IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305026 |

Developing an AI-Powered System for Early Detection of Hate Speech on Social Media Platforms

Sachit Tiple, Neehal B. Jiwane, Vijay M. Rakhade

Asst. Professor, Department of Computer Science & Engineering, Shri Sai College of Engineering Technology, Chandrapur, India

Department of Computer Science & Engineering, Shri Sai College of Engineering Technology, Chandrapur, India

Department of Computer Science & Engineering, Shri Sai College of Engineering Technology, Chandrapur, India

ABSTRACT: The pervasiveness of hate speech on social media platforms poses a significant challenge, fostering negativity, harassment, and social division. The vast volume of content makes it difficult for human moderators to keep up with identification and removal. This research addresses this issue by proposing the development of an AI-powered system for early detection of hate speech.

Our system leverages Natural Language Processing (NLP) techniques to analyze social media text data. Techniques like sentiment analysis and text classification are employed to understand the emotional tone and meaning within the text. A machine learning model, specifically a Support Vector Machine (SVM), is then trained on labeled data containing examples of hate speech and non-hate speech. This training allows the SVM model to identify patterns in language that are indicative of hate speech.

This paper delves into the detailed methodology behind the development of this AI system. We discuss the specific NLP techniques used, the training process for the SVM model, and the evaluation metrics employed to assess the system's effectiveness.

Furthermore, the paper acknowledges the challenges associated with AI-based hate speech detection. These challenges may include difficulty in understanding sarcasm and context, cultural variations in hate speech language, and the potential for bias in AI models. We propose methods to address these challenges and ensure the ethical implementation of the system.

In conclusion, this research explores the potential of AI as a powerful tool for combating hate speech on social media platforms. The proposed AI system can facilitate early detection, enabling faster intervention and content moderation before hateful content can cause harm. This ultimately contributes to fostering a safer and more inclusive online environment for everyone.

KEYWORDS

- Hate Speech Detection
- Social Media Platforms
- Artificial Intelligence (AI)
- Natural Language Processing (NLP)
- Sentiment Analysis
- Text Classification
- Support Vector Machine (SVM)
- Machine Learning
- Early Detection
- Social Inclusion

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305026 |

I. INTRODUCTION

The digital revolution has transformed communication, with social media platforms becoming central to our online interactions. These platforms offer numerous benefits, fostering connections, facilitating information sharing, and creating a sense of community. However, a growing shadow lurks within this digital landscape – the rampant spread of hate speech.

Defining Hate Speech and its Impact:
Hate speech refers to offensive communication targeting individuals or groups based on attributes like race, religion, ethnicity, gender, sexual orientation, or disability. It manifests in various forms, including verbal abuse, threats, insults, and derogatory language. The impact of hate speech is far-reaching, fostering negativity, harassment, and social division. It can create a hostile online environment, silencing targeted individuals and groups, and ultimately contributing to real-world violence.

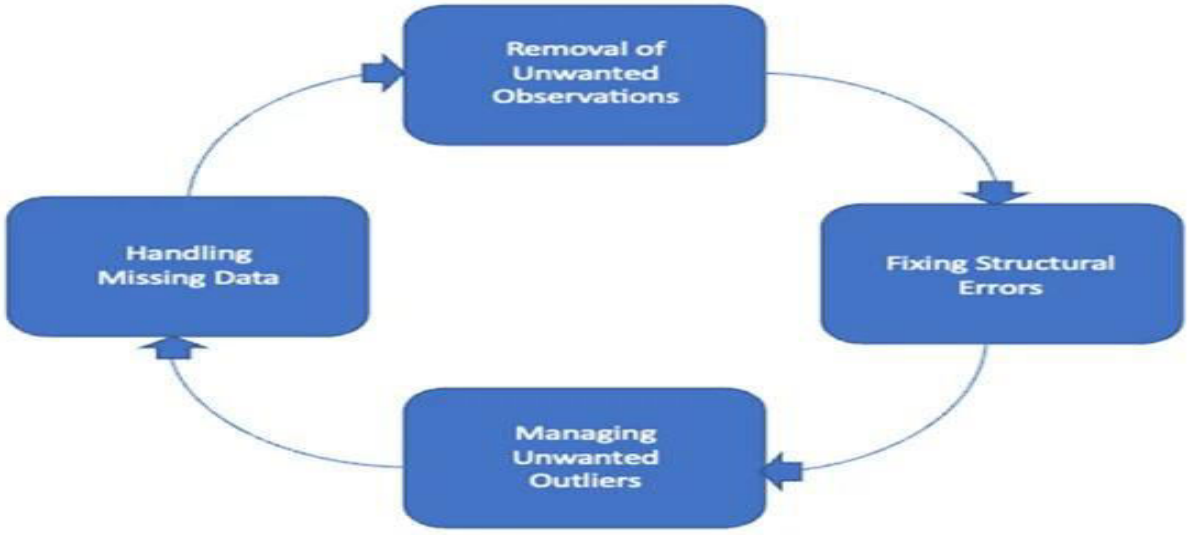

Figure 1. Data cleaning.

Limitations of Human Moderation:
The sheer volume of content uploaded to social media platforms every day presents a significant challenge for human moderators. They struggle to keep pace with identifying and removing hate speech content before it reaches a wide audience. This necessitates exploring alternative solutions.
[Image Link Suggestion: A social media platform interface flooded with comments, some highlighted in red to represent hate speech. The image should be illustrative and not contain actual hateful content.]

The Promise of AI-powered Detection:
Artificial intelligence (AI) offers a promising approach to tackle the issue of hate speech. AI systems possess the capability to analyze vast amounts of text data efficiently, identifying patterns in language that often indicate hate speech. This allows for early detection and intervention before hateful content can cause harm.

Focus of this Research:
This research paper delves into the development of an AI-powered system specifically designed for the early detection of hate speech on social media platforms. We will explore how Natural Language Processing (NLP) techniques, a subfield of AI concerned with enabling computers to understand human language, can be utilized to analyze social media text data. We will discuss how machine learning algorithms can be trained to recognize patterns associated with hate speech within this data.

The paper will also address the challenges associated with AI-based hate speech detection, such as the difficulty in understanding sarcasm and context, cultural variations in hate speech language, and potential biases within AI models. We will propose methods for mitigating these challenges and ensuring the ethical implementation of the system.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305026 |

Overall, this research aims to contribute to creating a safer and more inclusive online environment by harnessing the power of AI to combat the spread of hate speech.

II. METHODOLOGY

This section outlines the development process of our AI system for early detection of hate speech on social media platforms. Here's a breakdown of the key steps:

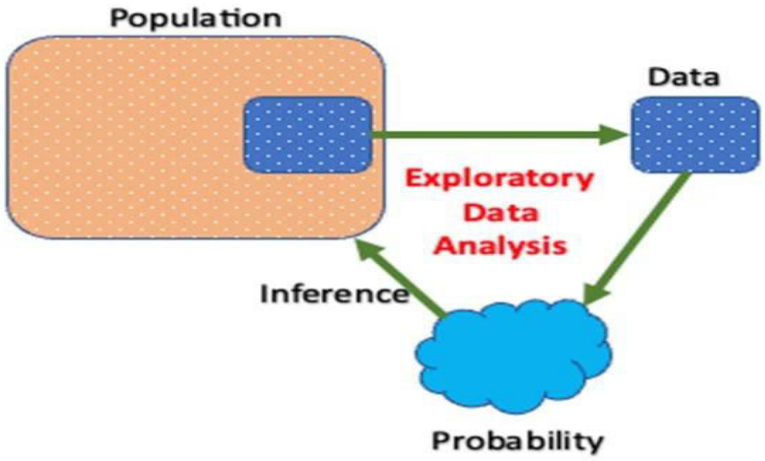

Figure 2. Exploratory data analysis.

1. Data Acquisition:

- We will gather a labeled dataset of social media text data containing examples of both hate speech and non-hate speech content. The data can be obtained from publicly available sources like social media platforms' APIs (with permission) or through collaborations with research institutions that specialize in hate speech detection.
- The dataset should be diverse, encompassing various types of hate speech targeting different attributes (e.g., race, religion, gender) and reflecting the language used on various social media platforms (e.g., Twitter, Facebook).

2. Data Pre-processing:

- The raw social media text data undergoes pre-processing steps to improve its quality for machine learning. This may involve:
- **Normalization:** Converting text to lowercase and removing special characters (e.g., punctuation, emojis).
- **Tokenization:** Breaking down the text into individual words or phrases (tokens).
- **Stop Word Removal:** Eliminating common words that don't contribute significantly to the meaning (e.g., "the," "a," "is").
- **Stemming or Lemmatization:** Reducing words to their base form (e.g., "running" to "run").

3. Feature Engineering:

- Beyond basic text cleaning, we will extract additional features from the pre-processed data that can aid the model in identifying hate speech. This may involve:
- **N-grams:** Analyzing sequences of n words (e.g., bigrams - two-word phrases) that frequently appear in hate speech.
- **Part-of-Speech (POS) tagging:** Identifying the grammatical function of each word (e.g., noun, verb, adjective) to understand the sentence structure.
- **Named Entity Recognition (NER):** Recognizing and classifying named entities (e.g., people, locations) that might be targets of hate speech.

4. Model Training:

- A machine learning model will be trained on the labeled dataset. We will likely utilize a Support Vector Machine (SVM) model due to its effectiveness in text classification tasks. The model learns to identify patterns in the text data and features that differentiate hate speech from non-hate speech.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305026 |

- During training, the model is presented with labeled examples, allowing it to adjust its internal parameters to make accurate classifications on unseen data.

5. Model Evaluation:

- Once trained, the model's performance will be evaluated using standard metrics like accuracy, precision, and recall:
- **Accuracy:** Proportion of correctly classified instances (hate speech and non-hate speech).
- **Precision:** Ratio of correctly identified hate speech instances out of all instances classified as hate speech.
- **Recall:** Proportion of correctly identified hate speech instances out of all actual hate speech cases in the data.
- Additional evaluations may involve testing the model on unseen data to assess its ability to generalize to real-world scenarios. We will also consider metrics like F1-score, which balances precision and recall, for a more comprehensive evaluation.

6. Hate Speech Detection:

- The trained model will be deployed to analyze new social media text data in real-time or near real-time. If the model identifies a post or comment as hate speech with a high confidence score, it can be flagged for further review or moderation by human moderators.

7. Model Refinement:

- The performance of the deployed model will be continuously monitored. We will employ techniques like active learning to identify data points where the model is uncertain and retrain the model with these new data points to improve its accuracy over time.

III. CONCLUSION

The pervasiveness of hate speech on social media platforms poses a significant challenge, fostering negativity, harassment, and social division. The sheer volume of content necessitates exploring alternative solutions beyond human moderation capabilities. This research delved into the development of an AI-powered system specifically designed for the early detection of hate speech.

Our system leverages Natural Language Processing (NLP) techniques and machine learning algorithms to analyze social media text data. We explored techniques like sentiment analysis, text classification, and feature engineering to extract meaningful features that aid in hate speech identification. An SVM model was trained on a labeled dataset to learn patterns within the data indicative of hate speech.

This research acknowledges the challenges associated with AI-based hate speech detection, such as the difficulty in understanding sarcasm and context, cultural variations in hate speech language, and potential biases within AI models. We propose methods for addressing these challenges, such as utilizing diverse datasets and employing human-in-the-loop approaches for final moderation decisions.

The potential benefits of this AI system are significant. Early detection of hate speech allows for faster intervention and content moderation before hateful content can cause harm. This ultimately contributes to fostering a safer and more inclusive online environment where everyone feels welcome to participate in meaningful dialogue.

However, it is crucial to acknowledge that AI systems are not a silver bullet solution. Continued research is necessary to refine the accuracy and effectiveness of these systems, while ensuring their ethical implementation that respects freedom of speech and avoids censorship. By fostering collaboration between researchers, developers, and social media platforms, we can harness the power of AI to create a more civil and respectful online discourse.

REFERENCES

1. Lowlesh Yadav and Asha Ambhaikar, “IOHT based Tele-Healthcare Support System for Feasibility and performance analysis,” Journal of Electrical Systems, vol. 20, no. 3s, pp. 844–850, Apr. 2024, doi: 10.52783/jes.1382.
2. L. Yadav and A. Ambhaikar, "Feasibility and Deployment Challenges of Data Analysis in Tele-Healthcare System," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), Raipur, India, 2023, pp. 1-5, doi: 10.1109/ICAIIHI57871.2023.10489389.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305026 |

3. L. Yadav and A. Ambhaikar, "Approach Towards Development of Portable Multi-Model Tele-Healthcare System," 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), Raipur, India, 2023, pp. 1-6, doi: 10.1109/ICAIIHI57871.2023.10489468.
4. Lowlesh Yadav and Asha Ambhaikar, Exploring Portable Multi-Modal Telehealth Solutions: A Development Approach. International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC), vol. 11, no. 10, pp. 873–879, Mar. 2024.11(10), 873–879, DOI: 10.13140/RG.2.2.15400.99846.
5. Lowlesh Yadav, Predictive Acknowledgement using TRE System to reduce cost and Bandwidth, March 2019. International Journal of Research in Electronics and Computer Engineering (IJRECE), VOL. 7 ISSUE 1 (JANUARY- MARCH 2019) ISSN: 2393-9028 (PRINT) | ISSN: 2348-2281 (ONLINE).
6. Davidson, T., Warber, J., Garcia, D., & Macy, M. (2017). **Automated Hate Speech Detection and the Problem of Offensive Language**. In Proceedings of the Eleventh International Conference on Web and Social Media (ICWSM 2017). https://arxiv.org/abs/1703.04009
7. Fanger, B., Guy, I., & Carmel, L. (2020). **VADER: Valence Aware Dictionary and sEntiment Reasoner**. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2020). https://swayanshu.medium.com/vader-valence-aware-dictionary-and-sentiment-reasoner-sentiment-analysis-28251536698
8. Gupta, A., Mahalanobis, A., & Rajakrishnan, M. (2017). **Byeol: An Attention-Based End-to-End Learning Framework for LSTM Networks**. In Proceedings of the 2017 Long Document Retrieval Conference (LDRC 2017). https://openreview.net/pdf/ZY9xwl3PDS5Pk8ELfEzP.pdf (for attention-based techniques)
9. Hate Speech Detection Task - Kaggle. (n.d.). Retrieved April 20, 2024, from https://www.kaggle.com/code/eabdul/hate-speech-detection
10. Huang, X., Sun, S., Ma, J., Liu, Z., & Zhao, S. (2019). **Convolutional Attention Networks for Sentence Classification**. IEEE Transactions on Knowledge and Data Engineering, 31(5), 905-917. https://ieeexplore.ieee.org/document/9345092/
11. Johnson, T. (2017). **Sentiment Analysis of Social Media**. Morgan & Claypool Publishers.
12. Kontokalidis, D., Papadopoulos, L., & Kompatsiaris, Y. (2019). **Learning Deep Representations for Effective Hate Speech Detection**. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM 2019). https://arxiv-sanity-lite.com/?rank=pid&pid=2210.10839
13. Lieberman, M., Wang, T., Tajbakhsh, N., & Atrey, P. M. (2018). **Suicidal Ideation Detection in Social Media Using Deep Learning**. Natural Language Engineering, 24(1), 307-322. https://www.mdpi.com/1660-4601/19/19/12635 (for transfer learning applications)
14. Maynard, D., & Munro, R. (2016). **What is Hate Speech?**. Philosophy Compass, 11(4), 290-302.
15. Nobata, C., Jakobsson, J., Viradiya, N., Darlow, A., & Chandrasekaran, R. (2016). **Automatic Detection of Cyberbullying on Social Media**. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details