Advancements in Hate Speech Detection: From Pre-LLM to LLM with a Novel Multi-Stage Filtering Methodology Using Intent, Aspect, and Dynamic Weighting

Raji S Pillai
Government Engineering College, Palakkad
APJ Abdul Kalam Technological University
St.Teresa's College (Autonomous),Ernakulam
Kerala, India

K R Remesh Babu
Government Engineering College, Palakkad
APJ Abdul Kalam Technological University
Kerala, India

***Abstract*—Social media profoundly influences interactions among individuals, dissemination of information, entertainment, and the development of online- communities, thereby affecting both personal and societal levels. The inherent anonymity and real-time nature of these platforms have facilitated the extensive propagation of hatespeech and biased language. We present a new framework for the detection of hatespeech that goes beyond conventional content-centric methods by focusing on revealing the underlying intent of the author and identifying specific elements within the text. The model differentiates between hatespeech and biasedspeech by interpreting the writer's purpose—such as discrimination, harassment, stereotyping, or misinformation—and by extracting the targeted attributes, including gender, race, disability, politics, or profession. Our multi-stage filtering method incorporates a context-sensitive weighted mechanism that adjusts the significance of intent classification and aspect identification, thereby improving the fine-grained detection of detrimental content. The framework is designed to foster safer online environments by effectively addressing both explicit and nuanced expressions of hatespeech.**

I. Introduction

hatespeech on social media appear in multitude of ways ranging from text in various languages, emojis, memes, or implicit or explicit targeting [1], [2]. It has a chilling effect, silencing marginalized voices and instilling fear and tension. In some cases, it leads to real-world violence, as seen in numerous mob lynchings triggered by online rumors [3]. Social media platforms often act as echo chambers, reinforcing prejudices and deepening societal divides [4].Online hate is widespread, with surveys across various countries indicating that 42%–67% of youth have encountered 'hateful and degrading content or speech online,' and 21% have experienced it as victims [5]. The negative effects of online hate impact both victims and bystanders, contributing to issues like depression, isolation, paranoia, social anxiety, self-doubt, and loneliness (social media and Online Hate). Given the importance of curbing hatespeech and ensuring accountability, extensive research is being conducted from multiple angles, leveraging cutting-edge technologies to combat online hatespeech. [3], [6], [7]

A. Approaches to Hatespeech Detection

Hatespeech can be detected using various approaches, ranging from lexical analysis of individual words to a comprehensive technical evaluation of user behavior and multi modal communication [8]. Each approach provides valuable insights for identifying hatespeech in diverse contexts. By combining techniques from these different approaches, the accuracy and contextual understanding for effective hatespeech detection can be enhanced.

1) Lexical Analysis: Lexical analysis in hatespeech detection involves examining text at the word or token level to identify patterns or linguistic features that signal hatespeech. The presence of general hateful terms strongly suggests hatespeech, and predefined lexicons like those found in Hate Base are often used to detect such language. Many studies have leveraged these hatespeech lexicons, one of the foundational works being the study by Davidson et al [9]. In their research, they used a lexical detection method to differentiate between hatespeech and offensivelanguage. The team collected tweets containing specific hatespeech terms and crowd-sourced their classification into three categories: hatespeech, offensivelanguage, and neither. They then trained a multi-class classifier to classify among these classes. Their results showed that tweets with racist and homophobic content were more likely to be classified as hatespeech, whereas sexisttweets were often categorized as offensivelanguage.

2) Comment Thread Analysis: Detecting hatespeech in online media is difficult with just token-level or word-level analysis because context is very important. hatespeech often does not appear in just one comment but develops in replies or during negative conversations [10]. So, looking at the entire comment thread, rather than individual comments, can be more

effective in identifying hatespeech. In their study, Xinchen Yu, Eduardo Blanco, and Lingzi Hong emphasized the importance of context in identifying hatespeech. In their study the previous message in a conversation thread is identified as context. [10].

3) User-level and Network-level Analysis: Identifying users who spread toxic content and analyzing their behavior patterns can greatly enhance hatespeech detection. Examining how these users interact, the type of content they engage with, and their responses to others can significantly aid in mitigating hatespeech effectively. In their work, Ribeiro et.al analyzed user behavior, rather than focusing solely on content to detect hatespeech. Their study showed that hateful users exhibit weird activity patterns, they use some specific word in their text, and a distinct network structures. [11] .The authors found that hateful users have dense connections, allowing their content to reach a large audience with high spreading velocity.

4) Multimodal Analysis: In the current scenario, multimodal hatespeech that is combining text, images, and videos is very common and thus poses a significant challenge. To effectively address this issue, it's essential to focus on detecting hatespeech across various content types. Identifying and mitigating hatespeech within multimodal content requires advanced approaches that can analyze not just text, but also visuals and audio, ensuring a comprehensive strategy for curbing this growing problem. Kiela et.al In their research, analyzed a multimodal dataset to identify hatespeech, illustrating how the insertion of a meme can alter the meaning of a simple sentence. [1] The complexity level rises as the approaches move from analysing individual words to understanding deeper meanings, social context, and multimodal data. As complexity increases, so does the accuracy, since more features are considered for classifying hatespeech, beyond simple keyword matching. More complex approaches also leverage datasets with richer features, such as context, conversation threads, emojis, and memes, contributing to improved detection performance.

Given the growing importance of addressing hatespeech and ensuring a safer online environment, research in this area has intensified, particularly with the onset of large language models (LLMs). In our review, we analyse various hatespeech detection and classification methods, categorizing them into two phases: pre-LLM approaches which primarily focuses on transformer-based models and those developed in the LLM era. The complexity level rises as the approaches move from analysing individual words to understanding deeper meanings, social context, and multimodal data. As complexity increases, so does the accuracy, since more features are considered for classifying hatespeech, beyond simple keyword matching. More complex approaches also leverage datasets with richer features, such as context, conversation threads, emojis, and memes, contributing to improved detection performance.

II. Evolution of HateSpeech Detection: Pre-LLM and LLM Era Approaches

Given the growing importance of addressing hatespeech and ensuring a safer online environment, research in this area has intensified, particularly with the onset of large language models (LLMs). In our survey, we analyse various hatespeech detection and classification methods, categorizing them into two phases: pre-LLM approaches which primarily focuses on transformer-based models and those developed in the LLM era.

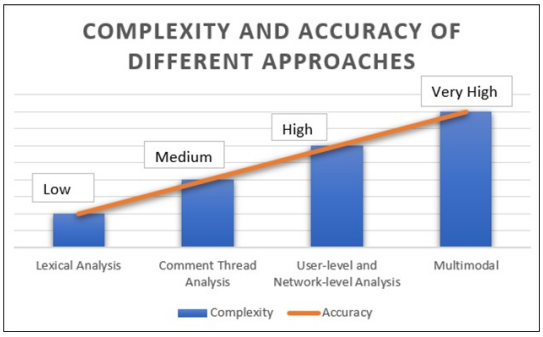

Fig. 1. Complexity and Accuracy of Different hatespeech Detection Approaches

A. Pre-LLM Era (TransformerBased)

With the invent of transformer-based models and attention mechanism the hatespeech detection process has achieved remarkable accuracy increase. They had an intense mechanism to cognize the context and semantic relationships in a document or set of tokens.

In hatespeech detection, it is crucial to not only identify context and semantic relationships but also evaluate the intent of the sentence and the specific aspect or category it targets. In their research, [12] Mazari et al. employed multi-aspect hatespeech detection by classifying text into multiple labels, including 'identity hate,' 'threat,' 'insult,' 'obscene,' 'toxic,' and 'severely toxic.' They utilized word embedding techniques such as FastText and GloVe for encoding, and developed a deep learning model combining Bi-LSTM and Bi-GRU layers to effectively classify hatespeech across these labels.

As online hatespeech frequently appears in a combination of various languages or regional dialects, several transformer-based studies are being conducted to enhance its detection and classification. Additionally, significant research is being pursued on cross-domain datasets. [13] Singhal et al. proposed an ensemble-based approach to automatically detect caste and migration-related hatespeech in Tamil, leveraging multiple versions of pre-trained BERT models. [14] Bilal et al. found that transformer-based models exhibited superior generalization across cross-domain datasets. Similarly, [15] Similarly, Bansal et al. found that XLM-RoBERTa, indic-BERT, MuRIL-BERT, and mBERT are the most effective transformer models for detecting toxic text across 13 Indic languages.

B. LLM Era

Transformer-based models possess the extensive capability to understand context across various regional languages

equipping them with excellent capacity in detecting hatespeech. But in the case of sarcasm, coded language or subtle biases they tend to be less effective. This may be because this type of hatespeech requires a deeper understanding of social context, cultural preferences etc. This is where the importance of large language models becomes evident.

LLMs are advanced transformer-based models where massive amount of text data is utilised for training which composed of billions of parameters. LLMs have evolved through a set of stages starting from statistical language models, neural language models , pre- trained language models and finally LLMs. The first stage which is called statistical language learning the LLMs are used to predict the next word by estimating the likelihood of next word based on the likelihood of previous words.

Neural language models predict the next token by analysing the embedding vectors of preceding tokens using neural networks. They calculate the semantic similarity between these vectors based on their distance, making the process task-specific.

Pre-trained language models, initially trained on large web text corpora, can be fine-tuned for specific tasks using small amounts of labelled data, thereby making them tailored to the task at hand. Tanmay et al. fine-tuned a large language model (LLM) using the LoRA and Adaptor techniques to enhance its overall performance in hatespeech classification. [16]

In India, hatespeech is addressed under various sections of the Bhartiya Nyaya Sanhita, [17]including offenses related to religion, personal insults, defamation against individuals, and defamation targeting Scheduled Castes and Scheduled Tribes. To effectively address hatespeech under the law, it is crucial to not only determine the context but also to assess the intent such as whether it is discriminatory or is it misinformation etc and relevant aspects, such as gender, race, disability, political views, and profession.

III. Proposed Methodology

Hatespeech is legally punishable, whereas biased speech, though not subject to legal consequences,remains undesirable due to its potential psychological impact. Our research aims to address both hate and biased speech by introducing a novel framework that leverages Intent and Aspect in a Multi-Class Hierarchical Approach for classification. The proposed framework utilizes a single model with shared layers to enable multi-task learning, creating an integrated classification pipeline that consists of Intent Classification, Aspect Extraction, and hatespeech Detection.

In this methodology, the model first undergoes training to classify intent. If the predicted intent is negative, aspect extraction follows. Finally, based on the identified intent and extracted aspect, the model classifies the text as hatespeech or biased speech. The loss function weights are dynamically

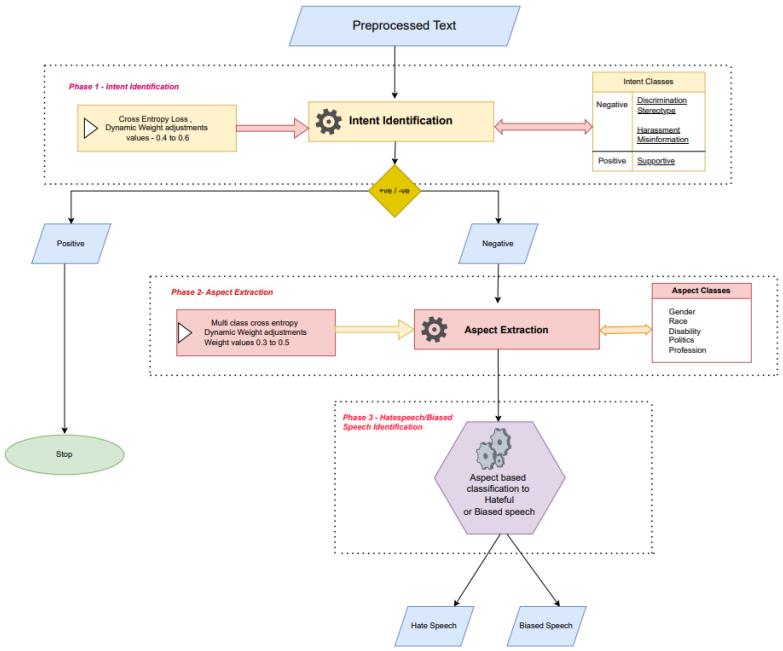

Fig. 2. Block Diagram for the proposed architecture

tuned to balance the contribution of each task to the overall loss, thus optimizing the hierarchical process.

The Intent Classification stage serves as the initial filter, distinguishing negative intents such as discrimination, harassment, misinformation, and stereotyping, while discarding positive intents. Cross-Entropy Loss is applied here to predict discrete intent labels. Subsequently, the Aspect Extraction stage uses Multi-Class Cross-Entropy Loss to assign a single aspect label to each text input. The final hatespeech Detection stage employs Binary Cross-Entropy Loss, enabling binary classification to discern hatespeech from biased speech.

In our experiments, we adjust loss function weights, assigning a range of 0.4 to 0.6 for Intent Classification, 0.3 to 0.5 for Aspect Extraction, and 0.4 to 0.6 for hatespeech Detection, to balance the impact of each task across the multi-stage filtering pipeline.

The weights are set in the mid-range to ensure each task receives balanced importance, avoiding undue prioritization of one task over another—a key factor in multi-stage filtering. This range also allows slight adjustments if a particular task, like intent identification, requires higher emphasis within the pipeline. For instance, intent identification is given a slightly higher weight (up to 0.6) due to its priority in this context, without significantly impacting the overall balance. Additionally, mid-range values support model convergence during training, and they facilitate testing task impact by allowing performance fine-tuning in small, manageable steps.

TABLE I
SUMMARY OF RELATED WORKS

Related Works	Methodology	Features
[18]	Combining textual understanding capabilities of LLMs and discriminative power of advanced classifiers.	Ensuring faithful interpretability, integrating the strengths of LLMs with advanced classifiers, and enhancing overall detection performance.
[19]	Analysing role of LLMs as classifiers	Directionality and Hate target are being analysed.. GPT-3.5 and Llama 2 exhibits commendable performance.
[12]	The deep learning model combines Bi-LSTM and Bi-GRU layers, leveraging GloVe and FastText embeddings, with additional performance enhancement through BERT integration.	Multi-aspect hate-speech detection based on classifying text into multi-labels.
[13]	A majority voting approach was used, combining XMLR, mBERT, and MuRIL models.	Ensemble Model with Majority Voting to detect Tamil hatespeech.
[20]	Pre-trained BERT embeddings with LSTM, BiLSTM, BiLSTM with an attention layer, and CNN.	Roman – Urdu multilingual HS classification and generalized the model using cross-domain dataset.
[6]	The classification efficacy and model complexity of four distinct Deep Neural Network models were evaluated: CNN (baseline), bidirectional LSTM with attention, pre-trained BERT, and fine-tuned RoBERTa transformer models.	Ternary fine-tuned RoBERTa transformer leveraging the Adam optimizer.
[21]	Hyperparameter tuning was performed on six different transformer models: AraBERT, AraElectra, ALBERT-Arabic, AraGPT-2, mBERT, and XLM-RoBERTa.	Detecting Offensivelanguage and identifying HS in Arabic dataset by means of majority vote and highest sum.
[22]	Experimental evaluation of XLM-RoBERTa, indic-BERT, MuRIL-BERT, and mBERT	Trained on a combined dataset of 13 Indic languages. Incorporating emoji embeddings enhance the performance of XLM-RoBERTa.
[23]	A pre-trained multilingual Transformer-based text encoder XLM-RoBERTa was used to classify tweets as hatespeech, offensive, or profane.	Demonstrates The usefulness and efficacy of language models trained with multilingual objectives across various languages.
[24]	Fine-tuned and modified multilingual Transformer models (mBERT, XLM-RoBERTa).	DBinary classification using a cross lingual approach.

IV. CONCLUSION AND FUTURE WORK

As digital platforms play a crucial role in individuals lives across all ages, ensuring their safety and entertainment is essential. As social networking platforms provide a space for everyone to express their thoughts and ideas freely, the presence of hatespeech and biased speech is widespread, which can negatively impact users and society, emphasizing the need for more effective detection techniques. In our work, we review the evolution of hatespeech detection methods, comparing transformer approaches in the pre-LLM era to the advanced capabilities of contemporary large language models (LLMs). While LLMs have excelled in detecting explicit hatespeech, we propose methodology for detecting implicit hatespeech by identifying the intent and aspect and finally classifying the text based on intent and aspect included in the text. In future work we are planning to implement this methodology allowing adjustments in weights to facilitate model convergence and fine tuning for optimal pipeline performance.

REFERENCES

[1] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: detecting hate speech in multimodal memes," in *Proceedings of the 34th International*

Conference on Neural Information Processing Systems, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[2] G. Arya, M. K. Hasan, A. Bagwari, N. Safie, S. Islam, F. R. A. Ahmed, A. De, M. A. Khan, and T. M. Ghazal, "Multimodal hate speech detection in memes using contrastive language-image pre-training," *IEEE Access*, vol. 12, pp. 22 359–22 375, 2024.

[3] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," *IEEE Access*, vol. 6, pp. 13 825–13 835, 2018.

[4] L. Nath, P. Mishra, R. Singh, S. Jain, A. Singh, and S. M. Benedict, "'online hate speech in inda: Legal reforms and social impact on social media platforms'," *SSRN Electronic Journal*, 01 2024.

[5] J. B. Walther, "Social media and online hate," *Current Opinion in Psychology*, vol. 45, p. 101298, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352250X21002505

[6] U. Mittal, "Detecting hate speech utilizing deep convolutional network and transformer models," in *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, 2023, pp. 1–4.

[7] V. Bansal, M. Tyagi, R. Sharma, V. Gupta, and Q. Xin, "A transformer based approach for abuse detection in code mixed indic languages." *ACM Transactions on Asian and Low-Resource Language Information Processing*, 11 2022.

[8] P. P. Jemima, B. R. Majumder, B. K. Ghosh, and F. Hoda, "Hate speech detection using machine learning," in *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, 2022, pp. 1274–1277.

[9] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, 03 2017.

[10] X. Yu, E. Blanco, and L. Hong, "Hate speech and counter speech detection: Conversational context does matter," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5918–5930. [Online]. Available: https://aclanthology.org/2022.naacl-main.433

[11] M. Horta Ribeiro, P. Calais, Y. dos Santos, V. Almeida, and W. Meira Jr, "Characterizing and detecting hateful users on twitter," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, 06 2018.

[12] A. C. Mazari, N. Boudoukhani, and A. Djeffal, "Bert-based ensemble learning for multi-aspect hate speech detection," *Cluster Computing*, vol. 27, no. 1, p. 325–339, Jan. 2023. [Online]. Available: https://doi.org/10.1007/s10586-022-03956-x

[13] K. Singhal and J. Bedi, "Transformers@LT-EDI-EACL2024: Caste and migration hate speech detection in Tamil using ensembling on transformers," in *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, B. R. Chakravarthi, B. B, P. Buitelaar, T. Durairaj, G. Kovács, and M. Á. García Cumbreras, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 249–253. [Online]. Available: https://aclanthology.org/2024.ltedi-1.32

[14] M. Asghar, A. Khan, S. Jan, S. Musa, and S. Ali, "Roman urdu hate speech detection using transformer-based model for cyber security applications," *Sensors*, vol. 23, p. 3909, 04 2023.

[15] V. Bansal, M. Tyagi, R. Sharma, V. Gupta, and Q. Xin, "A transformer based approach for abuse detection in code mixed indic languages." *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Nov. 2022, just Accepted. [Online]. Available: https://doi.org/10.1145/3571818

[16] T. Sen, A. Das, and M. Sen, "Hatetinyllm : Hate speech detection using tiny large language models," 2024. [Online]. Available: https://arxiv.org/abs/2405.01577

[17] BharatiyaNyayaSanhita, "Bharatiya nyaya sanhita," Government of India, 2023.

[18] A. Nirmal, A. Bhattacharjee, P. Sheth, and H. Liu, "Towards interpretable hate speech detection using large language model-extracted rationales," in *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, Y.-L. Chung, Z. Talat, D. Nozza, F. M. Plaza-del Arco, P. Röttger, A. Mostafazadeh Davani, and A. Calabrese, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 223–233. [Online]. Available: https://aclanthology.org/2024.woah-1.17

[19] T. Kumarage, A. Bhattacharjee, and J. Garland, "Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection," *ArXiv*, vol. abs/2403.08035, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268379355

[20] M. Asghar, A. Khan, S. Jan, S. Musa, and S. Ali, "Roman urdu hate speech detection using transformer-based model for cyber security applications," *Sensors*, vol. 23, p. 3909, 04 2023.

[21] A. F. Magnossão de Paula, P. Rosso, I. Bensalem, and W. Zaghouani, "UPV at the Arabic hate speech 2022 shared task: Offensive language and hate speech detection using transformers and ensemble models," in *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, H. Al-Khalifa, T. Elsayed, H. Mubarak, A. Al-Thubaity, W. Magdy, and K. Darwish, Eds. Marseille, France: European Language Resources Association, Jun. 2022, pp. 181–185. [Online]. Available: https://aclanthology.org/2022.osact-1.23

[22] V. Bansal, M. Tyagi, R. Sharma, V. Gupta, and Q. Xin, "A transformer based approach for abuse detection in code mixed indic languages." *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, Nov. 2022, just Accepted. [Online]. Available: https://doi.org/10.1145/3571818

[23] S. G. Roy, U. Narayan, T. Raha, Z. Abid, and V. Varma, "Leveraging multilingual transformers for hate speech detection," *ArXiv*, vol. abs/2101.03207, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231573370

[24] T. Tita and A. Zubiaga, "Cross-lingual hate speech detection using transformer models," 11 2021.