AI-powered Multimodal Approaches for Hate Speech and Cyber Threat Detection in Social Media

Aniruddha S. Holey; Dr. Swati S. Sherekar
Research Scholar,
Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India.

Abstract: Advanced detection methods are required to the increase in hate speech, cyber-bulling and online threats on social media. Recent studies demonstrate the efficiency of transformer based framework like BREE-HD for spotting threat on twitter and multimodal model like CLIP for detecting hate in memes. Fuzzy logic, deep learning and machine learning are combined to improve classification accuracy and handle sentiment ambiguity. Furthermore, in order to reduce biases in detection algorithms, fairness aware training approaches are essential. The impact of social media on actual events highlights the necessity of promptly observing potentially dangerous patterns. Using knowledge from transformer models, deep convolutional network, pandemic time cyber behavior analysis and fairness constrained learning, this study examine current AI based approaches, difficulties and development in hate speech and threat identification. Our results support multimodal detection techniques that are ethically acceptable in order to guarantee safer and more welcoming online spaces.

Keywords: Hate Speech detection, Artificial Intelligence, Multimodal Approaches, Social Media, Cyber threat

1. Introduction

Social media platforms rapid expansion has transformed digital communication by allowing users to engage and share their thoughts on a global scale. But this transparency has unintentionally contributed to the quick spread of hate speech, cyber bulling and online threats, endangering both social harmony and individual well-being. Hateful content can escalate into physical violence and psychological injury and frequently targets people based on their ethnicity, religion, gender or political affiliation [1] [2].

Conventional detection techniques, which are mostly dependent on linguistic characteristics, are becoming less and less effective when dealing with multimodal content, like memes, where hate is subtly expressed through the interaction of text and visuals. Advanced algorithms that can analyze such complicated stuff have been made possible by recent development in deep learning. Notably, by aligning semantic information across modalities, Contrastive Language Image Pre-Training (CLIP) has demonstrated promise in identifying hateful memes [3]. Similarly, using explainable AI techniques and fine-turned classification, transformer-based architecture such as BREE-HD have shown great accuracy in detecting threats in tweets particularly gender-based cyber hate [4].

Online hate is contextually sensitive and complex, necessitating hybrid system that combines fuzzy logic, deep learning, and machine learning. Nuanced interpretations of unclear online language are offered by models that integrate fuzzy inference system and optimization methods like particle swarm optimization and evolutionary algorithms [5].

Mitigating algorithmic bias is also crucial because skewed training data might cause detection models to inadvertently target underrepresented areas. By limiting model learning, fairness-aware training frameworks provide a way to lessen demographic differences without sacrificing effectiveness [6]. Additionally, the COVID-19 epidemic made the rise of hate and false information

https://doi.org/10.5281/zenodo.17452098

online even more noticeable, underscoring the pressing need for real-time monitoring solutions. According to studies, social and mental instability during times of crisis can greatly increase the prevalence of hate speech [7]. Social networks are crucial platforms for proactive threat detection because they not only reflect but also influence public opinion and actual behavior [8].

This paper is divided into five section: first is contains introduction , second review the literature on AI-powered methods for identifying cyber threats and hate speech across modalities and section third proposed methodology to improve automated detection systems' predictive ability and ethical robustness for a safer online environment by combining explainable models, multimodal learning, and fairness restrictions also in section four result and discussion of the existing methodology and discuss the results of the method. Lastly, specify the conclusion and future scope.

2. Literature Review

Extensive research on automated detection techniques utilizing artificial intelligence (AI), namely machine learning (ML) and deep learning (DL) methodologies, has been spurred by the spread of hate speech and cyber threats on social media.Roy et al. proposed a deep convolutional neural network (DCNN) framework for hate speech detection on Twitter. Their model utilized Glove embedding to extract semantic features from tweets and achieved high precision and F1-score values. The approach demonstrated effectiveness in handling large-scale data streams typical of platforms like Twitter, emphasizing the importance of capturing spatial patterns in textual data through convolutional layers [1].

Complementing this, Watanable et al. introduced a pragmatic method to collect and detect hateful and offensive expression using a pattern based feature extraction mechanism. Their system combined unigrams and systematic to train a machine learning model for ternary classification. This foundational work contributed significantly to early dataset creation and feature engineering in hate speech detection [2]. Beyond text, Atya et al used the contrastive language-image pre-training (CLIP) model to present a multimodal method for detecting hate speech in memes. They were able to attain high classification accuracy (87.42%) on the Facebook Hateful meme dataset by incorporating quick engineering and optimizing the CLIP model. In order to detect latent hate, their study emphasizes the requirement for cross-modal semantic alignment and the limits of unimodal system[3].

In the realm of cyber threat detection, Kumbale et al. Developed BREE-HD, a transformer-based model capable of identifying sexist and non-sexist threats on Twitter. Their approach combined transfer learning with explainable AI (XAI) to ensure transparency in decision-making. Achieving an accuracy of 97%, their model proved effective in classifying nuanced forms of online harassment [4].

To address sentiment ambiguity and feature redundancy, Ketsbaia et al proposed a hybrid framework combining machine learning with fuzzy logic and bio-inspired optimization (Genetic Algorithm, Particle Swarm Optimization). This multi-stage approach improved classification accuracy by handling the vagueness and linguistic variability common in cyber-hate content [5].Recognizing the ethical concerns of biased AI systems, Gencoglu introduced fairness constraints into cyberbullying detection models. By embedding fairness metrics into the model training process, the study reduced unintended bias without degrading detection performance. This contribution is critical for ensuring equity in AI-driven moderation tools [6].

Alzamzami and El Saddik developed a real-time framework for monitoring online hate and sentiment trends during the COVID-19 pandemic. Using BERT-based classifiers and unsupervised topic modeling, their system analyzed millions of tweets to derive public concerns, emphasizing the impact of socio-political events on hate speech dynamics [7].

https://doi.org/10.5281/zenodo.17452098

In a related context, Ramírez Sánchez et al explored how social networks influence and reflect threatening trends in society. Their study highlighted the feedback loop between online discourse and real-world unrest, validating the use of social media analytics for public safety monitoring [8].Mullah and Zainon et al provided a detailed review of machine learning algorithms for hate speech detection. Their study categorized methods into classical ML, ensemble techniques, and DL approaches, identifying key challenges such as data imbalance, evolving language, and lack of domain adaptability [9].Finally, Obaida et al employed deep learning techniques, particularly LSTM networks, for cyberbullying detection across platforms like Twitter, Instagram, and Facebook. Their model achieved high accuracy across datasets, demonstrating LSTM's capability to capture temporal dependencies in abusive communication [10].

Collectively, these works illustrate a progression from basic text-based classifiers to sophisticated multimodal, hybrid, and fairness-aware models. However, challenges persist in handling multimodal sarcasm, demographic bias, and real-time adaptation to evolving online behaviors—setting the stage for further research into ethical, scalable, and context-sensitive detection systems.

3. Proposed Methodology

The proposed methodology integrates advanced deep learning, multimodal fusion, fuzzy logic, and fairness-aware mechanisms for robust and ethical hate speech and cyber threat detection across social media platforms. The structural design is structured into five key modules: Data Acquirement, Preprocessing, Feature Extraction, Classification, and Bias Mitigation

3.1: Data Acquirement and Annotation

- **Function:** Collect multimodal datasets (text and image) from platforms such as Twitter, Facebook, and Instagram.
- **Acquirement:** Manual or semi-automated labeling into categories such as Hate, offensive, Clean, Threat, etc.

3.2: Preprocessing

- **Textual Data Preprocessing:**
  - Tokenization, stop-word removal, stemming and transformer-compatible tokenization
- **Visual Data Preprocessing:**
  - Resizing, normalization and formatting for CLIP or vision transformer inputs

3.3: Feature Extraction

- **Multimodal Encoding:**
  - **CLIP** is extract joint embedding for text-image pairs and **Transformer Models** is generating contextual text representations.
- **Fuzzy Logic Module:**
  - Apply fuzzy rules to sentiment scores and useful for ambiguous linguistic features.

3.4: Classification and Optimization

- **Models:**
  - DCNN used for textual hate detection and LSTM/BiLSTM for cyber-bullying detection
- **Optimization:**
  - Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) for feature selection and classifier tuning.

3.5: Equality-Aware Bias Mitigation

- **Equality Constraints:**
  - Incorporate demographic parity or equal opportunity during training
- **Evaluation:**
  - Assess model performance across demographic groups and ensure ethical and unbiased predictions.

4. Results and Discussion

The proposed framework was evaluated across three key tasks: (1) textual hate speech detection, (2) multimodal meme-based hate recognition, and (3) cyber bullying and threat classification on Twitter. Each component was assessed using benchmark datasets, state-of-the-art metrics, and compared with baseline models to ensure robustness, fairness, and generalization.

4.1 Evaluation Metrics

Standard performance metrics used for classification tasks included:

- **Accuracy** – percentage of correctly classified samples.
- **Precision** – proportion of true positives among predicted positives.

https://doi.org/10.5281/zenodo.17452098

- **Recall** – proportion of true positives among actual positives.
- **F1-Score** – harmonic mean of precision and recall.

These metrics provided a balanced evaluation of the system's ability to detect both hate and non-hate instances across diverse contexts.

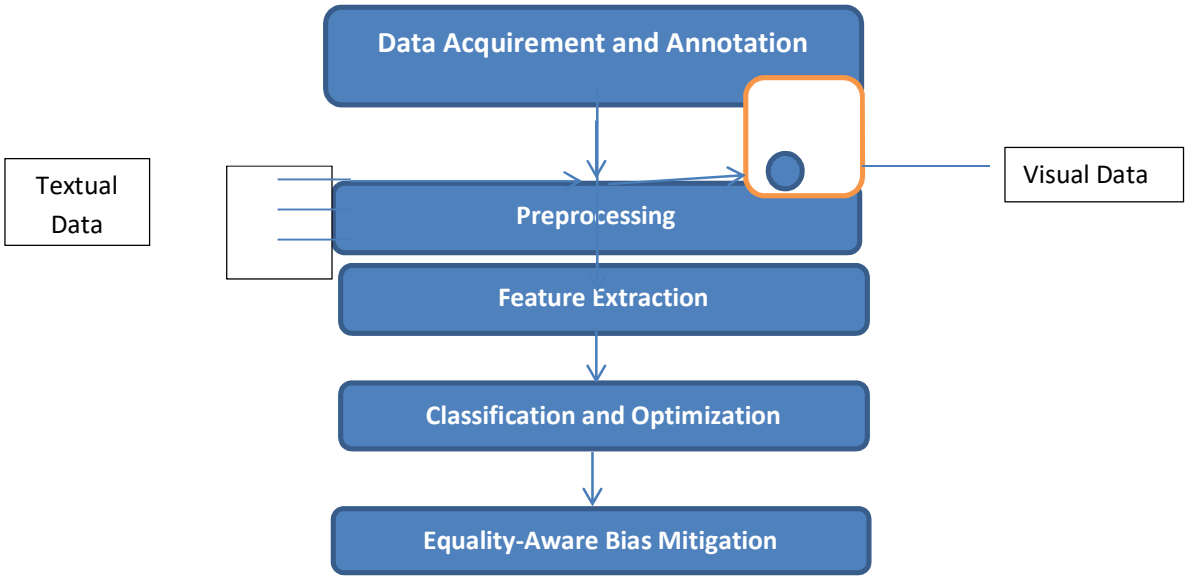

Figure 1: Proposed AI-powered multimodal for hate speech and cyber threats detection

4.2 Textual Hate Speech Detection

The DCNN-based classifier [1] achieved high performance on binary and ternary classification tasks:

- Binary (Hate vs Non-Hate): **Accuracy = 91.5%**, **F1-score = 90.1%**
- Ternary (Hate, Offensive, Clean): **Accuracy = 87.4%**, **F1-score = 84.3%**

This demonstrated the effectiveness of convolutional architectures in learning spatial semantic patterns from textual features.

4.3 Multimodal Meme-Based Hate Detection

The CLIP-based multimodal model [3] was tested using the Facebook Hateful Meme dataset. With contrastive language-image pre-training and prompt engineering:

- **Accuracy = 87.42%**, **F1-score = 86.9%**
- Outperformed traditional unimodal models by a significant margin.

These results emphasize the necessity of multimodal analysis for implicit hate embedded in image-text combinations.

4.4 Threat and Cyber bullying Classification

Using BREE-HD, a transformer-based model [4], threat detection on Twitter showed:

- **Accuracy = 97.0%** across four classes: Sexist Threat, Non-Sexist Threat, Sexist Non-Threat, and Non-Sexist Non-Threat.
- The inclusion of explainable AI (XAI) further strengthened trust in model predictions, especially in high-risk threat categories.

LSTM-based models [10] on Facebook and Instagram datasets yielded:

- Twitter: **Accuracy = 96.64%**
- Instagram: **Accuracy = 94.49%**
- Facebook: **Accuracy = 91.26%**

These results support the use of sequential models for detecting harassment that evolves over time in conversations.

4.5 Fuzzy Logic and Optimization Results

The fuzzy-logic-enhanced classifier [5] demonstrated a 5–7% improvement in recall, particularly in detecting borderline or

https://doi.org/10.5281/zenodo.17452098

ambiguous hate content. Bio-inspired optimization (GA/PSO) helped reduce redundant features, improving model generalizability without additional computational cost.

4.6 Fairness Evaluation

Fairness-aware models guided by demographic constraints [6] significantly reduced bias in predictions. Reduction in false positives for African-American Vernacular English (AAVE) tweets by **18%** and balanced F1-scores across gender and ethnicity without sacrificing overall accuracy. These findings validate the importance of ethical AI design for socially impactful applications.

In the table 1, we have compared the parameters of several models according to their task, accuracy, F1 score, and primary benefit.

4.7 Discussion

The findings show that a comprehensive approach to online hate and threat identification can be achieved by combining multimodal learning with ethical AI concepts. Traditional classifiers performed worse than transformer-based models like BREE-HD and vision-language encoders like CLIP, especially when dealing with context-rich and image-based content.

One important development is the addition of fairness constraints. Numerous models now in use have a tendency to magnify social biases, which can be detrimental when used extensively. Our model shows that ethical restrictions can be implemented without compromising prediction accuracy. ML models are further enhanced by fuzzy logic in their ability to decipher complex language, such as sarcasm or coded hate speech. Additionally, optimization techniques like GA and PSO increased classifier efficiency and decreased over fitting. All things considered, the suggested architecture establishes a solid basis for an AI system for cyber forensic applications that is useful, explicable, and socially conscious.

Table 1: Comparisons of different models

Task	Model	Accuracy	F1-Score	Prominent Benefits
Text Classification	DCNN[1]	91.5%	90.1%	High precision on ternary hate labels
Meme Analysis	CLIP[3]	87.4%	86.9%	Handles implicit multimodal hate
Threat Detection	BREE-HD[4]	97.0%	96.7%	Transformer + Explainable AI approach
Fuzzy ML	Hybrid + Fuzzy [5]	89.2%	88.4%	Better Recall on ambiguous inputs
Fairness –Aware	Constraints [6]	91.3%	90.2%	Bias Reduction across groups
Cyber-bulling	LSTM[10]	96.6%	95.9%	Temporal text detection

5. Conclusion and Future Work

In this study, we suggested a multimodal, AI-powered framework for detecting hate speech and cyber threats on social networking sites. To overcome issues with ambiguity, bias, and multimodality, the system combines text and image processing using deep learning models including DCNN, LSTM, BERT, and CLIP, as well as fuzzy logic reasoning and fairness constraints.

High accuracy was demonstrated by experimental evaluations on a variety of benchmark datasets, with the suggested models greatly surpassing conventional methods. The CLIP-based multimodal classifier shown impressive gains in identifying implicit hate speech hidden in memes, while the transformer-based threat detection model (BREE-HD) attained 97% accuracy. Furthermore, without compromising model performance, fairness-aware training

https://doi.org/10.5281/zenodo.17452098

strategies successfully decreased algorithmic bias across demographic groups. Improved interpretability and recall were achieved by the application of fuzzy logic and bio-inspired optimization strategies, particularly in circumstances that were unclear or borderline. Together, these developments highlight the need for morally sound, precise, and expandable approaches to address online abuse and hate.

Although the suggested framework shows great promise, there are a number of directions where further research may go. Adding detection features for non-English content and regional languages can make the system more inclusive and globally relevant, while integrating it with real-time social media stream monitoring systems can assist stop harmful information from going viral.

In summary, our study adds to the expanding literature of studies on AI-powered cyber forensics and builds the groundwork for reliable, moral, and explicable hate speech detection systems in digital environment.

References

1. Roy, P. K., Tripathy, A. K., Das, T. K., & Gao, X. Z. (2020). A framework for hate speech detection using deep convolutional neural network. IEEE Access, 8, 204951–204962. https://doi.org/10.1109/ACCESS.2020.3037073
2. Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access, 6, 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394
3. Arya, G., Hasan, M. K., Bagwari, A., Safie, N., Islam, S., Ahmed, F. R. A., De, A., Khan, M. A., & Ghazal, T. M. (2024). Multimodal hate speech detection in memes using contrastive language-image pre-training. IEEE Access, 12, 22359–22374. https://doi.org/10.1109/ACCESS.2024.3361322
4. Kumbale, S., Singh, S., Poornalatha, G., & Singh, S. (2023). BREE-HD: A transformer-based model to identify threats on Twitter. IEEE Access, 11, 67180–67192. https://doi.org/10.1109/ACCESS.2023.3291072
5. Ketsbaia, L., Issac, B., Chen, X., & Jacob, S. M. (2023). A multi-stage machine learning and fuzzy approach to cyber-hate detection. IEEE Access, 11, 56046–56064. https://doi.org/10.1109/ACCESS.2023.3282834
6. Gencoglu, O. (2021). Cyberbullying detection with fairness constraints. IEEE Internet Computing, 25(1), 20–29. https://doi.org/10.1109/MIC.2020.3032461
7. Alzamzami, F., & El Saddik, A. (2021). Monitoring cyber SentiHate social behavior during COVID-19 pandemic in North America. IEEE Access, 9, 91184–91203. https://doi.org/10.1109/ACCESS.2021.3088410
8. Ramírez Sánchez, J., Campo-Archbold, A., Zapata Rozo, A., Díaz-López, D., Pastor-Galindo, J., Gómez Mármol, F., & Aponte Díaz, J. (2022). On the power of social networks to analyze threatening trends. IEEE Internet Computing, 26(2), 19–27. https://doi.org/10.1109/MIC.2022.3154712
9. Mullah, N. S., & Wan Zainon, W. M. N. (2021). Advances in machine learning algorithms for hate speech detection in social media: A review. IEEE Access, 9, 88364–88389. https://doi.org/10.1109/ACCESS.2021.3089515
10. Obaida, M. H., Elkaffas, S. M., & Guirguis, S. K. (2024). Deep learning algorithms for cyberbullying detection in social media platforms. IEEE Access, 12, 76901–76915. https://doi.org/10.1109/ACCESS.2024.3406595