iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60284

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

An Enhanced System to Detect Cyberbullying and Automate Reporting on Twitter Using Text Based Pattern Recognition Technique

Dr. P. Sumathi[1], Dhakshinya Marudhavanan[2], M. Raghul[3], S. Rajarajan[4], S. Sivasaamy[5]
[1]*Head of the department,* [2, 3, 4, 5]*B.Tech student, Department of Artificial intelligence & Data science, SNS College of Engineering, Sathy road, Saravanampatti post, Coimbatore- 641035.*

Abstract: ***The increasing prevalence of cyberbullying on social media platforms necessitates effective detection and response mechanisms. This paper presents an enhanced system for detecting cyberbullying directed at politicians on Twitter and automating the reporting process. Utilizing advanced text-based pattern recognition techniques, the systePm identifies potentially harmful content and automatically reports it to a designated bot account for further action. We detail the system's architecture, the machine learning algorithms employed, and the performance of the system in terms of accuracy and speed. The proposed solution not only automates the detection and reporting processes but also contributes to safer online environments for politically active individuals.***

Keywords: ***Cyberbully, politician, twitter, report, multimonial naive bayes, API, text based.***

I. INTRODUCTION

In recent years, the pervasive influence of social media platforms like Twitter has become increasingly intertwined with political discourse, providing a platform for politicians to engage with constituents and share their perspectives on various issues. However, alongside the opportunities for communication and engagement, social media also presents significant challenges, particularly in the form of cyberbullying. Politicians, in particular, face a relentless barrage of online harassment, ranging from personal attacks to threats of violence, which can have profound consequences on their mental health, public perception, and even their ability to perform their duties effectively. The development of an enhanced report system to detect cyberbullying for politicians on Twitter using text-based pattern recognition techniques represents a critical step towards creating a safer and more inclusive online environment. By empowering politicians to report abusive content more effectively, facilitating proactive intervention by platform administrators, and upholding principles of privacy and security, this system aimsto mitigate the harmful effects of cyberbullying and safeguard the integrity of political discourse on social media platforms.

II. PROBLEM STATEMENT

The rise of social media has democratized communication, allowing individuals from diverse backgrounds to connect and express themselves on a global scale. Politicians, recognizing the potential of these platforms to reach and mobilize voters, have increasingly embraced social media as integral components of their political strategies. However, the open nature of social media also exposes politicians to a range of risks, chief among them being cyberbullying. Unlike traditional forms of harassment, cyberbullying transcends physical boundaries and operates in a virtual realm where anonymity and impunity reign supreme. As such, politicians are vulnerable to a relentless barrage of online abuse, ranging from personal attacks and character assassinations to threats of violence and defamation.

III. PROPOSED SYSTEM

Developing an algorithm capable of accurately identifying cyberbullying instances within tweets is paramount in combating the pervasive issue of online harassment. By leveraging advanced natural language processing (NLP) techniques, our algorithm meticulously analyzes tweet content to discern patterns associated with harmful behavior, encompassing a spectrum of cyberbullying manifestations, including direct threats, offensive language, and targeted attacks. The implementation of an automated reporting mechanism further fortifies our system's efficacy, facilitating the swift generation of detailed reports containing pertinent information such as the offending tweet, user details, and timestamp, thereby expediting intervention by relevant authorities or platform administrators.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

A. System requirements

Win 10 or 11 / MacOS sonomo or higher8 GB RAM
Enabled Path Environments for IDE
IDE-RequirementsPandas
Scikit-learn
TF-ID vectorizerTextblob Tweepy
Yaml
FrameworksRandom Flask Streamlit

B. Data-Training set

To develop a robust cyberbullying detection model tailored for politicians on Twitter, a meticulous approach to dataset curation, annotation, and model training is imperative. The foundation of this endeavor lies in the procurement of a comprehensive dataset encompassing a diverse array of tweets targeting politicians, spanning both abusive and non-abusive content to enable the development of discerning algorithms. The dataset assembly commences with the scrupulous collection of tweets via Twitter's public API, targeting accounts affiliated with politicians from varied political affiliations, jurisdictions, and ideological stances. This endeavor entails strategic querying utilizing keywords, hashtags, and user handles pertinent to politics and government, ensuring a comprehensive representation of political figures to encapsulate a spectrum of perspectives and experiences.

C. Algorithm

Multinomial Naive Bayes (MNB) serves as a cornerstone for text classification, playing a pivotal role in sifting through the vast volume of tweets on Twitter to distinguish between normal content and instances of cyberbullying. The system is meticulously designed to harness the capabilities of MNB within this framework, analyzing the textual content of tweets and classifying them into appropriate categories based on learned patterns. Before MNB assumes its pivotal role, the system embarks on a journey of data collection, amassing a diverse dataset of tweets that spans the spectrum from typical social media discourse to instances of cyberbullying, laying the foundation for training and testing the MNB classifier.

As shown in fig:1, the text preprocessing pipeline comes into play next, applying a suite of techniques to standardize the data, encompassing tokenization, lowercasing, removal of stop words, and potentially stemming or lemmatization, thereby ensuring consistency and enhancing the quality of the input data for subsequent classification. Armed with the preprocessed data, the system proceeds to the crucial phase of feature extraction, where textual data is transformed into a numerical format using methods such as the bag-of-words model or TF-IDF, enabling the MNB classifier to comprehend and analyze the data effectively.

Before MNB comes into play, the system collects a diverse dataset of tweets, encompassing both typical social media discourse and instances of cyberbullying. This dataset forms the basis for training and testing the MNB classifier. Text preprocessing techniques are then applied to standardize the data, including tokenization, lowercasing, removing stop words, and potentially stemming or lemmatization. These steps ensure consistency and enhance the quality of the input data for classification.

With the preprocessed data in hand, the next step involves feature extraction. Textual data is represented in a numerical format using methods such as the bag-of-words model or TF-IDF. These techniques convert text into numerical feature vectors, allowing the MNB classifier to understand and analyze the data. Despite its simplicity, MNB performs well in text classification tasks, particularly in scenarios where the data is sparse and high-dimensional, as is often the case with text data.

The heart of the system lies in training the MNB classifier. Multinomial Naive Bayes is a probabilistic classifier based on Bayes' theorem, with the "naive" assumption of feature independence given the class. Despite this simplification, MNB is effective in text classification tasks because it can handle large feature spaces efficiently and is well-suited for modeling word frequencies. During the training phase, the classifier learns from the labeled dataset, adjusting its parameters to maximize the likelihood of observing the training data given the class labels.

Despite its apparent simplicity, MNB demonstrates remarkable efficacy in text classification tasks, particularly in scenarios characterized by sparse and high-dimensional data, a frequent occurrence with text data. At the heart of the system lies the training of the MNB classifier, a process wherein Multinomial Naive Bayes, a probabilistic classifier grounded in Bayes' theorem with the "naive" assumption of feature independence given the class, emerges as a formidable tool. Despite the inherentsimplification, MNB excels in text classification tasks owing to its adeptness at efficiently handling large feature spaces and its aptitude for modeling word frequencies.

©IJRASET: All Rights are Reserved

During the training phase, the classifier imbibes insights from the labeled dataset, adeptly adjusting its parameters to maximize the likelihood of observing the training data given the class labels, thereby honing its discriminatory prowess. Once the MNB classifier completes its training regimen, it undergoes a rigorous evaluation process aimed at gauging its performance across various metrics such as accuracy, precision, recall, and F1-score, to ascertain its efficacy in discerning between normal tweets and instances of cyberbullying. This evaluative phase serves as a crucible for fine-tuning the classifier, ensuring its adeptness in reliably identifying cyberbullying content while minimizing the occurrences of false positives and false negatives.

In the context of our project, the trained MNB classifier seamlessly integrates into the larger system for cyberbullying detection and automated reporting on Twitter. In real-time, incoming tweets undergo scrutiny by the classifier, which adeptly identifies potential instances of cyberbullying based on learned patterns and linguistic cues. Upon detecting cyberbullying content, the system springs into action, automatically generating reports and flagging the offending tweets for further review by moderators or authorities, thereby facilitating swift intervention to safeguard users from online harassment and fostering a safer and more inclusive online environment. Multinomial Naive Bayes stands as a linchpin in our project, furnishing an efficient and effective means of text classification for cyberbullying detection on Twitter. Leveraging the potent amalgam of machine learning and natural language processing techniques, our endeavor aims to engender a robust system that not only identifies cyberbullying content with precision but also expedites timely intervention and support for affected users, thereby ushering in a safer and more harmonious online ecosystem.

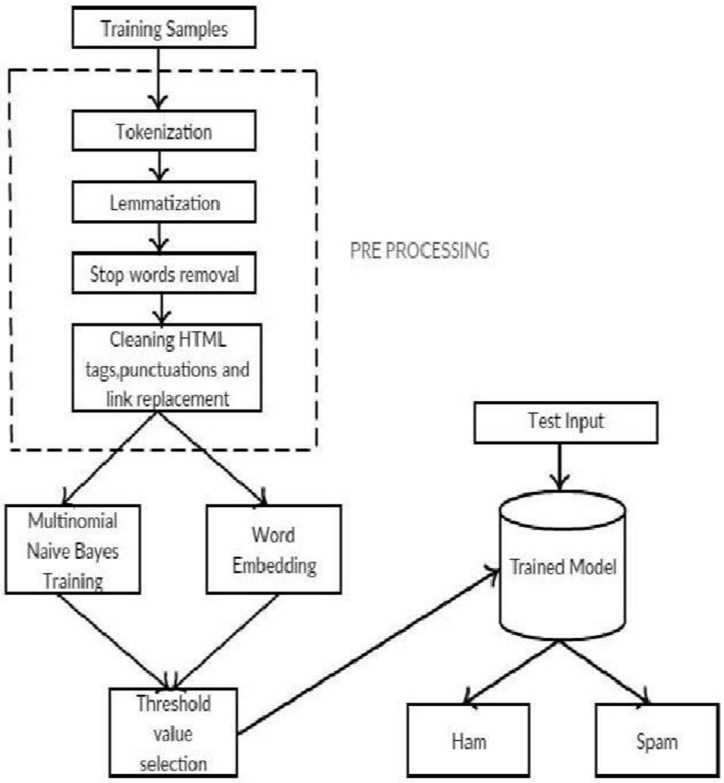


Fig. 1 THE WORKING OF THE ALGORITHM IN A SEQUENCE

D. Working

1) Classification of dataset – Using MNB algorithm Randomly selected tweets(5) and classified as Bully and Non Bully

2) Trained using ML algorithm(MNB) for text classification and NLP algorithm (Textblob) for sentiment analysis topredict labels for new tweets

3) X API CHECK v2, POST tweet

4) Twitter API authentication and posting messages using Tweepy, this phase is specifically for establishing a connectionto the Twitter API and posting a message on Twitter

5) Integrating Classification with X API – Shows classification accuracy, Prediction accuracy, Filters only Bully toreport, Input selection, Automates Report by posting tweet

Fig. 2 EXAMPLE OF POSTING MESSAGE USING TWEEPY

IV. CONCLUSIONS

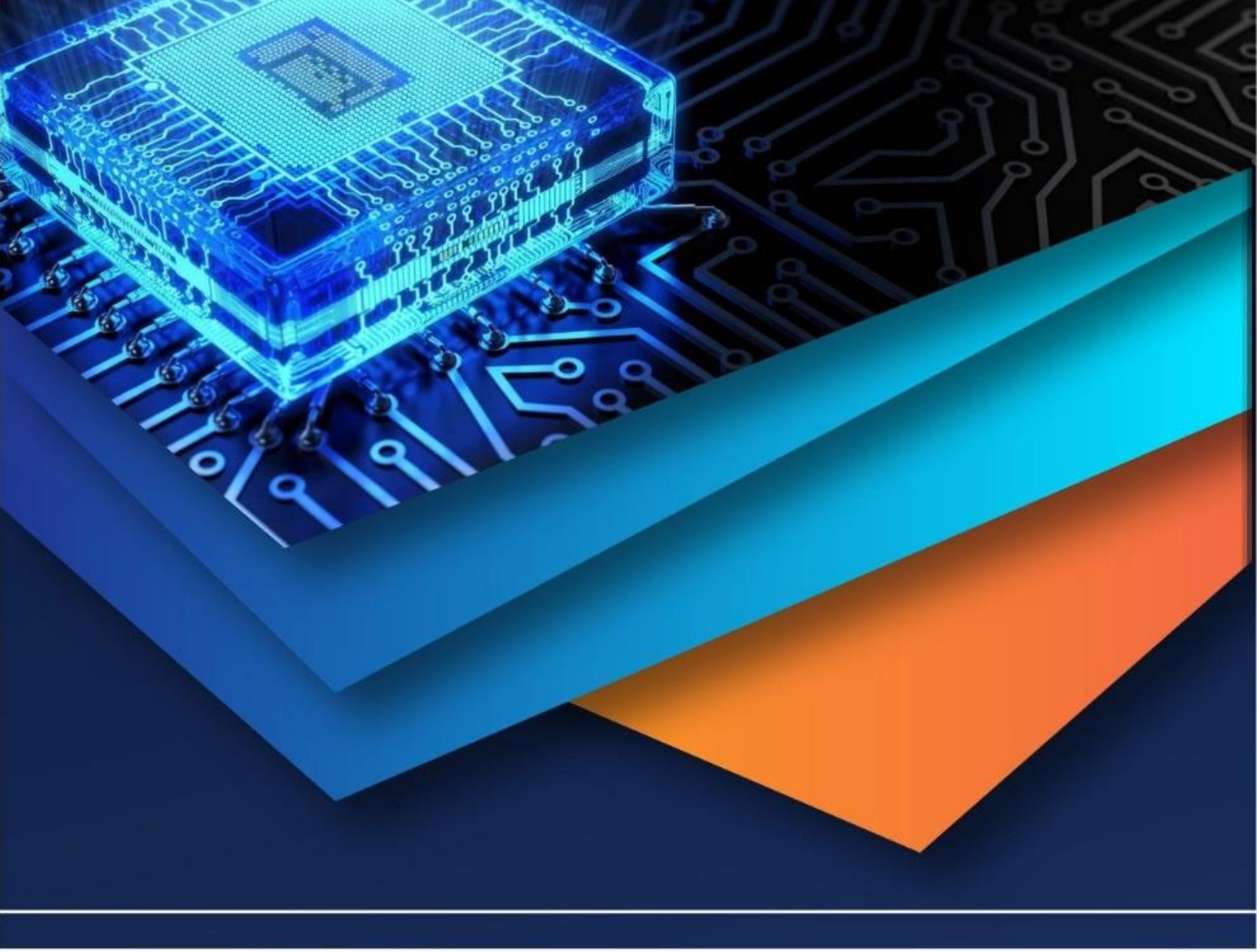
In conclusion, the development of a cyberbullying detection system targeting politicians on Twitter represents a significant step towards creating a safer and more respectful online environment. Throughout this project, we have outlined the objectives, scope, and software requirements necessary for the successful implementation of such a system. By focusing on accurately reporting the bully, providing the bully's ID, and highlighting the specific bullied word or comment, our proposed work aims to address the pervasive issue of cyberbullying directed at politicians.

Furthermore, our proposed work prioritizes privacy preservation and ethical considerations to ensure the protection of individuals' identities and personal information involved in cyberbullying incidents. By adhering to ethical guidelines and data protection regulations, we aim to mitigate the risk of unintended consequences and harm while maintaining transparency and accountability in our cyberbullying detection efforts.

REFERENCES

[1] P. Burnap and M. L. Williams, "Cyber hate speech on Twitter: An application of machine learning and statistical techniques," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 178, no. 2, pp. 435-465, 2015.
[2] S. Kumar, D. Barbosa, and J. Benevenuto, "Detecting harassment on social media," in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, 2014, pp. 819-828.
[3] R. R. McCrae, J. Allik, and J. Lönnqvist, "The need for cross-cultural research on the mechanisms of cyberbullying," in Cyberbullying: From theory to intervention, Springer, Cham, 2019, pp. 207-220.
[4] M. R. Smith, K. D. J. Ruddick, and S. H. Jun, "An evaluation of machine learning models for detecting cyberbullying on Twitter," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, no. 1, pp. 618-629, 2020.
[5] D. West, T. King, and J. M. Watson, "Exploring cyberbullying detection on social media using deep learning neural networks," Information Sciences, vol.559, pp. 21-38, 2021.
[6] S. S. Kang and H. H. Hwang, "A review of cyberbullying detection using machine learning techniques," International Journal of Information Management, vol. 55, pp. 102154, 2020.
[7] C. Silva, T. Ribeiro, and P. Benevenuto, "A comprehensive survey on content-based cyberbullying detection," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-38, 2021.
[8] S. Huang, W. Wang, and K. Li, "Detecting cyberbullying on social media with deep learning and data augmentation," Information Processing & Management, vol. 58, no. 6, pp. 102687, 2021.

[9] D. Chatzakou et al., "Mean birds: Detecting aggression and bullying on Twitter," in Proceedings of the 2017 ACM on Web Science Conference, 2017, pp.13-22.
[10] F. T. O'Nell et al., "A survey of cyberbullying detection techniques," Computers in Human Behavior, vol. 88, pp. 337-345, 2018.
[11] E. Grinberg et al., "Extracting sociodemographic and socioeconomic patterns from online social networks and big data sources," Social Science Computer Review, vol. 33, no. 5, pp. 617-637, 2015.
[12] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.
[13] G. Brantner et al., "Finding meaningful communities in link collections," in Proceedings of the 15th International Conference on World Wide Web, 2006,pp. 431-440.
[14] C. Shah and H. Panchal, "A comprehensive review of cyberbullying detection techniques using machine learning algorithms," in Proceedings of the 20182nd International Conference on Inventive Systems and Control, 2018, pp. 1323-1328.
[15] A. K. Samara and A. I. Al-Khasawneh, "A review of cyberbullying detection and mitigation methods," Journal of Network and Computer Applications, vol. 174, pp. 102919, 2021.
[16] L. E. Almeida et al., "Cyberbullying detection on social media: A systematic review," Expert Systems with Applications, vol. 165, pp. 113824, 2021.
[17] B. M. Loria and D. W. McDonald, "Who do you bully and why?: Contextualizing cyberbullying perpetration in interpersonal relationships," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1 -14.
[18] A. C. L. Cunha, J. M. N. Silva, and G. F. V. Medeiros, "A survey on machine learning techniques applied to the cyberbullying detection problem," Journal of King Saud University- Computer and Information Sciences, 2020.
[19] K. O. Şahin and H. A. N. Yücebaş, "Sentiment analysis of cyberbullying on social media," in Proceedings of the International Conference on Soft Computing Models in Industrial and Environmental Applications, 2017, pp. 560-571.
[20] L. A. Moore, C. A. Nocera, and A. L. Zeichick, "Understanding bystander behaviors to cyberbullying: A meta-analytic review," Social Media + Society, vol. 6, no. 2, pp. 1-11,2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)