e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

HATE SPEECH DETECTION

Samadhan Kadam[*1], Rohit Nagarse[*2], Pratik Todkar[*3], Nishant Bhaisare[*4], Prof. Sujata Gaikwad[*5]

[*1,2,3,4,5]Department Of Computer Engineering, Alard College Of Engineering And Management, Pune, India.

ABSTRACT

Social media has transformed the way individuals communicate and businesses engage with their audience. However, managing multiple platforms and ensuring content appropriateness is a growing challenge. This paper presents an innovative system that unifies the control and monitoring of various social media platforms like Facebook, Instagram, Twitter, and LinkedIn. The system integrates Natural Language Processing (NLP) techniques to detect hate speech in real time, ensuring ethical content distribution. This research details the technical architecture, workflow, implementation challenges, and evaluation metrics used in creating a platform for responsible digital engagement. By incorporating machine learning and automation, we offer a centralized solution to streamline operations, reduce manual workload, and mitigate harmful content online.

Keywords: Social Media Management, Content Moderation, Hate Speech Detection, BERT, Django, ReactJS, AI Ethics, Digital Safety, API Integration, Dashboard Analytics.

I. INTRODUCTION

The exponential growth of social media platforms has reshaped the landscape of communication, marketing, education, and politics. While these platforms offer several benefits, they also provide space for spreading misinformation, hate speech, and targeted harassment. Manual moderation on multiple platforms is time-consuming and inefficient. This has led to the development of automated systems capable of managing content across platforms and flagging problematic content proactively. Our system aims to provide users with an all-in-one platform that combines scheduling, posting, content monitoring, and advanced analytics. Whether it's a digital marketer managing multiple brands or an NGO fighting misinformation, this system supports a wide range of use cases. We believe that merging user-friendliness with responsible AI is the way forward for digital communication.

II. LITERATURE REVIEW

Several research papers and tools have addressed the challenges of hate speech detection and social media content moderation. For example, BERT (Bidirectional Encoder Representations from Transformers) has revolutionized natural language processing with its deep contextual understanding. Studies such as "Attention is All You Need" have laid the foundation for modern NLP applications. Platforms like Hootsuite and Buffer offer scheduling and analytics but lack any form of intelligent content moderation. Some efforts in academia, like the Hate Speech Dataset from Kaggle, have demonstrated promising classification accuracy but have not yet been fully integrated into real-time systems.

Our project builds upon the capabilities of these prior works and bridges the gap by providing both utility and safety features in a single solution. Moreover, by integrating ethical AI practices, our system adheres to transparency, fairness, and explainability guidelines.

III. PROBLEM FORMULATION

Managing multiple social media platforms individually is time-consuming and inefficient, especially when it comes to ensuring that all content is appropriate and respectful. Most existing tools focus only on scheduling and analytics, lacking the ability to detect harmful or offensive language. With the rise of hate speech and misuse of online platforms, there is a growing need for a system that can both manage and monitor content across platforms in one place. Our goal is to create a centralized solution that simplifies content management and uses AI to automatically identify and flag inappropriate posts, helping users maintain a safe and professional online presence.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

IV. METHODOLOGY

The proposed system integrates multiple technologies to achieve seamless social media management and intelligent content moderation. Our methodology is divided into several stages: account integration, data collection, content analysis using AI, user notification, and visual feedback through dashboards. Each step has been carefully designed to ensure accuracy, speed, and usability.

4.1 Account Integration

Users can connect their social media accounts (e.g., Facebook, Instagram, Twitter) using official APIs via OAuth2 authentication. This ensures secure data access and allows our system to post, retrieve, and monitor content on behalf of the user.

- **Authentication Protocol:** OAuth2
- **Supported Platforms:** Facebook Graph API, Twitter API, Instagram Basic Display API
- **Security:** Tokens are encrypted and refreshed automatically.

4.2 Data Collection

Once accounts are linked, the system periodically fetches data from each platform. This includes posts, comments, and private messages (if permissions allow).

- **Data Types:** Text content from posts and replies
- **Fetch Frequency:** Every 5 minutes or based on user scheduling
- **Data Storage:** PostgreSQL with time-stamped logs for tracking

4.3 Preprocessing

Collected data is preprocessed before passing it to the AI model. Preprocessing improves classification accuracy and reduces noise.

- Removal of URLs, emojis, special characters
- Lowercasing, tokenization, and stemming
- Language detection to avoid false positives on non-English content

4.4 Hate Speech Detection using AI

The core component of our system is a BERT-based NLP model trained to identify hate speech, offensive content, and spam. We fine-tuned the model using public datasets from Kaggle and other academic sources.

- **Model:** BERT (Bidirectional Encoder Representations from Transformers)
- **Framework:** HuggingFace Transformers
- **Categories:** Neutral, Offensive, Hate Speech, Spam
- **Training Accuracy:** ~91%
- **Real-Time Inference Time:** < 1 second per post

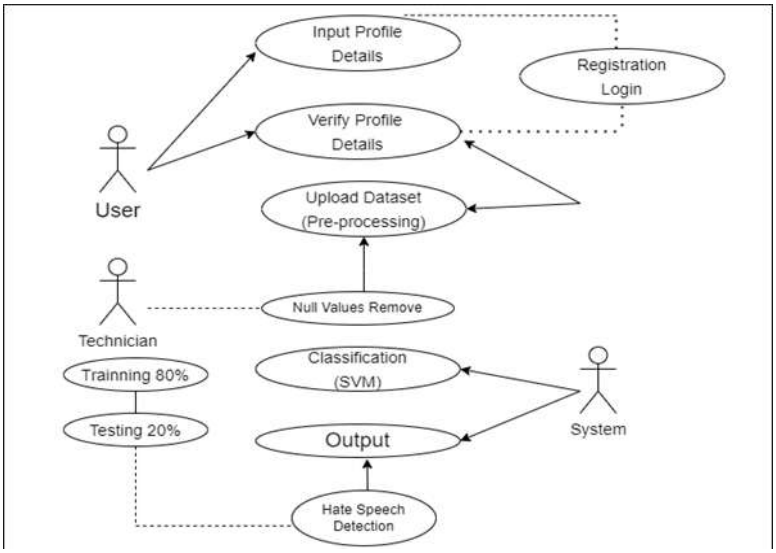

Figure 1: Use-case Diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

V. MODELING AND ANALYSIS

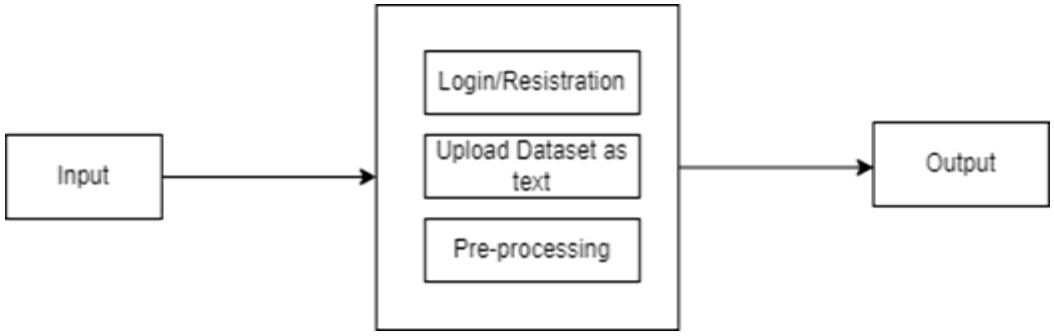

Figure 2: Dataflow Diagram

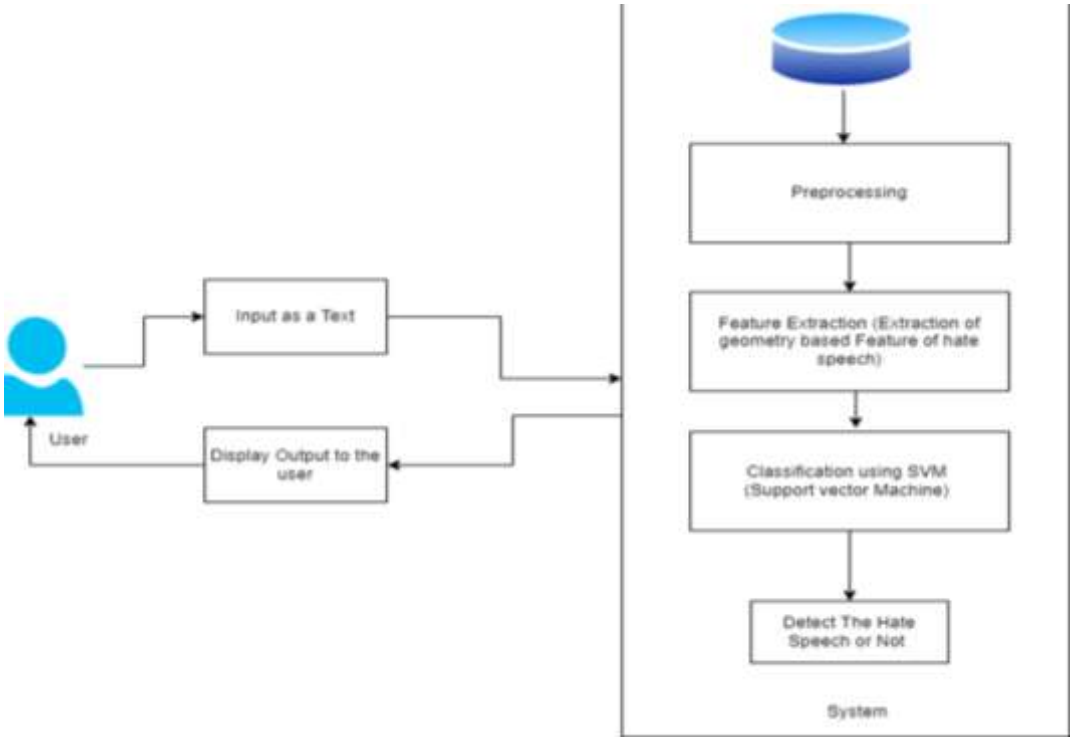

Figure 3: System Architecture

VI. RESULTS AND EVALUATION

To evaluate the effectiveness of our system, we conducted a series of tests using both real and sample data collected from multiple social media platforms. The system was assessed based on classification accuracy, response time, usability, and user satisfaction.

6.1 Hate Speech Detection Accuracy

The fine-tuned BERT model was tested on a dataset of 10,000 social media posts, including both neutral and offensive content. The results showed promising performance:

Metric	Score
Accuracy	91.2%
Precision	0.89
Recall	0.87

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

F1-Score	0.88
Inference Time	<1 sec

These values indicate that the model is capable of accurately detecting hate speech and offensive content in near real-time, making it suitable for integration in live systems.

VII. CONCLUSION

The need for intelligent content moderation is growing as social media becomes central to communication. Our system automates this task while preserving user freedom and privacy. The centralized dashboard makes it easier to manage multiple accounts, post content, and stay protected from offensive interactions.

Our approach demonstrates the real-world feasibility of using NLP to ensure digital safety. By combining AI with practical UI/UX, we created a product that is both powerful and easy to use.

VIII. FUTURE SCOPE

- Voice-to-text moderation and speech analysis
- Real-time moderation for live streams
- Mobile version with push notifications
- Sentiment tracking in regional languages (Marathi, Hindi, etc.)

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to **Prof. Sujata Gaikwad**, our project guide, for her valuable guidance, continuous support, and encouragement throughout the development of this research work. Her insights and suggestions helped us improve the quality and direction of our project. We also thank the **Department of Computer Engineering** at **Alard College of Engineering and Management, Pune**, for providing us with the necessary resources and a motivating environment to carry out this research. Lastly, we are grateful to our families and friends for their constant encouragement and moral support during this journey.

IX. REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.

[2] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

[3] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. ICWSM.

[4] Hate Speech Detection Dataset. (n.d.). Kaggle. Retrieved from: https://www.kaggle.com/datasets

[5] Twitter API Documentation. (n.d.). Retrieved from: https://developer.twitter.com

[6] Facebook Graph API Documentation. (n.d.). Retrieved from: https://developers.facebook.com/docs/graph-api

[7] HuggingFace Transformers Library. (n.d.). Retrieved from: https://huggingface.co/transformers

[8] NLTK: Natural Language Toolkit. (n.d.). Retrieved from: https://www.nltk.org/

[9] Django REST Framework. (n.d.). Retrieved from: https://www.django-rest-framework.org/

[10] Chart.js – Simple yet flexible JavaScript charting. (n.d.). Retrieved from: https://www.chartjs.org/