ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

REPHRASE AI : A SENTIMENT AWARE APPROACH TO ONLINE CIVILITY ENHANCEMENT

CH NagaJyothi

21N81A6767

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Nadergul, Hyderabad,501510

chikirapallynagajyothigmail.com

M Keerthi

21N81A6780

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Nadergul, Hyderabad,501510

kirticou03@gmail.com

G Sudheeshna Reddy

21N81A6793

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Nadergul, Hyderabad,501510

gaddamsudheeshnareddy@gmial.com

P Abhijith Reddy

21N81A67B3

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Nadergul, Hyderabad,501510

abhijithreddy0703@gmail.com

Mr Mohd Miskeen Ali

Assistant Professor

Computer Science and Engineering (Data-Science)

Sphoorthy Engineering College,

Nadergul, Hyderabad,501510

info.miskeen@gmail.com

ABSTRACT: Social media platforms increasingly face challenges from toxic and offensive language, which can lead to cyberbullying, emotional harm, and reduced user engagement. To address this, RephraseAI offers a web-based comment moderation tool that leverages Natural Language Processing (NLP), accurate toxic comment classification, assigning a toxicity score and model confidence score to each input. Through integrated sentiment analysis and intelligent text generation, the tool rewrites harmful content into respectful alternatives while preserving the original meaning. A Flask backend ensures fast, performance. Additionally, the (NLP), Machine Learning, and Deep Learning to detect and rephrase toxic comments in real time. The system employs transformer-based models like BERT from Hugging

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

Face Transformers to perform accurate toxic comment classification, assigning a toxicity score and model confidence score to each input. Through integrated sentiment analysis and intelligent text generation, the tool rewrites harmful content into respectful alternatives while preserving the original meaning. A Flask backend ensures fast, secure processing with robust user input sanitization. Designed with a user-friendly interface, RephraseAI is suitable for integration into social media platforms, educational forums, and other interactive environments, promoting healthier online communication through real-time, ethical, and scalable comment moderation.

Keywords: *Toxic Comment Classification, Natural Language Processing (NLP), Sentiment Analysis, Text Rephrasing, Machine Learning, Hugging Face Transformers, Deep Learning, Text Generation, User Input Sanitization, Model Confidence Score, Toxicity Score, Comment Moderation Tool*

1. INTRODUCTION

In today's hyperconnected digital era, the internet has become a powerful tool for global interaction, knowledge sharing, and social engagement. Platforms like social media networks, online forums, comment sections, and collaborative discussion portals have become central to how people communicate and express themselves . However, alongside these benefits has emerged a significant challenge toxic and offensive language. The anonymity and speed of online interactions have made it easier for individuals to post harmful, a disrespectful, or emotionally damaging comments without immediate accountability. These toxic comments not only affect the well-being of users but also degrade the quality of online discourse, leading to an increase in cyberbullying, reduced user engagement, and erosion of trust in digital platforms. Moderation tools aim to detect and manage inappropriate or hostile language, helping platform administrators maintain community standards and ensure user safety. However, most existing systems rely heavily on static keyword filters, rule-based detection, or human moderators. These approaches can be limited in scalability, accuracy, and adaptability. , RephraseAI is not just a comment moderation tool it is a step toward more respectful, inclusive, and meaningful communication online. By integrating the latest advancements in Natural Language Processing, Machine Learning, and Deep Learning, and combining them with a strong focus on sentiment analysis, text generation, and toxicity management, RephraseAI reimagines how digital platforms can respond to online hostility. Its design reflects the evolving need for intelligent, ethical, and user-centered moderation systems that go beyond detection to promote positive transformation in language and behavior.As the digital world becomes more central to personal, educational, and professional interactions, tools like RephraseAI will play a vital role in shaping the quality of online discourse. With its unique blend of cutting-edge AI and human-centric design, RephraseAI offers a practical and forward-thinking solution to one of the most urgent challenges in digital communication today.

2. LITERATURE REVIEW

[1] Kubade, **Rewati A. Kawale**, **Ishika S. Kahar**, **Jatin P. Bais**, **Adarsh A. Nimkar**, and **Manish V. Dhoble** published in May 2024, focuses on the pressing issue of toxic language in digital environments. Their paper, *"Toxic Comment Classification,"* proposes a machine learning-based

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

system designed to identify harmful online content and enhance digital safety. Recognizing the shortcomings of traditional moderation methods like manual reporting or keyword filters, the authors adopt a data-driven approach that leverages supervised machine learning models for more accurate and scalable toxic comment detection.The methodology follows a structured machine learning pipeline, starting with data preprocessing to clean and transform raw text into a machine-readable format. Techniques such as tokenization, stop-word removal, and TF-IDF vectorization are employed to extract meaningful features from the text. These features are then fed into classification algorithms primarily Logistic Regression and Support Vector Machines (SVMs) which are widely used for text classification due to their balance between performance and interpretability. The model is trained to differentiate between toxic and non-toxic comments, and possibly even between different types of toxicity such as insults, threats, or hate speech.

Evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure the effectiveness of the models. The authors acknowledge the challenges posed by class imbalance, which is common in toxicity datasets where non-toxic comments often dominate. Therefore, they stress the importance of using balanced metrics rather than relying on accuracy alone. The reported results suggest that their system performs reliably and could be deployed in real-world applications like comment sections or online forums to automatically flag or block harmful content.While the study presents a solid foundation for toxic comment classification, it has certain limitations. Primarily, the system is designed only to detect and block toxic content, with no provision for guiding users toward respectful communication. It lacks functionality for rephrasing or providing constructive alternatives to offensive messages. Moreover, the use of traditional ML models, while practical, limits the system's ability to understand context or handle subtle toxicity such as sarcasm. The model also does not adapt to different user tones or cultural contexts, making it rigid in its moderation decisions.

[2] Navoneel Chakrabarty, in his 2019 study titled *"A Machine Learning Approach to Comment Toxicity Classification"*, which addressed the growing challenge of identifying toxic content in online platforms. His research utilized the Jigsaw Wikipedia Comment dataset, a widely adopted benchmark for toxicity detection, which contains millions of annotated comments labeled with multiple categories of toxicity such as identity-based hate, obscenity, threats, insults, and general toxicity.Chakrabarty's key innovation was the use of a six-headed machine learning architecture based on TF-IDF (Term Frequency-Inverse Document Frequency) feature extraction. TF-IDF is a classical natural language processing technique that converts textual data into weighted numerical vectors, highlighting words that are particularly informative in distinguishing different comment categories. Each "head" in his model was trained separately as a binary classifier, targeting a specific type of toxicity. This modular design allowed the model to specialize in detecting nuances associated with different toxic behaviors rather than treating toxicity as a monolithic category.The ensemble of these classifiers collectively enhanced the robustness and accuracy of the system. The reported performance metrics were impressive, with a mean validation accuracy of 98.08% and an absolute validation accuracy of 91.61%. These results underscored the

power of combining multiple specialized classifiers and classical machine learning techniques to effectively identify toxic comments, even without resorting to more complex deep learning architectures.However, the study also implicitly pointed to challenges such as handling subtle and context-dependent toxicity and the ethical implications of automated moderation.

Despite its success, Chakrabarty's approach is limited by its reliance on traditional TF-IDF features and separate classifiers that do not capture deeper semantic understanding or contextual nuances of language. Additionally, while the model excels at detection, it does not address post-detection user engagement, such as encouraging better

[3] C. Lakshmi, **D. Ananya**, **M. Sreevani**, and **M. Sreedevi**, in their 2021 research, developed a multi-headed machine learning framework aimed at advancing toxicity detection in social networking environments. Their research was grounded in the use of the Kaggle Jigsaw dataset, a comprehensive resource similar to the Wikipedia Comments dataset, but designed to challenge models with various toxic comment categories and multilabel classification tasks. Their framework explored the impact of both word-level and character-level inputs, recognizing that toxic content often includes misspellings, slang, or obfuscated offensive words that can bypass simplistic keyword detection. The use of character-level features enabled the model to be more resilient to such manipulations, improving detection accuracy.

However, despite these technical advancements, their system primarily focuses on detection and restriction of toxic content either by flagging or blocking. The framework lacks mechanisms for interpretability at the user level, personalization of moderation feedback, or guidance to promote healthier communication patterns. Like many toxicity detection systems, it does not engage users constructively, missing an opportunity to reduce toxicity through education or feedback loops.The study thus highlights a persistent gap in toxicity moderation research: while detection models become increasingly accurate, future systems need to evolve beyond mere identification to foster positive user interactions and enhance transparency and user trust.

Multi-headed machine learning framework combining logistic regression and Long Short-Term Memory (LSTM) networks. They trained their models on the Kaggle Jigsaw dataset and explored both word-level and character-level inputs for better classification results. Their system was designed to detect a broad range of toxic behaviors, using binary and multi-label classification schemes to improve performance and accuracy in social networking environments.Despite the progress achieved by these systems, they primarily focus on detection and restriction—either blocking or flagging toxic content. They rarely guide users toward better communication or offer constructive feedback. Most systems lack interpretability and personalization, which can hinder user experience and acceptance.

3. METHODOLOGY

The methodological framework of **RephraseAI** revolves around a two-phase Natural Language Processing (NLP) system that addresses a critical gap in online communication: the need to not only detect toxic content but also rephrase it into a more civil and respectful form. Traditional content moderation systems usually rely on deletion or blocking, which can frustrate users and do little to encourage respectful

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

engagement. RephraseAI aims to reshape this process by using intelligent, interpretable, and real-time moderation strategies powered by advanced machine learning and deep learning techniques.The project uses a blend of classical NLP tasks, transformer-based deep learning models, and web development technologies to create a system that is scalable, user-friendly, and ethically aware.

1.Toxicity Detection: The first phase in RephraseAI's pipeline is identifying toxic or harmful language in user-generated comments. This step ensures that inappropriate content is recognized before it is published, thus maintaining the health of online platforms.For the task of toxic comment classification, the system uses the **unitary/toxic-bert** model a transformer-based architecture derived from BERT (Bidirectional Encoder Representations from Transformers). BERT is known for its bidirectional training of transformers on masked language modeling and next sentence prediction tasks, allowing it to capture deep contextual relationships in text.Tokenization involves converting the raw input text into numerical tokens that the model can understand. These tokens are then passed through the model to obtain toxicity predictions.model inference, the output toxicity score is compared against a predefined threshold (e.g., 0.5). Based on this comparison

- If the score is below the threshold, the comment is marked non-toxic and shown to the user.
- If the score is above the threshold, the comment is flagged as toxic and passed to the rephrasing phase.

2.Comment Rephrasing: Once a comment is identified as toxic, the second phase is triggered: Rephrasing the input to maintain the semantic meaning but in a polite, non-offensive manner. This component addresses the core limitation of most existing systems, which typically do not offer any feedback or reformative action.For rephrasing, the system can use a transformer-based text generation model (such as GPT-2, T5, or a distilled version fine-tuned on paraphrasing datasets). The goal is to generate text that:

- Preserves the intent of the original comment
- Removes harmful, aggressive, or disrespectful language
- Aligns with ethical and civil communication norms
- Altering sentence structure to reduce negative sentiment
- Removing second-person attacks (e.g., "You are stupid") and replacing with passive or neutral tone

  Such rules provide a backup to handle simple toxic constructs efficiently without always relying on computationally expensive models.

Disadvantages:

I**n** existing system the issues they face are :

1. **Over-blocking**:
   Harmless comments get blocked due to key-word filters, without considering the context of the message.
2. **Under-detection**:
   Fails to catch toxic comments that are subtly written, sarcastic, or emotionally nuanced.

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

3. **UserGuidance**:
   The system removes or blocks content without educating users on how to communicate more respectfully.

Proposed System:

To overcome the limitations of existing systems, the proposed system RephraseAI offers a novel approach that not only detects toxic comments but also rephrases them into polite, respectful alternatives in real time. This dual-functionality sets it apart by promoting ethical communication without censorship.The application is built using Flask as the backend framework and integrates Hugging Face Transformers for deep NLP capabilities. It uses a fine-tuned BERT-based model to detect toxic comments and classify them with high accuracy. If a comment is deemed toxic, it is passed through a text generation model that rewrites the statement in a non-toxic manner, preserving the original intent but removing offensive language.

Key Advantages:

1. Real-Time Detection and Rewriting: Unlike systems that simply block or delay content moderation, RephraseAI detects and rewrites toxic content instantly, enabling smooth, live interaction across digital platforms.

2. Ethical Moderation: Rather than censoring or punishing users, RephraseAI encourages more positive communication by showing users how to reframe their statements constructively.

3. Interpretability: Users can view the toxicity score and model confidence, offering transparency and helping them understand the logic behind the system's decisions.

4. User-Friendly Interface: The web application is built with HTML and CSS for a responsive and simple user experience that works across all modern browsers and devices.

5. Scalability and Flexibility: The system can be deployed on cloud platforms like Heroku or AWS and supports integration with multiple languages and social media APIs.

6. Safe and Constructive Communication: By turning negative comments into positive ones, the tool actively promotes digital well-being and healthier online interactions.

Overall, RephraseAI represents a paradigm shift from punitive to educational and supportive moderation, aligning with modern principles of ethical AI and user empowerment.

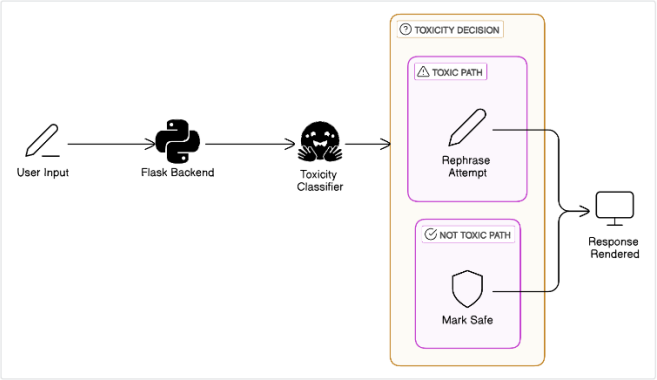

Fig.1: System architecture

MODULES:

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

To implement this project we have designed following modules.

- **User Input Module:** Accepts user comments via a web-based interface.Built using HTML and CSS for simplicity and ease of use.
- **Toxic Comment Detection Module :**Uses a pretrained BERT model from Hugging Face.Calculates toxicity and confidence scores for each input.
- **Text Rephrasing Module :** Activates if a comment is toxic.Uses a text generation model (e.g., T5) to rewrite the comment politely.
- **Flask Backend Module :** Manages routing, logic, and model interactions.Connects frontend and machine learning models.
- **Output & Feedback Module :** Displays either a "safe" comment or the rephrased version.Shows toxicity and confidence scores to the user.
- **Model Integration & Preprocessing Module :** Loads pretrained models and prepares inputs.Handles tokenization and formatting for inference.

4. IMPLEMENTATION

The implementation of **RephraseAI** combines modern web development tools with advanced Natural Language Processing (NLP) models to build a real-time toxic comment detection and rephrasing system. The application is divided into multiple modules, each responsible for a specific task to ensure a responsive, lightweight, and scalable experience.

1. Backend – Flask Framework

The backend is built using **Flask**, a Python web framework that manages routing, API endpoints, and business logic. Flask handles incoming user comments and connects them to the machine learning models. Its integration with Python-based libraries makes it ideal for this AI-driven application, ensuring fast and reliable processing.

2. Toxic Comment Detection Module

This module uses the **unitary/toxic-bert** model from Hugging Face to analyze user input. It classifies comments as toxic or non-toxic and returns a **toxicity score** and **confidence score**. These outputs help determine whether the comment should be rephrased and provide transparency to users.

3. Comment Rephrasing Module

If a comment is marked toxic, it is sent to the **rephrasing module**, which uses a **T5-base** model or a similar text-to-text generation model. This model rewrites the offensive comment into a respectful version while maintaining the original meaning. This approach supports constructive communication instead of simply censoring content.

4. Frontend – HTML/CSS with Bootstrap

The user interface is created using **HTML** and **CSS**, with optional styling from **Bootstrap**. It is designed to be clean, responsive, and accessible on all devices. Users can input comments, view results instantly, and understand the feedback clearly through simple design elements.

5. Optional Authentication – Flask-Login

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

If needed, **Flask-Login** can be integrated for user authentication. It allows login/logout and session tracking, which is useful for personalized experiences or when scaling the app for multi-user environments. While not mandatory for the basic version, it adds flexibility for future growth.

6. System Flow

The process begins when a user inputs a comment. Flask sends it to the toxicity detection model. If it's non-toxic, the comment is shown back as safe. If toxic, it is passed to the rephrasing model, which generates a polite version. The final output is then displayed along with toxicity and confidence scores for clarity.

7. Development and Deployment

Initially, RephraseAI runs locally for testing and development. For broader access, it can be deployed to cloud platforms like **Heroku**, **Render**, or **AWS**. This allows the system to handle more users and makes it suitable for integration with live comment streams on websites or social platforms.

5. EXPERIMENTAL RESULTS

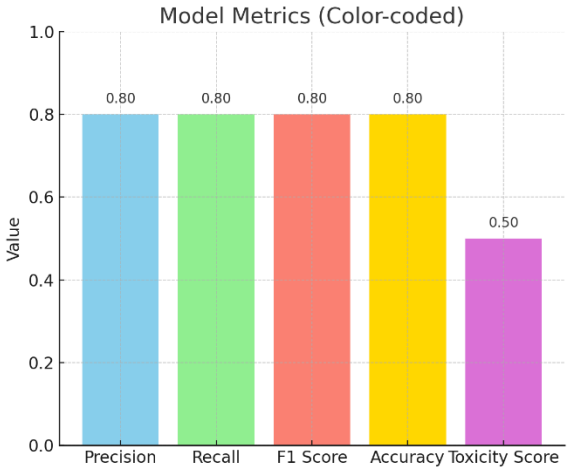

Fig.2: Graph

Precision : 0.80

Recall : 0.80

F1 Score:0.80

Accuracy:0.80

Toxicity Score:0.50

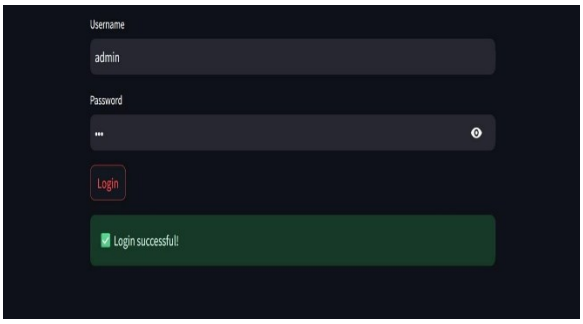

Fig.3: Output screen

Fig.4: Output screen

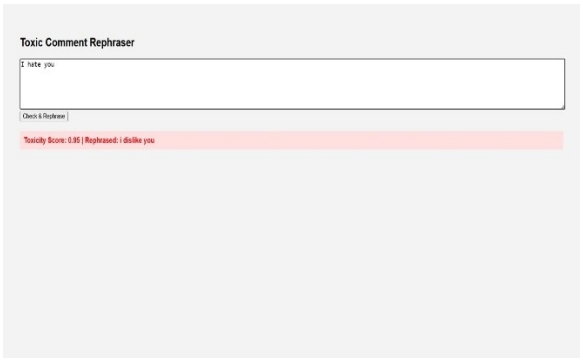

Fig.5: Output screen

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

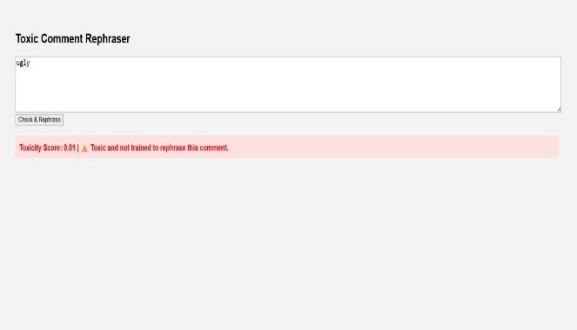

Fig.6: Output screen

6. CONCLUSION

In conclusion, RephraseAI revolutionizes online content moderation by shifting from censorship to constructive communication. Instead of deleting or blocking toxic content, it uses advanced Natural Language Processing (NLP), deep learning, and ethical AI principles to detect harmful language and offer real-time, respectful rephrasings. This approach preserves the user's intent while promoting empathy, accountability, and thoughtful dialogue. By transforming negative expressions into constructive ones, RephraseAI educates users and fosters a culture of emotional intelligence rather than fear or suppression. The system accurately identifies various forms of toxicity—such as hate speech, insults, and cyberbullying—using AI models trained on diverse datasets. Its transparent interface shows users suggested edits with explanations, encouraging learning and self-reflection. Scalable and easy to integrate, RephraseAI fits seamlessly into social media platforms, messaging apps, forums, and customer support tools. Its interpretable architecture supports trust and compliance by allowing stakeholders to understand how decisions are made. By enhancing digital civility without compromising freedom of speech, RephraseAI enables users to express themselves responsibly while contributing to safer and more inclusive online communities. With its real-time functionality and user-centric design, RephraseAI is well-suited for the fast-paced nature of modern digital interactions, making it a powerful tool for improving user experience.

REFERENCES

[1] H. M. Kubade, R. A. Kawale, I. S. Kahar, J. P. Bais, A. A. Nimkar, and M. V. Dhoble, "Toxic Comment Classification," International Journal of Creative Research Thoughts (IJCRT), May 2024.

[2] N. Chakrabarty, "A Machine Learning Approach to Comment Toxicity Classification," arXiv preprint arXiv:1903.06765, Feb. 2019.

[3] C. Lakshmi, D. Ananya, M. Sreevani, and M. Sreedevi, "Toxic Comments Classification in Social Networking," International Journal, Feb. 2021.

[4] S. Ahmed, J. F. Esha, M. S. Rahman, M. S. Kaiser, A. S. M. S. Hosen, D. Ghimire, and M. J. Park, "Exploring deep learning and machine learning approaches for brain hemorrhage detection," IEEE Access, vol. 12, pp. 45060-45083, Mar. 2024.

[5] S. Gilda, M. Silva, L. Giovanini, and D. Oliveira, "Predicting different types of subtle toxicity in unhealthy online conversations," Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 2021, pp. 1–10.

[6] É. Brassard-Gourdeau and R. Khoury, "Impact of sentiment detection to recognize toxic and subversive online comments," Proceedings of the 2018 IEEE International Conference on Big Data, Seattle, WA, USA, 2018, pp. 1–10.

[7] T. Maheshwari, A. N. Reganti, S. Gupta, A. Jamatia, U. Kumar, B. Gambäck, and A. Das, "A societal sentiment analysis: Predicting the values and ethics of individuals by analysing social media content," Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 2017, pp. 731–741.

[8] V. Singh, A. Md. Khan, and A. Ekbal, "Indian Institute of Technology-Patna: Sentiment Analysis in Twitter," Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 2014, pp. 341–345.

[9] R. Badlani, N. Asnani, and M. Rai, "An ensemble of humour, sarcasm, and hate speech for sentiment classification in online reviews," Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), Hong Kong, China, 2019, pp. 337–345.

[10] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," arXiv preprint arXiv:1802.00385, 2018.

[11] O. Kolchyna, T. T. P. Souza, P. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," arXiv preprint arXiv:1507.00955, 2015. [12] P. Duc Tong et al., Brain Hemorrhage Diagnosis by Using Deep Learning. 2017.

[13] V. G. Vinodhini and R. M. Chandrasekaran, "Sentiment analysis and opinion mining: A survey," International Journal of Computer Applications, vol. 47, no. 18, pp. 1–6, 2012.

[14] S. Kiritchenko and S. M. Mohammad, "Examining gender and race bias in two hundred sentiment analysis systems," arXiv preprint arXiv:1805.04508, 2018.

[15] A. Mozafari, D. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," Proceedings of the 12th International Conference on Web and Social Media (ICWSM), 2018, pp. 1–11.

[16] S. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," ACM Computing Surveys, vol. 51, no. 4, pp. 85:1–85:30, July 2018.

[17] F. Z. Mohamed, A. Y. Mokhtar, and A. H. Mohamed, "Sentiment Analysis Techniques and Applications: A Survey," International Journal of Computer Applications, vol. 182, no. 5, pp. 8–16, Dec. 2018.

[18] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM), 2017, pp. 512–515.

[19] M. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proceedings of the NAACL Student Research Workshop, 2016, pp. 88–93.

[20] F. B. Hovy and E. S. Søgaard, "Tagging Performance Correlates with Author Age," Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016, pp. 483–488.

[21] M. K. Kaur and J. S. Nagra, "A survey on sentiment analysis and opinion mining," International

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Journal of Computer Applications, vol. 109, no. 5, pp. 12–18, Jan. 2015.

[22] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, May 2012.

[23] E. Fersini, A. Messina, and F. Falchi, "Automatic Detection of Offensive Language in Tweets: A Survey," Journal of Internet Services and Applications, vol. 9, no. 1, pp. 1–25, Dec. 2018.

[24] K. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," Proceedings of the First Workshop on NLP and Computational Social Science, 2016, pp. 138–142.

[25] Z. Qian, D. R. Medimorec, and T. Caldeira, "Benchmarking hate speech detection: How biases limit model performance," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 9157–9169.