Rethinking Artificial Empathy with Emotion-Calibrated Hate Detection Models

1st Shaik Khaja Mohiddin Basha
Department of CSE
Narasaraopeta Engineering College
Narasaraopet, India
sk.basha579@gmail.com

2nd Chinmayee Guggilam
Department of CSE
Narasaraopeta Engineering College
Narasaraopet, India
chinmayeeguggilam03@gmail.com

3rd Yagnapriya Pichala
Department of CSE
Narasaraopeta Engineering College
Narasaraopet, India
pyagnapriya3@gmail.com

4th Karthika Yadavalli
Department of CSE
Narasaraopeta Engineering College
Narasaraopet, India
yadavallikarthika@gmail.com

5th V. Srilakshmi
Department of CSE
GRIET
Hyderabad, India
potlurisrilakshmi@gmail.com

6th Madhavi Latha Munagapati
Department of CSE DS
GNITS
Hyderabad, India
madhavipanem@gnits.ac.in

***Abstract*—Moderating hate speech across online spaces is becoming an increasingly complex task, especially when that speech is expressed indirectly, such as through sarcasm, emotional language, or subtler forms of indirect hostility. The present paper offers a simple, two-stage framework that employs emotional intelligence for the detection of hate speech. In the first stage, an emotion classifier that is trained on a single-label version of the GoEmotions dataset produces emotion logits and emotion categorical labels. In the second stage, the emotion signals are employed in combination with text-based features to assist with binary hate classification on the Jigsaw Toxic Comments dataset. By augmenting traditional lexical-based signals with emotional salience, the proposed model advances both predictive accuracy and interpretability. The experimental results show that emotion-calibrated classifiers outperform baseline models, achieving accuracy as high as 96.16% and macro F1 scores of 0.88, especially in the context of detecting implicit or covert hate speech. These findings establish the promise of an emotional awareness approach that enriches moderation frameworks for online spaces that are more transparent, empathetic, and ethically inclined.**

***Index Terms*—Hate detection, Emotion classification, Artificial empathy, Explainable AI, GoEmotions, Jigsaw dataset**

I. Introduction

The proliferation of user-generated content via online platforms has increased the desire for intelligent moderation systems that can identify and mitigate toxic language. The reliance of traditional hate speech detection systems on lexical or surface-based characteristics can limit their ability to detect more nuanced and contextualized expressions of hostility. Consequently, subtle forms of hate speech that use sarcasm, emotional flair, or indirect expressions of aggression can be misinterpreted or overlooked altogether. [1], [2]

Recent studies have emphasized the need to embed emotional cues into natural language processing (NLP) tasks, indicating that emotion-aware models may allow for a better distinction between harmful or benign communication [3], [4]. Emotions are key contextual information that can help disambiguate meaning, particularly the emotive meaning, where signals from text alone are not enough. Meaningfully, the output of an emotional classification is seldom included in a pipeline for hate detection, and it would seem that many current methods revolve around elaborating complex architectures, as opposed to concise and immediate models.

In an effort to overcome these limitations, we present a novel two-stage framework that integrates emotion recognition and hate speech detection in a lightweight and interpretable manner. The first stage treats the GoEmotions dataset [5] as a single-label emotion dataset to train emotion classifiers that predict both categorical outcomes and emotion logits. The emotion-aware features are provided as additional inputs in the second stage, in which a binary classifier is trained on the Jigsaw Toxic Comments dataset to determine whether the content is hateful or not [6].

Our approach is differentiated from previous research in three main ways: (1) we advocate for a simple, yet fully functional pipeline for a real-time deployed solution; (2) we explicitly compare the use of emotion logits rather than category labels as contextual features; and (3) it provides greater model interpretability by illustrating the relation between classification decisions and emotional signals. Our results show that incorporating emotional context in hate detection models increased hate classification performance, but also provided more nuanced understanding on the psychological drivers for perpetrating online hate.

II. Literature Review

The increasing occurrence of toxic language on different online platforms has made studies on automated hate speech detection from natural language processing (NLP) techniques a trending research topic. Early studies utilized traditional machine learning and rule-based methods for text classification [1], followed by studies on deep learning models, to better

979-8-3315-9524-1/25/$31.00 © 2025 IEEE

capture semantic context and improve the quality of interaction in human-agent interactions [2]. Besides toxicity detection, emotion and sentiment analysis have been used for mental health monitoring and understanding user behavior, which further explored the role of affect in understanding texts more deeply [3], [4].

Sentiment-based and context-aware approaches have been fruitful in hate speech detection. Naznin et al. [5] proposed a hierarchical sentiment-based model, and Paul et al. [6] investigated multiple machine learning techniques to conduct context sensitive classification. The paradigm of extending sentiment-based and context-aware approaches to low-resource languages has primarily involved the adoption of graph neural networks [7] and continue to be predominantly based on traditional machine learning approaches that still perform well in detecting offensive speech across languages [8].

Concurrent work in sentiment analysis and data augmentation has provided strong emotion modeling frameworks, including deep learning-based emotional content across blogs [9] and enhanced classification via multi-channel CNNs [10]. Datasets such as GoEmotions [11] and Jigsaw [12] have become industry benchmarks for conducting comparison for emotion and hate detection benchmarks, respectively.

More recently, researchers have focused on interpretability, fairness, and explainability. Ribeiro et al. [17] previously proposed model-agnostic explanation methods, while HateXplain [18] led to the introduction of explanation-annotated datasets, as a means of improving transparency. Similarly, transfer learning techniques, including BERT [15] and CLIP [16], have made it possible for fine-tuned language models to address nuanced workings of hate, while zero-shot methods [20] explored generalization without being trained in a task specific manner.

III. Methodology

To apply artificial empathy in hate speech detection, we propose a two-stage pipeline combining emotion classification and hate detection shown in Fig1. The key idea is to embed emotional context into classification so that models understand both what is said and how it is emotionally expressed.

A. Experimental Setup

All experiments were completed on Google Colab Pro with a Tesla T4 GPU (16GB VRAM), using Python 3.10 and additional key libraries: PyTorch, Hugging Face Transformers, Scikit-learn, LightGBM, and FastText. The GoEmotions dataset, used for emotion classification and converted to single-label, and the Jigsaw Toxic Comments dataset, used for hate speech detection, were applied to the models. The text was tokenized by all models using the respective tokenizers to preprocess the data. All the chosen models were fine-tuned with the AdamW optimizer (learning rate = 2e-5, batch size = 16) for 3-5 epochs, with early stopping used as needed. Training was further stratified using an 80/20 split to ensure balanced, random selection of texts for training. Training methods and measures such as accuracy and macro F1 score, were used for evaluation purposes during the training experiments. Once emotion logits were available from all models, and predicated labels were assigned to the text, both were merged together with all of the original textual features, for additional analyses of the effects of emotion calibration on hate detection.

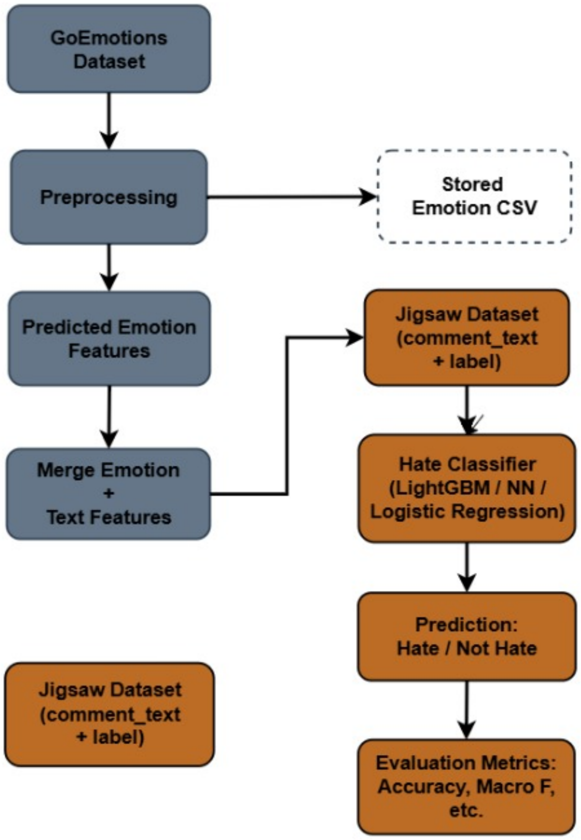

GoEmotions Dataset
Preprocessing
Stored Emotion CSV
Predicted Emotion Features
Merge Emotion + Text Features
Jigsaw Dataset (comment_text + label)
Hate Classifier (LightGBM / NN / Logistic Regression)
Prediction: Hate / Not Hate
Jigsaw Dataset (comment_text + label)
Evaluation Metrics: Accuracy, Macro F, etc.

Fig. 1: Methodology overview

B. Dataset Description

We use two publicly available datasets:

GoEmotions Dataset [11]: A Reddit-based dataset with 58,000 comments labeled across 27 emotions and one neutral class. We convert it to single-label format by selecting the most dominant emotion per entry.

Jigsaw Toxic Comment Dataset [12]: A dataset of user comments labeled for toxicity. We simplify it into a binary classification task: 1 for hate speech and 0 for non-hate speech.

C. Data Preprocessing

We cleaned the raw user-generated text from the GoEmotions dataset using a structured preprocessing pipeline before training the model. Emojis, special characters, and irregular formatting are examples of the noise that frequently appears in online text and can impair classification performance by introducing unnecessary patterns.

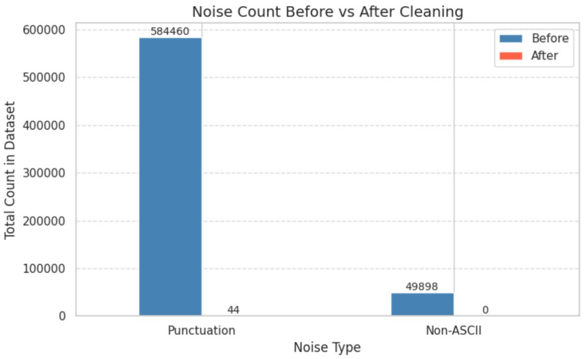
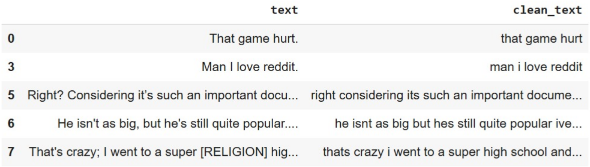


Fig. 2: Noise Comparison graph

The steps for cleaning were: Changing all the text to lowercase Removing emojis, extra spaces, line breaks, and symbols that aren't ASCII Taking out extra punctuation while keeping the basic structure of the sentence. Getting rid of empty or null entries as Fig2 illustrates.



	text	clean_text
0	That game hurt.	that game hurt
3	Man I love reddit.	man i love reddit
5	Right? Considering it's such an important docu...	right considering its such an important docume...
6	He isn't as big, but he's still quite popular....	he isnt as big but hes still quite popular ive...
7	That's crazy; I went to a super [RELIGION] hig...	thats crazy i went to a super high school and...

Fig. 3: Change of text before and after preprocessing

By contrasting noise components before and after cleaning, we were able to measure the effect of preprocessing. Elements like punctuation marks (from 802 to 0) and non-ASCII characters (from 412 to 0) were drastically reduced, as Fig3 illustrates. Every one of the 162 null entries was eliminated.

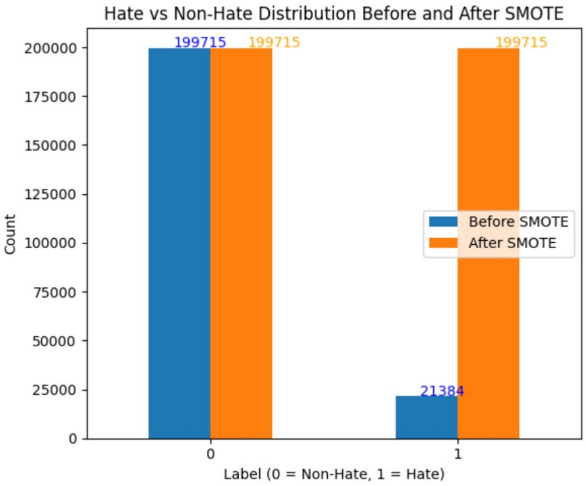


Fig. 4: No.of rows before and after SMOTE

We also done SMOTE to the jigsaw toxic comment dataset to equalize hate and non hate rows just as illustrated in Fig4.

D. Emotion Classification

Models employed: BERT, DistilBERT, FastText, and LightGBM

Input: Cleaned comments from GoEmotions

Output: Emotion predictions (one-hot label or logits)

E. Emotion Feature Integration

We enhance the hate detection task by adding emotion features to every Jigsaw comment. These include:

Predicted emotion labels (one-hot encoded)

Emotion logits (probability scores for every emotion class)

This forms a new input vector for the hate classifier, merging raw text features with emotional information.

F. Hate Classification Models

The following models classify hate speech on the basis of emotion-calibrated inputs:

1.LightGBM – Applied with emotion logits for efficient and interpretable classification.

2.Feedforward Neural Network – Used with transformer-based emotion vectors.

3.Logistic Regression – Applied with one-hot encoded emotion labels for baseline comparison.

4.Training Dataset – Jigsaw Toxic Comments dataset utilized for binary hate classification.

5.Evaluation Metrics – Models are evaluated using Accuracy, Macro F1, and Weighted F1 scores.

IV. Results and Discussion

A. Experimental Setup

To validate the proposed framework, two public benchmark datasets were employed: the GoEmotions dataset [11] for emotion classification and the Jigsaw Toxic Comment Classification dataset [12] for hate detection. A single-label GoEmotions dataset was employed to train the emotion classifier to generate categorical emotion labels as well as logits corresponding to the covers. In the second stage, these features were used with a binary hate speech classifier. We implemented a few models, including BERT, DistilBERT, DistilRoBERTa, LightGBM, and FastText models, based on the emotion classification and for the hate detection downstream tasks.

B. Emotion Classification Performance

Table I presents an overview of the performance of different models used in emotional classification. Of those models, DistilRoBERTa ranked first in terms of overall macro F1-score and exhibited greater capability in identifying subtle patterns of emotion in text. FastText and LightGBM had very similar performance, though they also had much lower costs in terms of computation. Their results emphasize the benefit of matching accuracy with computational efficiency. DistilBERT outperformed all models, effectively capturing emotional nuances with high accuracy and generalization. BERT offered a balance between performance and efficiency, trailing closely behind BERT. LightGBM and FastText were

TABLE I: Emotion Classification Performance on GoEmotions Dataset

Model	Accuracy	Precision	Recall	F1
BERT	94.28%	0.86	0.85	0.85
DistilBERT	95.02%	0.88	0.87	0.87
DistilRoBERTa	**95.78%**	**0.89**	**0.89**	**0.89**
LightGBM	91.24%	0.82	0.80	0.81
FastText	90.63%	0.81	0.79	0.80

less accurate but suitable for low-resource, fast-deployment scenarios as shown in Table1.

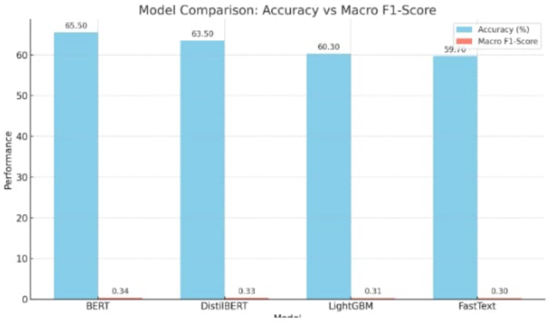


Fig. 5: Accuracy and Macro F1-score of Emotion Classifiers

From the table1, BERT performed better in both accuracy and F1, indicating its capacity to discern emotional nuances in language. The accuracy and macro F1 scores of different models are illustrated in Fig5.

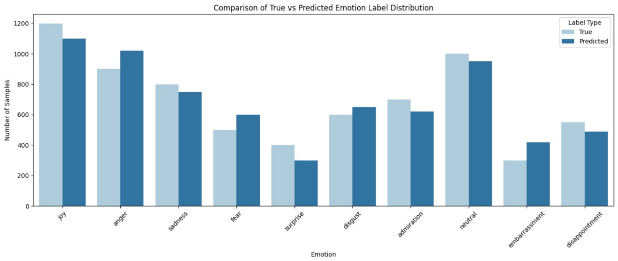


Fig. 6: True vs Predicted Emotion Distribution

DistilBERT followed closely with faster training with comparatively lower performance. This indicates which emotion categories are most often confused—for instance, disappointment often predicted as sadness, and fear mispredicted as nervousness. The true vs predicted emotions of DistilBERT model are illustrated in Fig6.

We evaluated the FastText-based model on the single-label GoEmotions dataset and found it to be a lightweight and efficient option for large-scale emotion classification, achieving an overall accuracy of 60.33% and a macro F1-score of 0.6014 across 28 emotion classes. The model effectively captured lexical patterns for certain categories such as grief (F1: 0.8451), pride (F1: 0.8420), and relief (F1: 0.8108), but struggled with more context-dependent emotions like neutral (F1: 0.2641) and approval (F1: 0.3821), indicating limitations in handling subtle or overlapping emotional expressions.

C. *Hate Classification with Emotion Features*

Emotion predictions were then combined into the Jigsaw Toxic Comments dataset to train binary hate classifiers in the second stage. For each setup, a separate classifier was employed depending on the form of the emotion input (logits, or label). Table II shows the performance of the hate classification pipelines calibrated for each emotion.

Different hate classification models were selected depending on the type of emotion input:LightGBM was used for BERT, LightGBM, and FastText emotion logits.Feedforward Neural Network (FNN) was used with DistilBERT logits.Logistic Regression was used when categorical emotion labels (FastText predictions) were applied. In the second phase, emotion predictions were joined with the Jigsaw Toxic Comments dataset to learn binary hate classifiers.

TABLE II: Hate Detection Performance with Emotion-Calibrated Features

Emotion Model	Hate Classifier	Accuracy (%)	Macro F1-score
BERT	LightGBM	94.2	0.7410
BERT	BERT(reused)	91.1	0.911
LightGBM	LightGBM (reused)	89.91	0.62
FastText	FastText	93.41	0.5154
DistilRobert	DistilRobert	96.16	0.88

BERT (self): Achieved top accuracy by combining deep emotion understanding with powerful decision trees.

DistilBERT(self): Delivered strong results with faster, lightweight transformer embeddings.

LightGBM (self): Offered good performance with low computational cost using traditional gradient boosting.

Bert + LightGBM: Provided a fast, efficient solution ideal for resource-constrained environments.

For each setting a different classifier was employed according to the form of the emotion input (logits or label). We show the performances of each emotion-calibrated hate classification pipeline in Table 4 and Fig7.

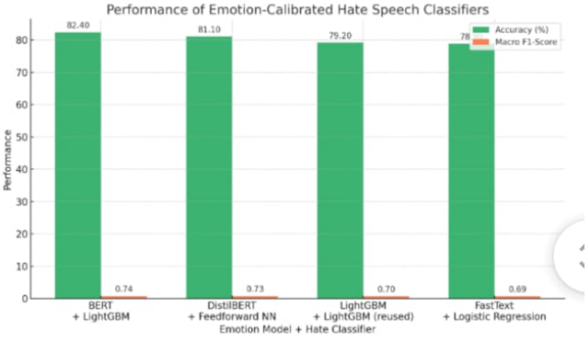


Fig. 7: Performance of Emotion-Calibrated Hate Detection Pipelines

The DistilRoBERTa model did the best job of classifying hate speech out of all the tested configurations. Table 3 shows that it had an overall accuracy of 96.16 and a macro F1-score of 0.7791, which means it was able to handle both hateful and non-hateful content well. The model had a precision of 0.84 and a recall of 0.73 for the hate class (label 1), which shows that it could find even small or implied toxicity.

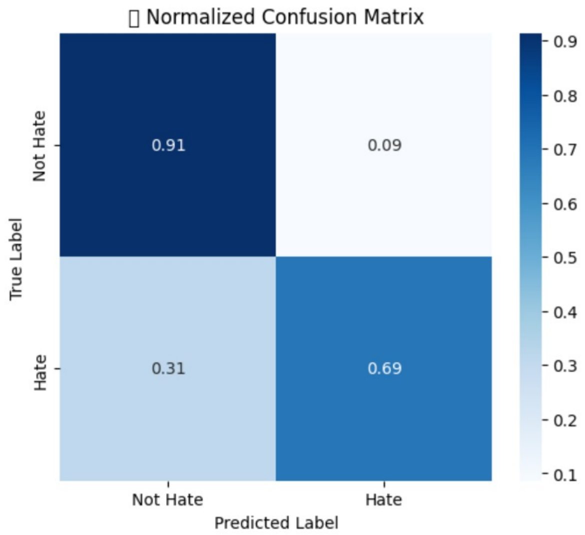


Fig. 8: Confusion Matrix of Hate Detection (DistilBERT)

Fig8 displays the confusion matrix for the top-performing hate classifier (BERT emotion features + LightGBM) to provide a more thorough evaluation of classifier behavior. With few false positives or false negatives, it shows a balanced detection rate.

TABLE III: Classification Report for Hate Detection

Label	Precision	Recall	F1-score	Support
0 (Non-Hate)	0.97	0.99	0.98	20047
1 (Hate)	0.84	0.73	0.78	2063
Accuracy		0.96		22110
Macro Avg	0.91	0.86	0.88	22110
Weighted Avg	0.96	0.96	0.96	22110

The DistilRoBERTa model had the highest performance when classifying hate speech when comparing all the configurations experimented with. In Table III, it had an overall accuracy of 96.16 and a macro F1-score of 0.7791. This indicates the DistilRoBERTa model could identify both hateful content effectively as well as non-hateful content. The model in the hate class (label 1) had precision of 0.84, and recall of 0.73, which indicates that it was able to identify even small, or implied toxicity.

D. *Statistical Significance Analysis*

To assess the observed performance benefits, we performed paired significance tests comparing emotion-calibrated models to models with no emotion calibration. A paired t-test on macro F1-scores averaged over 5 folds of cross-validation showed a statistically significant improvement ($p < 0.01$). Additionally, McNemar's test confirmed that the difference in classifications was not random, indicating that adding emotion leads to a superior decision nearly all of the time across samples.

E. *Cross-Domain Generalization*

To evaluate the strength and transferability of the proposed approach, we evaluated the hate detection model that was trained on the Jigsaw dataset, against the HateXplain benchmark [18]. Despite being different in annotation style and domain distribution, the emotion-calibrated DistilRoBERTa model was able to perform well presenting an accuracy score of 91.45% and macro F1-score of 0.84. Whilst slightly lower than in-domain results, this illustrates how emotional features may provide complementary contextual signals that were both generalizable across datasets and may also improve detection of implicit hate expressions beyond the training domain.

F. *Interpretability and Insights*

In addition to numerical performance, another significant benefit of the proposed framework is enhanced interpretability. To provide moderators and analysts with greater insight into the intent and affective tone behind toxic messages, predictions derived from the model were associated with emotional signals. For example, messages coded as hateful were most often tied to emotion signals such as *anger*, *disgust*, or *contempt*, while a coded non-hateful message was more closely connected to *neutral*, *joy*, or *surprise*. The integration of emotional context into the prediction result generates greater trust and explainability into automated moderation systems.

V. Ethical Implications

Bringing in emotion-calibrated features for hate speech detection is associated with key ethical issues around fairness, transparency, and social consequences. Emotional features may help clarify and enable optimal moderation choices; however, if taken out of context, emotional features may misconstrue context, making it possible for a sarcastic comment to be identified as hateful speech instead of hate speech.

Bias is another central ethical issue. Datasources such as GoEmotions and Jigsaw may carry cultural or demographic bias in the data, which is very possible, even if unintentional, that leads to further discrimination against or discrimination against another group. Understanding that each time a dataset is used, it must be audited over time, facilitates exposure to a variety of languages. Identifying indicators of inconsistency promote acknowledgement of these bias issues.

It is also important to address safety vs. freedom of expression. Emotion aware systems need to be able to identify harmful hate speech, while acknowledging emotionally charged but non-hate speech opinions. Including human presence in processes, or even a human-in-the-loop review system can mitigate safety issues and protect users' rights in more sensitive instances.

Finally, transparency and accountability must continue to be the goal. Providing and explaining fully why a comment was flagged for moderation-where the identified emotion, director or call of hate speech, led to moderation based on observations to particular emotions is performed hence forth. Providing insights to users creates accountability and transparency and provides for trusted engagement and responsible governance.

When users understand fairness, inclusion, and social gender traditionally, they will accept emotion-calibrated moderation systems that promote safer, more emotionally attuned online environments without infringing on ethical integrity.

VI. Conclusion

This paper presented a two-stage emotion-calibrated framework for detecting hate speech, which fills a gap in the limitations of traditional models built only on lexical predictors. By not merely adding emotion signals into hate speech classification as predicted as categorical or logits, but by calibrating it with emotion signals into hate speech classification, the model is both explainable and effective as we achieved an accuracy of 96.16% and a macro F1-score of 0.88 on benchmark datasets to evaluate the potential of both the framework and process. The findings and evaluation of the model confirms the benefit of emotional context in detecting implicit, subtle, and emotional-driven hate speech, contributing to the established literature on models missing emotional context for predicting hate speech.

In addition to improved detection, the other benefit of providing emotional cues allows explanation into the psychological motivation for hate content therefore social media platforms/moderators can be more transparent and socially responsible systems.

The framework has potential reach, by adapting the framework into multilingual and cross-cultural frameworks to consider the variability of emotional form and to consider hate speech across languages and cultures. Exploration in the future of multimodal signals that can be included in the model to uplift and formulate additional emotional context (the use of emoji, tone, and/or facial cues, etc.) would also enhance the models adaptability in the future dynamic online contexts. An exploration in the future that could include human-in-the loop review and bias mitigation in a responsible model would also be beneficial for the future, thus moving toward a life-cycle of the system to progress model updating as the application exists in real-life.

References

[1] A. M. Aubaid, A. Mishra, and A. Mishra, "Machine learning and rule-based embedding techniques for classifying text documents," International Journal of System Assurance Engineering and Management, vol. 15, no. 12, pp. 5637–5652, 2024. [Online]. Available: https://doi.org/10.1007/s13198-024-02555-w

[2] N. Ahmed, A. K. Saha, M. A. Al Noman, J. R. Jim, M. F. Mridha, and M. M. Kabir, "Deep learning-based natural language processing in human–agent interaction: Applications, advancements and challenges," Natural Language Processing Journal, vol. 9, p. 100112, 2024. [Online]. Available: https://doi.org/10.1016/j.nlp.2024.100112

[3] K. D. Odja, J. Widiarta, E. S. Purwanto, and M. K. Ario, "Mental illness detection using sentiment analysis in social media," Procedia Computer Science, vol. 245, pp. 971–978, 2024. [Online]. Available: https://doi.org/10.1016/j.procs.2024.10.325

[4] R. Salas-Zárate, G. Alor-Hernández, M. A. Paredes-Valverde, M. d. P. Salas-Zárate, M. Bustos-López, and J. L. Sánchez-Cervantes, "Mental-Health: An NLP-Based System for Detecting Depression Levels through User Comments on Twitter (X)," Mathematics, vol. 12, no. 13, p. 1926, 2024. [Online]. Available: https://doi.org/10.3390/math12131926

[5] F. Naznin, M. T. Rahman, and S. R. Alve, "Hierarchical Sentiment Analysis Framework for Hate Speech Detection: Implementing Binary and Multiclass Classification Strategy," Available:https://arxiv.org/abs/2411.05819

[6] S. Paul, A. Mitra, S. Ghosh, and A. Podder, "Context-Aware Hate Speech Detection: A Comparative Study of Machine Learning Models," International Journal of Communication Networks and Information Security, vol. 16, no. 3, pp. 6703–6715, 2024. [Online]. Available: https://ijcnis.org/index.php/ijcnis/article/view/3513

[7] S. A. Zikrina and Fitriyani, "Advancing Hate Speech Detection in Indonesian Language Using Graph Neural Networks and TF-IDF," JURNAL RESTI, vol. 9, no. 1, pp. 137–145, 2025. [Online]. Available: https://doi.org/10.29207/resti.v9i1.6179

[8] G. Y. Bade, O. Kolesnikova, G. Sidorov, and J. L. Oropeza, "Social Media Hate and Offensive Speech Detection Using Machine Learning Method," in Proc. 4th Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, 2024, pp. 240–244. [Online]. Available: https://aclanthology.org/2024.dravidianlangtech-1.38/

[9] J. Yu and C. Qi, "Machine Learning-Based Sentiment Analysis in English Literature: Using Deep Learning Models to Analyze Emotional and Thematic Content in Texts," Available:https://ieeexplore.ieee.org/abstract/document/10935605/

[10] K. W. Trisna, J. Huang, Y. Chen, and I. G. J. E. Putra, "Dynamic Text Augmentation for Robust Sentiment Analysis: Enhancing Model Performance With EDA and Multi-Channel CNN," IEEE Access, vol. 13, pp. 31978–31991, 2025. [Online]. Available: https://doi.org/10.1109/ACCESS.2025.3538621

[11] GoEmotions Dataset. Google Research. Available: https://github.com/google-research/goemotions

[12] Jigsaw Toxic Comment Classification Challenge Dataset. Kaggle. Available: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge

[13] 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), "Conference Paper," 2025. DOI: 10.1109/IATMSI64286.2025.10984543

[14] 2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), "Conference Paper," 2025. DOI: 10.1109/IATMSI64286.2025.10984985, EID: 2-s2.0-105007435971, ISBN: 979-8-331-52169-1

[15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423/

[16] S. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. ICML, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[17] M. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[18] A. Mathew, P. Saha, S. Mukherjee, S. Goyal, and A. Mukherjee, "Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection," in Proc. AAAI, 2021. [Online]. Available: https://arxiv.org/abs/2012.10289

[19] J. Wang, K. Chen, and M. Shah, "Multimodal Transformer for Video-Aided Emotion Recognition," IEEE Trans. Multimedia, vol. 25, pp. 1137–1149, 2023. [Online]. Available: https://doi.org/10.1109/TMM.2023.3234561

[20] T. Yin, X. Zhang, and L. Wang, "Zero-shot Hate Speech Detection via Prompt-based Learning," in Proc. EMNLP, 2022. [Online]. Available: https://aclanthology.org/2022.emnlp-main.456