IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11443

Social Media-Based Hate Speech And Stress Identification Through Machine Learning And Natural Language Processing (NLP)

Mrs. Sharon D'Souza[1], Ashwin Shetty[2], Jeevan M[3], Nishal SP Karkera[4], Rahul D Shetty[5]

Assistant Professor, Computer Science and Engineering, AJIET, Mangalore, India[1]

Student, Computer Science and Engineering, AJIET, Mangalore, India[2-5]

Abstract: The proliferation of hate speech on social media has become a pressing societal concern, prompting the need for effective identification and mitigation strategies. This abstract outlines a novel approach utilizing machine learning (ML) and natural language processing (NLP) techniques to detect hate speech and assess its impact on inducing stress among users. The study focuses on the development of an ML-based model trained on a diverse dataset of social media content to accurately identify hate speech. Leveraging NLP, the model aims to comprehend linguistic nuances, context, and sentiment within textual data, enabling it to distinguish between normal discourse and potentially harmful language. Furthermore, the research extends beyond mere identification, aiming to gauge the psychological impact of hate speech by analyzing its correlation with stress levels among social media users. By employing sentiment analysis and stress identification algorithms, the study aims to quantify the emotional toll experienced by individuals exposed to such content. The abstract emphasizes the interdisciplinary nature of the research, bridging the gap between computer science, linguistics, and psychology. The proposed methodology holds promise in aiding social media platforms, policymakers, and mental health professionals in devising targeted interventions to combat hate speech and mitigate its adverse effects on users' well being. Through this holistic approach, this study endeavors to contribute to the development of proactive strategies for early detection, intervention, and support mechanisms, fostering a safer and healthier online environment for all users.

Keywords: Stressfull comments, hate speech, personal assaults, healthier online environment.

I. INTRODUCTION

Over the last decade social media has acquired a lot of attraction both positively and negatively way with the fast growth of social networking. People can communicate with one other via many social media platforms. In these the digital age, social media platforms have become vibrant hubs for communication, fostering connectivity and information sharing. However, alongside these benefits, they have also become breeding grounds for hate speech and stress-inducing content. The pervasive nature of online interactions has highlighted the urgent need to develop robust mechanisms for identifying and mitigating harmful discourse.

Addressing this challenge requires innovative solutions that leverage machine learning to detect hate speech and stress indicators embedded within the vast expanse of social media content. This research endeavors to create a sophisticated machine learning framework capable of discerning hate speech and stress markers within the complex fabric of online conversations. By employing cutting-edge algorithms in Machine learning, natural language processing (NLP) and sentiment analysis, this system aims to detect subtle linguistic cues, contextual nuances, and patterns indicative of hate speech or emotional distress.

Through the analysis of user-generated content across various social media platforms, this technology seeks to categorize and flag harmful discourse, enabling swift interventions to maintain a healthier online environment. This project addresses the pervasive challenges of hate speech and stress on social media platforms through an innovative machine learning solution.

Leveraging a multi-modal approach that analyzes text, and images, our system dynamically adapts to changing sentiment landscapes. It goes beyond traditional sentiment analysis by prioritizing contextual awareness, considering broader conversations, user interactions, and historical context. Transparency is emphasized through explain ability in model predictions, fostering user trust.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11443

II. PROBLEM STATEMENT

Identifying hate speech and stress on social media through machine learning (ML) and natural language processing (NLP) involves tackling a multifaceted challenge that intersects technological, ethical, and social considerations. Ethical considerations play a pivotal role in this domain.

Handling sensitive content like hate speech or distressing language demands a delicate balance between ensuring user safety and privacy while also mitigating potential biases within the algorithms. Ensuring fairness, transparency, and accountability in the models' decision-making processes is essential. Developing robust ML models involves creating algorithms that can learn patterns and features indicative of hate speech or stress.

This requires an extensive, diverse, and carefully annotated dataset for training. These datasets should encompass a wide range of demographics, languages, and social contexts to ensure the models are sensitive to various cultural nuances and aren't biased towards specific groups. One significant challenge is adapting to the constantly evolving nature of language and online behavior.

New words, phrases, or expressions frequently emerge, necessitating continuous model updates and retraining to maintain accuracy and relevance. In conclusion, the identification of hate speech and stress on social media via ML and NLP is a complex and interdisciplinary challenge. Developing accurate, fair, and culturally sensitive models that can navigate the subtleties of language and user intent is pivotal to fostering a safer and more supportive online environment

III. OBJECTIVE

1.The objective is to develop a robust machine learning system that effectively discerns hate speech and stress markers in social media content.

2. By leveraging algorithms capable of analyzing linguistic patterns and contextual cues, this system aims to accurately detect and categorize harmful or distress-inducing language.

3. Finding the stressful posts on social media platform which may causes psychological instability of people’s.

4. Data-Driven Decision Making Utilize AI/ML algorithms to analyze feedback data, extract meaningful patterns, and provide actionable insights. This objective aims to empower decision-makers with data-driven information for strategic human resource planning.

IV. EXPECTED OUTCOMES

Hate Speech Detection Outcomes

1. Increased Accuracy in Hate Speech Detection:Detail the expected improvement in model accuracy based on selected features and algorithms. Discuss the significance of achieving higher precision and recall rates in hate speech identification.

2. Enhanced Contextual Understanding in Hate Speech Detection: Anticipate improvements in the model's ability to comprehend nuanced language and context.Discuss the potential reduction in false positives and negatives through enhanced contextual analysis.

3. Cross-Cultural Adaptability in Hate Speech Detection:Outline the expected success in adapting the model to diverse cultural contexts. Discuss the impact on mitigating biases and improving the inclusivity of hate speech detection.

Stress Detection Outcomes

1. Effective Identification of Stress Indicators: Anticipate the model's capability to identify linguistic patterns indicative of stress. Discuss the potential applications for user well-being and mental health awareness.

2. Correlation Analysis between Hate Speech and Stress: Expect insights into how instances of hate speech may contribute to increased stress levels.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11443

V. LITERATURE SURVEY

[1] **Filip Klubicka, Raquel Fernandez "Examining a hate speech corpus for hate speech detection and popularity prediction"** Paper discussion environment online can become abusive, hateful, and toxic, especially when user anonymity is added. In order to identify, study, and ultimately contain this problem, such negative environments and the language used within them are studied under the name of hate speech. Consequently, and especially when user anonymity is added to the mix, online discussion environments can become abusive, hateful and toxic. To help identify, study, and ultimately curb this problem, such negative environments and the language used within are being studied under the name hate speech.

[2] **Raquel Fernandez et al. Published "Hate Speech Corpus Research for Detecting Hate Speech and Predicting Popularity".** Their study found that Twitter has a sizeable Black following, but anti-Black people We've found that racist tweets are particularly damaging to the Twitter community. In doing so, they can provide data on the sources of hate speech against black people. This study aims to address the quality of datasets, which is a major concern raised by many of the problems that have been brought to light. This paper also addresses the second issue, which is that the best characteristics for hate speech identification must be investigated and determined before developing a suitable classifier. For this reason, datasets tend to fall into one of these categories.

[3] **Pete Burnap and Matthew L. Williams "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modelling for Policy and Decision Making".** Through the analysis of user-generated content across various social media platforms, this technology seeks to categorize and flag harmful discourse, enabling swift interventions to maintain a healthier online environment. This study aims to address the quality of datasets, which is a major concern raised by many of the problems that have been brought to light. This paper also addresses the second issue, which is that the best characteristics for hate speech identification must be investigated

[4] **E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, "Statistical approaches to concept-level sentiment analysis".** The paper contributes to natural language processing and conversational AI by utilizing AIML. In this work, we propose a novel brain-inspired sentiment analysis framework to help machines emulate human inference of sentiment from natural language. By merging, for the first time, CI, linguistics, and common-sense computing, the proposed paradigm exploits the relations between concepts and linguistic patterns in text to reveal the flow of sentiment from concept to concept, and improve our understanding of how sentiment is conveyed in a sentence.

[5] **Florio, Komal, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. "Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media".** The paper contributes to the field by leveraging Convolutional Neural Networks and employing Recurrent Neural Networks (RNNs) to advance the state-of-the-art in text emotion recognition. The increasing availability of textual data from social media platforms is essential to the development of training datasets for Natural Language Processing (NLP) prediction tasks. In particular, hate speech detection is the NLP task that aims at classifying segments of text based on their hateful content. The abundance of data allows the research community to tackle more in-depth long-standing questions such as understanding, measuring, and monitoring users' sentiment towards specific topics or events.

[6] **Poletto, F., Basile, V., Sanguinetti, M. et al. Resources and benchmark corpora for hate speech detection: a systematic review. "With the announcement of Hate Speech Detection in Conventional Languages on Social media Using Machine Learning."** The need to automate the process of classifying hate speech data arises, the system detects hate speech based on the dataset applied to the English system. Hate Speech (HS), lying at the intersection of multiple tensions as expression of conflicts between different groups within and across societies, is a phenomenon that can easily proliferate on social media. It is a vivid example of how technologies with a transformative potential are loaded with both opportunities and challenges. Implying a complex balance between freedom of expression and defense of human dignity, HS is hotly debated and has recently gained traction in the AI community, that can play a leading role in developing tools to confront pervasive dangerous trends such as the escalation of violence and hatred in online communication, or the spread of fake news.

[7] **Damian. Detecting Emotions in Text. "Detected stress from multimodal signals using a machine learning framework, which includes feature identification, outlier detection, imputation, and classification".** The authors effectively addressed missing data and outliers. Within the field of AI, and Natural Language Processing (NLP) in particular, techniques for tasks related to Sentiment Analysis and Opinion Mining (SA&OM) grew in relevance over the past decades. Such techniques are typically motivated by purposes such as extracting users' opinion on a given product or polling political stance. Robust and effective approaches are made possible by the rapid progress in supervised learning technologies and by the huge amount of user-generated contents available online, especially on social media.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11443

[8] **Boiy E, Moens MF. "A machine learning approach to sentiment analysis in multilingual web texts."** The idea for exploring emotion in social media came from an interest in wanting to understand people, given the once in a lifetime opportunity to explore and calibrate our data to a significant global event. Sentiment analysis, also called opinion mining, is a form of information extraction from text of growing research and commercial interest. In this paper we present our machine learning experiments with regard to sentiment analysis in blog, review and forum texts found on the World Wide Web and written in English, Dutch and French. We train from a set of example sentences or statements that are manually annotated as positive, negative or neutral with regard to a certain entity.

[9**] Dmitry D, Oren T, Ari R. Enhanced sentiment learning using twitter hashtags and smileys**. Sentiment analysis, also called opinion mining, is a form of information extraction from text of growing research and commercial interest. In this paper we present our machine learning experiments with regard to sentiment analysis in blog, review and forum texts found on the World Wide Web and written in English, Dutch and French. Sentiment extraction systems usually require an extensive set of manually supplied sentiment words or a handcrafted sentiment-specific dataset. With the recent popularity of article tagging, some social media types like blogs allow users to add sentiment tags to articles. This allows to use blogs as a large user-labeled dataset for sentiment learning and identification. However, the set of sentiment tags in most blog platforms is somewhat restricted.

[10**] Krestel R, Fankhauser P. Personalized topic-based tag recommendation**. This paper focuses on the key subtask in sentiment analysis: aspect-based sentiment analysis. Unlike feature-based traditional approaches and long short-term memory network based models, our work combines the strengths of neural network based model for aspect-based sentiment analysis. Social Media Analytics (SMA) is the process of collecting information on various social media platforms, websites and blogs and evaluating that, to successful business decisions. The use of social media has become quite commonplace in today's world. SMA is not only a collection of likes and comments shared by people but also a platform for many advertising brands.

VI. REQUIREMENT SPECIFICATION

Hardware requirements

This application is designed to run on the minimum possible configuration of hardware.

- RAM: 8GB
- Processor: AMD Ryzen 5 10th generation
- Hard disk: compatable

Software requirements

- Tool: PyCharm
- Language: Python 3.6
- Tensorflow
- Keras
- Pandas
- Flask
- Scikit-Learn, Matplotlib and Librosa

VII. SYSTEM DESIGN

A. A Flow Diagram

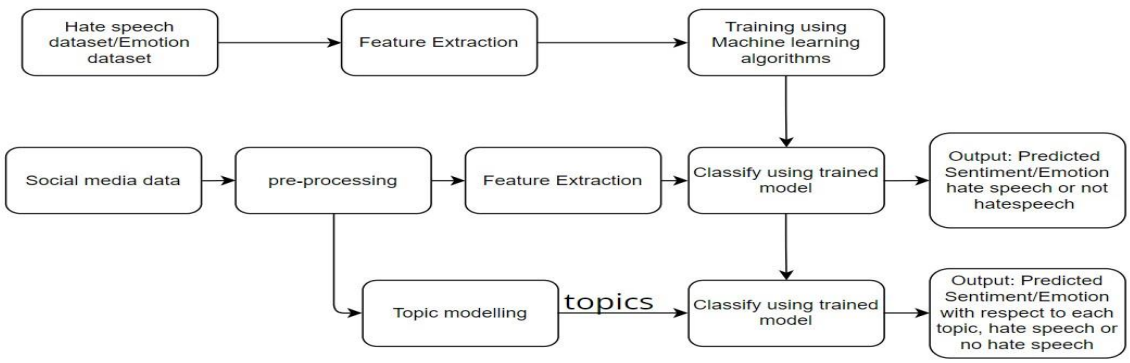

Fig:Flow diagram

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11443

The flowchart in Figure above outlines the process of toxic comments detection and classification

Hate Speech Detection Module:

• Algorithm Selection: Discuss the choice of machine learning algorithms for hate speech detection. Explain how the selected algorithms contribute to accurate detection.

• Feature Engineering: Outline the features used for hate speech detection. - Discuss how linguistic and contextual features are extracted and processed.

• Model Training and Validation:

Detail the process of training hate speech detection models. - Discuss validation techniques to ensure robust performance.

• Cross-Cultural Adaptability: Explore strategies for adapting the hate speech detection module to diverse cultural contexts. Discuss any challenges and solutions related to cultural nuances in language

Stress Detection Module:

• Feature Selection: Discuss the features used to identify stress indicators in textual content. - Explore linguistic patterns and psychological cues for effective stress detection.

• Correlation Analysis: Explain the methods employed to conduct correlation analysis between hate speech and stress. Discuss how insights from this analysis inform stress detection.

• User Well-being Metrics: Define metrics for assessing user well-being based on stress detection - Discuss the implications for mental health awareness and community support

VIII. RESEARCH METHODOLOGY

The research methodology for the development of a Social Media-Based Hate Speech and Stress Identification System using Machine Learning and Natural Language Processing entails several key steps. Initially, data collection is conducted from various social media platforms, encompassing text data containing instances of hate speech and indicators of stress. Following data collection, preprocessing techniques such as text normalization, tokenization, and noise removal are applied to clean the text data

Next, feature engineering is performed to extract relevant features from the text data, including linguistic features, sentiment analysis, and contextual information. These features are then utilized to train machine learning models, such as classification algorithms, to distinguish between hate speech and non-hate speech content, as well as to identify indicators of stress.The trained models are evaluated using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score, through techniques like cross-validation to ensure robustness and generalizability. Furthermore, techniques for handling class imbalances and bias mitigation are employed to enhance the model's performance and fairness.User feedback and validation are crucial aspects of the methodology, involving experts in linguistics, psychology, and social sciences to assess the system's effectiveness and interpretability. Additionally, ethical considerations regarding privacy, data protection, and potential biases are carefully addressed throughout the research process.

Finally, the developed system undergoes rigorous testing in real-world social media environments to validate its efficacy in identifying hate speech and stress-inducing content, thereby contributing to the mitigation of online toxicity and promoting mental well-being in digital communities

IX. CONCLUSION

The Expected Outcomes Specification is a pivotal aspect of the project, outlining the anticipated results and impacts across multiple facets. In the domain of hate speech detection, the project strives for heightened accuracy by leveraging advanced algorithms and features.

The objective is to surpass conventional benchmarks, enhancing the system's ability to discern and categorize hate speech with precision and recall rates that reflect a substantial improvement. Contextual understanding within hate speech detection is a critical focus, with the anticipation of reducing false positives and negatives through nuanced language comprehension.

The project's crosscultural adaptability outcome aims to navigate diverse cultural contexts effectively, mitigating biases and fostering inclusivity in hate speech detection. Turning to stress detection, the project envisions the effective identification of stress indicators, utilizing linguistic patterns to contribute to user well-being and mental health awareness. The machine learning integration goals center around achieving robust and adaptable models, demonstrating resilience to the dynamic nature of online communication.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 11, Issue 4, April 2024

DOI: 10.17148/IARJSET.2024.11443

REFERENCES

[1] Filip Klubicka, Raquel Fernandez "Examining a hate speech corpus for hate speech detection and popularity prediction" arXiv:1805.04661v1 [cs.CL] 12 May 2018.

[2] Raquel Fernandez et al. Published "Hate Speech Corpus Research for Detecting Hate Speech and Predicting Popularity". the Twitter corpus collected by Waseem and Hovy (2016).

[3] Pete Burnap and Matthew L. Williams "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modelling for Policy and Decision Making ." 1944-2866 # 2015 The Authors. Policy & Internet published by Wiley Periodicals, Inc. on behalf of Policy Studies Organization.

[4] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, "Statistical approaches to concept-level sentiment analysis," IEEE Intelligent Systems, vol. 28, no. 3, pp. 6–9, 2013 .

[5] Florio, Komal, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. "Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media" .Applied Sciences 10, no. 12: 4180. https://doi.org/10.3390/app10124180.Gaydhani, Aditya, et al. \"Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach.\" arXiv preprint arXiv: 1809.08651 (2018).

[6] Poletto, F., Basile, V., Sanguinetti, M. et al. "Resources and benchmark corpora for hate speech detection: a systematic review". Lang Resources & Evaluation 55, 477– 523 (2021). https://doi.org/10.1007/s10579-020- 09502-8.

[7] D. G. Haryadi, J. A. Orr, K. Kuck, S. McJames, and D. R. Westen-skow, "Partial co2 rebreathing indirect fick technique for non-invasive measurement of cardiac output." Journal of clinical monitoring and computing, vol. 16, no. 5-6, pp. 361–74, 2000.

[8] Dmitry D, Oren T, Ari R. "Enhanced sentiment learning using twitter hashtags and smileys". Coling 2010—23rd International Conference on Computational Linguistics, Proceedings of the Conference. 2. 2010; 241–249.

[9] Krestel R, Fankhauser P. Personalized topic-based tag recommendation. Neurocomputing. 2012;76:61–70. https://doi.org/10.1016/j.neucom.2011.04.034.

[10] Boiy E, Moens MF. A machine learning approach to sentiment analysis in multilingual web texts. Inf Retrieval. 2009;12:526–58. https://doi.org/10.1007/s10791-008-9070-z[N+1].