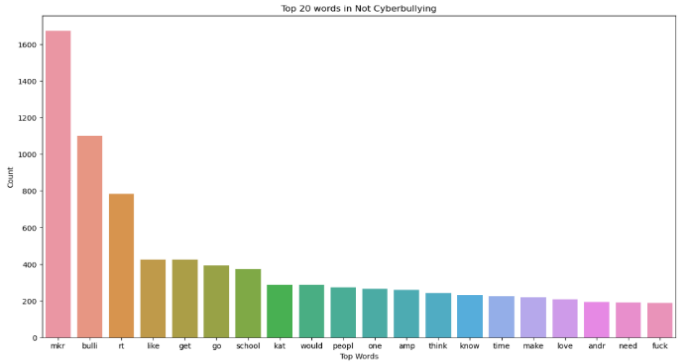

Fig5. Top words count of Other types of cyberbullying

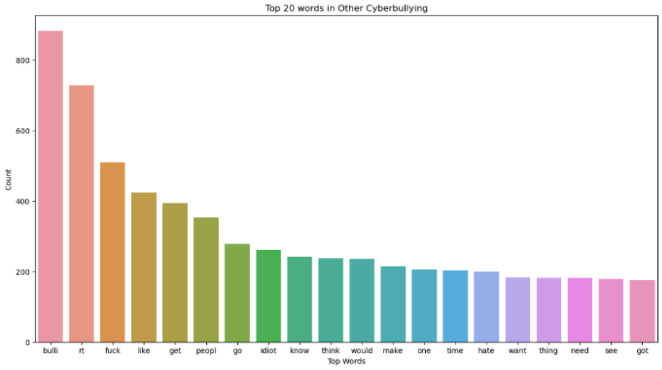

Fig6. Top words count of tweets without cyberbullying

Through exploratory data analysis and feature extraction techniques, we aim to identify patterns and linguistic markers associated with each type of cyberbullying. This nuanced approach allows us to not only recognize cyberbullying but also gain insights into the specific dimensions of online harassment that individuals may experience. Furthermore, our analysis includes the classification of tweets that do not exhibit cyberbullying behavior. By identifying and differentiating non-cyberbullying tweets, our goal is to establish a more comprehensive and precise cyberbullying detection system that reduces the occurrence of false positives and ensures a balanced approach to moderating online content. This multifaceted analysis will advance our comprehension of the intricate nature of cyberbullying and facilitate more precisely targeted interventions to combat its various manifestations.

4. Feature Representation with Bag of Words and TF-IDF Vectorization: To enhance the effectiveness of our cyberbullying tweet recognition system, we employ Bag of Words and Term Frequency-Inverse Document Frequency vectorization techniques in our architecture.

Bag of Words (BoW): BoW is a fundamental technique in natural language processing that transforms textual data into numerical feature vectors. In the context of our research, we utilize BoW to transform tweet text into

matrix representation, where each row corresponds to a tweet, and each column represents a distinct word or term from the entire dataset. This approach captures the word frequency within each tweet, supplying essential information for our machine learning models.

TF-IDF Vectorization: Term Frequency-Inverse Document Frequency (TF-IDF) is an alternate technique widely employed for extracting features from text. In contrast to BoW, TF-IDF takes into account not only the frequency of terms within a single tweet but also their significance in the entire dataset. TF-IDF assigns greater weights to terms that are frequent within a tweet but less common across the entire corpus, thereby emphasizing the unique words or phrases in each tweet. This allows us to capture the uniqueness of cyberbullying-related language and differentiate it from common language.

By incorporating both BoW and TF-IDF vectorization in our architecture, we enable our machine learning models to analyze tweet content more effectively. These vectorization techniques serve as a bridge between the raw text data and the models, providing a rich representation of the textual information. This approach enhances the accuracy and granularity of our cyberbullying tweet recognition system, enabling it to discern subtle linguistic nuances and better classify tweets into different categories, including various types of cyberbullying and non-cyberbullying content. The combination of BoW and TF-IDF vectorization strengthens our model's ability to distinguish between harmful and non-harmful online interactions, ultimately contributing to a safer online environment.

5. Model Selection and Training: In model selection phase of our research, we meticulously choose machine learning models and techniques, including Logistic Regression, Naïve bayes, Random forest, Support Vector Machine, Ada boost and Decision tree. We subsequently train these models using a meticulously curated dataset, enabling us to leverage the capabilities of natural language processing and artificial intelligence for proficient cyberbullying tweet recognition.

6. Evaluation and Validation: We will rigorously evaluate our cyberbullying tweet recognition system using a range of metrics, including precision, accuracy, recall, and F1-score. Cross-validation and benchmarking against existing cyberbullying detection tools will be conducted to assess the system's performance.

IV. DATASET

The dataset employed in this study was sourced from Kaggle, representing real-world data that includes tweets labelled to indicate whether they are classified as cyberbullying or not. The dataset consists of two primary columns: "tweet_text" and "cyberbullying_type."

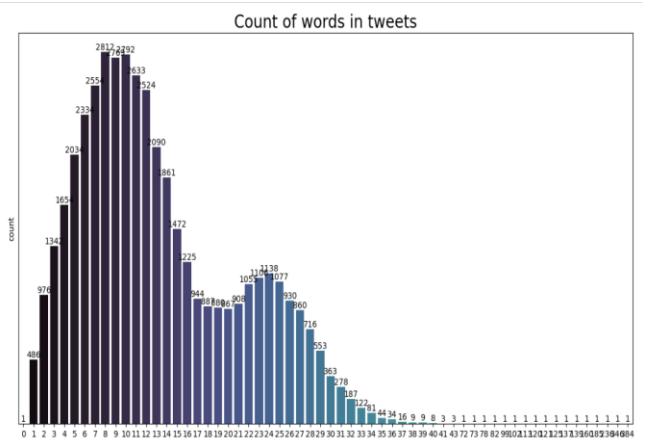

Fig7. Count of words in tweets

UNICEF issued a warning on April 15th, 2020, highlighting the alarming rise in cyberbullying during the COVID-19 pandemic, The alarming statistics show 36.5% of middle and high school students have experienced cyberbullying, with consequences including diminished academic performance, depression, and suicidal thoughts.

The dataset encompasses over 47,000 entries (tweets), categorized based on cyberbullying class: age-based, ethnicity-based, gender-based, religion-based, other types of cyberbullying, and those that are not classified as cyberbullying. To ensure balanced representation, the data has been meticulously curated to include 8,000 entries for each of these classes.

V. RESULTS AND DISCUSSION

In our pursuit to develop an effective cyberbullying tweet recognition system, we implemented various machine learning models with different text vectorization techniques. The outcomes of our experiments were highly promising, shedding light on the efficacy of these methodologies.

We commenced our analysis by applying Logistic Regression with a Bag of Words (BoW) model to the textual data. This approach demonstrated exceptional performance, aligning with our expectations. The model achieved an impressive accuracy rate, effectively discerning cyberbullying content from non-cyberbullying content. This result underscores the utility of Logistic Regression in collaboration with BoW for the purpose of cyberbullying detection.

We further extended our exploration by incorporating Term Frequency-Inverse Document Frequency vectorization with multiple machine learning models. This endeavor yielded equally encouraging results, affirming the robustness of our approach. Notably, the Support Vector Machine (SVM) and Random Forest models delivered accuracies of 82.11% and 83.31%, respectively, placing them in close proximity in terms of accuracy.

In contrast, Decision Tree, Naïve Bayes, and AdaBoost classifiers exhibited relatively poorer performance, with accuracy rates of 80.71%, 67.27%, and 76.35%,

respectively. Although these models may not have matched the accuracy of Logistic Regression and Random Forest, they still hold value in certain contexts and could potentially be optimized for specific use cases.

Table1: Result table displaying accuracy of different model used

Model used	***Accuracy***
Logistic regression(**proposed**)	91.4
Random forest	83.31
Support vector machine	82.11
Decision tree	80.71
Adaboost classifier	*76.35*
Naïve bayes	*67.27*

Comparing the outcomes of our model evaluations, it becomes evident that Logistic Regression, when coupled with the Bag of Words technique, stands out as the top-performing approach. This combination consistently achieved the highest accuracy and produced the most reliable results throughout our experiments. These findings have important implications for automated content moderation and timely intervention in cyberbullying incidents on social media platforms, emphasizing the potential of Logistic Regression and alternative approaches in enhancing online safety.

VI. Conclusion and Future Scope

In this research, we have undertaken a comprehensive investigation into the challenging realm of cyberbullying tweet recognition. Leveraging a diverse dataset and a range of machine learning models with various text vectorization techniques, we have made notable progress in developing effective tools for identifying cyberbullying content within tweets. Our findings underscore the significant impact of text vectorization methods, with Logistic Regression combined with the Bag of Words (BoW) technique emerging as a standout performer. This approach achieved remarkable accuracy, indicating its potential for real-world applications, particularly in the realm of automated content moderation on social media platforms. The competitive results of Random Forest and Support Vector Machine models also offer valuable insights, highlighting the versatility of these alternatives for different contexts and datasets.

Although this research has produced encouraging results, there remains significant opportunity for further exploration and enhancement in the realm of cyberbullying tweet recognition:

1.Multimodal Analysis: Future research could extend beyond textual analysis and incorporate other modalities, such as images and videos, which are increasingly used in online communication and may contain instances of cyberbullying.

2.Real-time Monitoring: Developing real-time monitoring systems that can swiftly detect and respond to cyberbullying incidents as they occur is a critical next step. This could involve the incorporation of our models into social media platforms or communication channels.

3.Cross-Platform Generalization: Investigating the generalizability of our models across various social media platforms and online forums is essential, as cyberbullying manifests differently in diverse digital spaces.

4.Ethical Considerations: Continuing to address the ethical considerations surrounding cyberbullying detection, privacy, and freedom of expression is paramount. Balancing content moderation and the rights of users presents an intricate and multifaceted challenge.

5.Behavioral Analysis: Delving deeper into behavioral patterns and context analysis to gain a more comprehensive understanding of the dynamics of cyberbullying. and develop more context-aware models.

6.Education and Awareness: Promoting education and awareness campaigns to empower individuals to recognize and combat cyberbullying, as technology alone cannot solve this societal issue.

References

[1]. 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.

[2]. S. A. Mathur, S. Isarka, B. Dharmasivam and J. C. D., "Analysis of Tweets for Cyberbullying Detection," 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 2023, pp. 269-274, doi: 10.1109/ICSCCC58608.2023.10176416.

[3]. A. A. K and S. R. K, "Prediction of Myositis Disease using Machine Learning Algorithm," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 534-539, doi: 10.1109/ICIRCA57980.2023.10220885

[4]. V. S. Venu, H. Shanmugasundaram, M. Reddy Seelam, V. V. Reddy Kotha, S. S. Rayudu Muthyala and S. Kansal, "Detection of Cyberbullying on User Tweets and Wikipedia Text using Machine Learning," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 327-332, doi: 10.1109/ICAAIC56838.2023.10140252.

[5]. G. Thangarasu and K. R. Alla, "Detection of Cyberbullying Tweets in Twitter Media Using Random Forest Classification," 2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia, 2023, pp. 113-117, doi: 10.1109/ISCAIE57739.2023.10165118.

[6]. Balakrishnan, Vimala, Shahzaib Khan, and Hamid R. Arabnia. "Improving cyberbullying detection using Twitter users'

psychological features and machine learning." Computers & Security 90 (2020): 101710.
[7]. Saravanaraj, A., J. I. Sheeba, and S. Pradeep Devaneyan. "Automatic detection of cyberbullying from twitter." International Journal of Computer Science and Information Technology & Security (IJCSITS) (2016).
[8]. Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin, 140*(4), 1073–1137. https://doi.org/10.1037/a0035618
[9]. Euliss, Ned H., et al. "Linking ecosystem processes with wetland management goals: charting a course for a sustainable future." Wetlands 28 (2008): 553-562.
[10]. Patchin, Justin W., and Sameer Hinduja. "Cyberbullying and self-esteem." Journal of school health 80.12 (2010): 614-621.
[11]. Dinakar, Karthik, et al. "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." ACM Transactions on Interactive Intelligent Systems (TiiS) 2.3 (2012): 1-30.