SemanticCuetSync@DravidianLangTech 2025: Multimodal Fusion for Hate Speech Detection - A Transformer Based Approach with Cross-Modal Attention

Md. Sajjad Hossain, Symom Hossain Shohan, Ashraful Islam Paran , Jawad Hossain and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology
{u1904031, u1904048, u1904029, u1704039}@student.cuet.ac.bd
moshiul_240@cuet.ac.bd

Abstract

The rise of social media has significantly facilitated the rapid spread of hate speech. Detecting hate speech for content moderation is challenging, especially in low-resource languages (LRLs) like Telugu. Although some progress has been noticed in hate speech detection in Telegu concerning unimodal (text or image) in recent years, there is a lack of research on hate speech detection based on multimodal content detection (specifically using audio and text). In this regard, DravidianLangTech has arranged a shared task to address this challenge. This work explores three machine learning (ML), three deep learning (DL), and seven transformer-based models that integrate text and audio modalities using cross-modal attention for hate speech detection. The evaluation results demonstrate that mBERT achieved the highest F-1 score of 49.68% using text. However, the proposed multimodal attention-based approach with Whisper-small+TeluguBERT-3 achieves an F-1 score of 43 68%, which helps us achieve a rank of 3rd in the shared task competition.

1 Introduction

Social media platforms have emerged as the focal point for information sharing in the rapidly evolving digital world, where individuals interact and communicate. On the one hand, increased connectivity and easier idea sharing have resulted from increased online activities, and it has also accelerated the spread of hate speech and other forms of internet harassment. Hate speech refers to communication, including speaking, writing, and symbolic expressions, that spreads hatred, slander, discrimination, and violence. It may be aimed at a specific person or group based on traits including race, color, ethnicity, religion, gender, sexual orientation, caste, country, or socioeconomic class (Nockleby, 1994; Keipi et al., 2016; Benikova et al., 2018). Due to the large volume of data, manually monitoring and identifying hate speech is impractical. That is why manual moderation is impractical. Thus an automatic system for hate speech detection is essential for real-time detection of harmful content and creating a safer online space.

The subjective and context-dependent nature of hate speech makes detecting hate speech a complex problem. As the meaning of specific phrases varies across cultures and social and situational factors, it becomes more challenging to understand the context. Sometimes, it is tough to distinguish between hate speech and legitimate expressions like satire or criticism. This problem often requires a nuanced understanding of the language. Also, certain words or slang in social media are uncommon in daily conversation, making it difficult to identify as hate speech. Various research has been conducted in the Natural Language Processing (NLP) domain to detect hate speech. Most previous work concentrated on a single domain like text or audio (Alkomah and Ma, 2022; Imbwaga et al., 2024). The multimodal aspects of the problem make it even more difficult. We proposed a cross-modal attention-based approach to fuse text and audio in this shared task on Multimodal Hate Speech Detection in Dravidian languages (Premjith et al., 2024a,b). The main contributions of this work are:

- Proposed a cross-modal attention-based approach to fuse two modalities for hate speech detection in Telugu.
- Investigated several transformers and DL models for hate speech detection in Telugu exploiting textual and audio features.

2 Related Work

Many studies have been conducted in recent years to identify hate speech. Sreelakshmi et al., 2024 presented a mix of multilingual transformer-based

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 567–573

embedding models with ML classifiers to detect hate speech and foul language in CodeMix Dravidian languages. After examining models such as MuRIL, BERT, and XLM, they discovered that MuRIL, combined with an SVM classifier, achieved the best performance across Kannada-English, Malayalam-English, and Tamil-English datasets, with accuracies up to 96%. Their study also featured a cost-sensitive learning strategy to address class imbalance, as well as a novel annotated Malayalam-English CodeMix dataset. Hakim et al., 2024 presented a combination of transformer and deep learning models to identify hate speech in Indonesian tweets. Combining IndoBERTweet, BiLSTM, and CNN resulted in an F-1 score of 85.06%

Talking about multiple modalities, Arya et al., 2024 have identified hate speech in memes using the Contrastive Language-Image Pre-Training (CLIP) model with prompt engineering. They have used the Facebook Hateful Meme dataset (Kiela et al., 2020), which contains two modalities (Text and Image). Their finetuned CLIP model scored F-1 score of 90.12%. Mandal et al., 2024 also proposed a technique for identifying hate speech using transformers. Their dataset also contains two modality, but this time, audio and text (English). They have used a new fusion technique called Attentive Fusion, which helped their model to get F-1 score of 92.70%. Similarly, Imbwaga et al., 2024 offered numerous machine learning-based approaches to identify hate speech in English and Kiswahili from audio. The Extreme Gradient Boosting Model achieved the highest F-1 score (96.10%) in Kiswahili, whereas Random Forest achieved the highest F-1 score (90.00%) in English.

There has been a lack of research on identifying hate speech in Telugu using audio and text. This work developed a multimodal framework leveraging transformers to bridge this gap.

3 Dataset and Task Description

This task (Lal G et al., 2025) mainly focused on creating models that accurately detect hate speech in Telugu speech and texts. This work used a multimodal hate speech dataset created by Anilkumar et al., 2024. The dataset includes five hate speech classes: Gender (G), Political (P), Religious (R), Personal Defamation (PD), and Non-hate (NH). The definition (Sharif et al., 2022) of the classes are illustrated in the following:

- **Gender (G):** Use offensive references to body parts, sexual orientation, sexuality, or other pornographic material to harm a person or group.
- **Political (P):** Criticize political ideologies, provoke party supporters, or stir people against the government and police enforcement.
- **Religious (R):** Provoke violence by insulting a religion, religious group, or religious beliefs (Catholic, Hindu, Jewish, or Islamic, among others).
- **Personal Defamation (PD):** Act of making false statements about an individual that harm their reputation.
- **Non-hate (NH):** Do not make any rude comments or convey any hostile intent to hurt other people mentally or physically.

Modality	**Train**	**Test**	**Total**
Text	556	50	606
Audio	551	50	601
Total	1107	100	1207
T_W **(Text)**	13170	1064	14234
T_{UW} **(Text)**	6598	696	7294
T_{avg} **(Text)**	23	21	–
A_{avg} **(Audio)**	1055	918	–
D_{avg} **(Audio)**	12	10	–

Table 1: Dataset statistics for Task-3. The symbols T_W and T_{UW} denote the total and unique words in the text, whereas A_{avg} indicates the average audio size in KB and D_{avg} indicates the average duration of audio in seconds.

The dataset comprises 556 texts and 551 audio in the training set and 50 texts and 50 audio in the test set. Task-3 concerns multimodal hate speech detection in Telugu. Table 1 shows the distribution of dataset into train, and test sets. The source code is publicly available at https://github.com/ashrafulparan2/SemanticCuetSync-DravidianLangTech-2025.

4 System Overview

This work exploited several transformer-based models to address task 3. Textual and audio features train the models and fuse the outputs with cross-modal attention. Figure 1 illustrates the

schematic configuration of the proposed multimodal hate speech detection solution.

4.1 Feature Extraction

The feature extraction involves two independent processes for text and audio modalities.

4.1.1 Text

We have investigated 8 transformer-based models and 1 DL model for textual feature extraction.

- **BERT:** This (Devlin, 2018) transformer-based model was pre-trained and self-supervised on a large corpus of English data. The training process used raw texts and was conducted without human labeling, employing two objectives: masked language modeling (MLM) and next-sentence prediction. As a result, this model has developed a robust understanding of the language's internal representation. In this task, this model was employed for feature extraction.

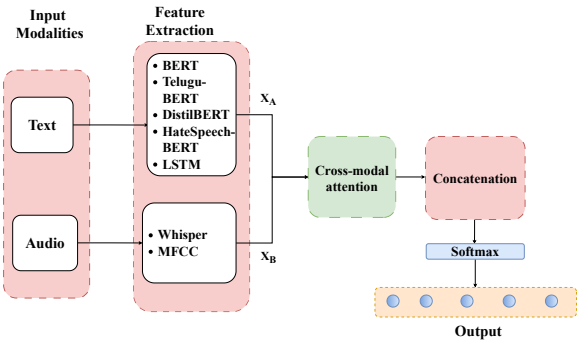

Figure 1: Schematic process for hate speech detection.

- **TeluguBERT (TBERT):** We used five versions of TeluguBERT (Joshi, 2022) for textual feature extraction in this task. This BERT model is trained on a publicly available Telugu monolingual dataset. The extensive training enables the model to capture the rich linguistic nuances, syntax, and semantic patterns unique to the Telugu language, which can be very useful for textual feature extraction.

- **DistilBERT (dBERT):** This (Sanh, 2019) is a smaller, faster, cheaper, and lighter version of the BERT model. A notable characteristic of this model is that it has 40% fewer parameters than the BERT models. As the number of parameters is lower, it is 60% faster. Most importantly, it maintains 95% of BERT's performance as measured on the GLUE language understanding benchmark. We used this model for textual feature extraction in this task.

- **HateSpeechBERT (HS-BERT):** This is a pre-trained BERT-based model specially fine-tuned for detecting abusive speech in Bengali, Devanagari Hindi, code-mixed Hindi, code-mixed Kannada, code-mixed Malayalam, Marathi, code-mixed Tamil, Urdu, and English. We used this model for textual feature extraction.

Table 2 illustrates the hyperparameters used in transformer-based models. The hyperparameters were tuned manually based on empirical observations and iterative experimentation.

Models	LR	WD	WS	EP
Unimodal (text)	5e-5	0.30	50	10
Unimodal (Audio)	3e-5	0.01	0	5
Bimodal	1e-5	0.00	0	10

Table 2: Hyperparameters for transformer-based models.

4.1.2 Audio Features

- **Whisper:** Whisper (Radford et al., 2023) is a state-of-the-art pre-trained model developed for automatic speech recognition (ASR). It is also trained for speech translation. This model is trained on approximately 680k hours of labeled data. This vast training corpus enables Whisper to demonstrate a strong ability to generalize to many datasets and domains. We used this model for auditory feature extraction because of its robust performance and multilingual capabilities.

- **MFCC:** Mel-frequency Cepstral Coefficients (MFCC) is another popular auditory feature extractor used in this task for detecting hate speech. It is designed to mimic the way humans perceive sound and speech. It analyzes the power spectrum of the audio signal and maps it to the Mel scale.

4.2 Cross-modal Attention

After the feature extraction steps, we used a cross-modal attention (Ye et al., 2019) mechanism between the audio-text pair. Cross-modal attention can be represented by the Eqs. (1)-(5).

1. Query, Key, and Value Projections:

$$Q = Z_A W_Q \quad (1)$$

$$K = Z_B W_K \quad (2)$$

$$V = Z_B W_V \quad (3)$$

2. Scaled Dot-Product Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (4)$$

3. Concatenation:

$$\text{Output} = \text{Concat}(\text{Attention}(Q, K, V), \dots) \quad (5)$$

The equation of concatenation:

$$\alpha_{\text{concat}} = [\alpha_1; \alpha_2; \cdots; \alpha_n] \quad (6)$$

Here, Z_A represents features from modality A, and Z_B represents features from modality B.

4.3 Fusion

In this step, we concatenated the output from the cross-modal attention layers. This is used to produce the final output. Equation 7, the early fusion approach, concatenates audio and text features.

$$F = [F_{\text{audio}} \oplus F_{\text{text}}] \quad (7)$$

Here, F is the fused feature representation, F_{audio} represents the feature vector extracted from the audio modality, F_{text} represents the feature vector extracted from the text modality, and $\oplus$ denotes the concatenation operation.

5 Results and Analysis

Table 3 demonstrates the evaluation results of unimodal and bimodal models on the test set.

Among unimodal (Text) models, mBERT surpasses all others with the highest F-1 score of 49.68%. dBERT scores the lowest, with an F-1 score of 17.46%. We analyzed numerous TBERT versions, and TBERT-5 had the highest F-1 score of 38.10%. Among unimodal (Audio) models, Hubert surpasses all others with an F-1 score of 22.94%.

For Bimodal (Audio+Text), we have explored several transformer-based models with early fusion. Whisper-small and TBERT-3 with early fusion outperform all other models with an F-1 score of 43.68%. However, for TBERT versions 4 and 5, the F-1 score decreases gradually. Whisper-small with HS-BERT results in the lowest F-1 score of 28.12%. Appendix D illustrates the detailed error analysis of the best-performed models (mBERT and Whisper-small+TBERT-3).

Unimodal (Text)				
Classifier	**Pr(%)**	**Re(%)**	**F1(%)**	**Ac(%)**
SVM	68.25	32.63	31.60	48.65
Random Forest	28.80	37.20	32.18	52.25
Logistic Regression	68.70	37.47	37.79	53.15
CNN	46.91	38.80	33.58	56.76
CNN + LSTM	30.07	38.00	31.01	51.35
CNN + BiLSTM	36.52	43.20	39.12	58.56
TBERT-1	36.42	40.00	37.20	40.00
TBERT-2	40.00	40.00	37.98	40.00
TBERT-3	34.00	34.38	34.00	33.57
TBERT-4	32.00	47.28	32.00	29.96
TBERT-5	40.61	38.00	38.10	38.00
dBERT	12.52	30.00	17.46	30.00
mBERT	50.94	50.00	**49.68**	50.00
Unimodal (Audio)				
Classifier	**Pr(%)**	**Re(%)**	**F1(%)**	**Ac(%)**
Whisper-small	14.86	20.00	17.00	20.00
Hubert	17.13	38.00	**22.94**	38.00
Wav2vec2	7.74	20.00	10.34	20.00
Bimodal				
Classifier	**Pr(%)**	**Re(%)**	**F1(%)**	**Ac(%)**
Whisper-small + BERT	34.88	38.00	33.58	38.00
Whisper-small + TBERT-1	42.69	38.00	32.78	38.00
Whisper-small + TBERT-2	40.00	40.00	37.98	40.00
Whisper-small + TBERT-3	43.44	46.00	**43.68**	46.00
Whisper-small + TBERT-4	21.87	36.00	27.05	36.00
Whisper-small + TBERT-5	31.28	46.00	35.21	46.00
Whisper-small + dBERT	32.90	46.00	36.47	46.00
Whisper-small + HS-BERT	40.00	22.18	28.12	40.00
MFCC + LSTM	13.63	18.00	9.25	18.00

Table 3: Performance of the employed models for the tasks.

6 Error Analysis

We have analyzed the proposed model's performance to illustrate a quantitative and qualitative error analysis.

Quantitative Analysis

Figure 2 depicts the confusion matrix for the test set, categorizing speeches into their appropriate classes. The findings suggest that 23 out of 50 speeches were properly predicted. Among the five categories, "Personal Defamation" was the most precisely identified, while "Gender" and "Political" hate speeches were only correctly identified once each. Overall performance was unsatisfactory, owing to the dataset's limited size.

Qualitative Analysis

Figure 3 displays predicted outputs and their corresponding true labels for some randomly selected samples, demonstrating the proposed model's performance. The model frequently struggles to appropriately understand the intent underlying the tone of a speech. The same speech may convey multiple meanings, depending on the tone in which it is delivered. For example, in the second case, the model failed to catch the subtle tone of the speech,

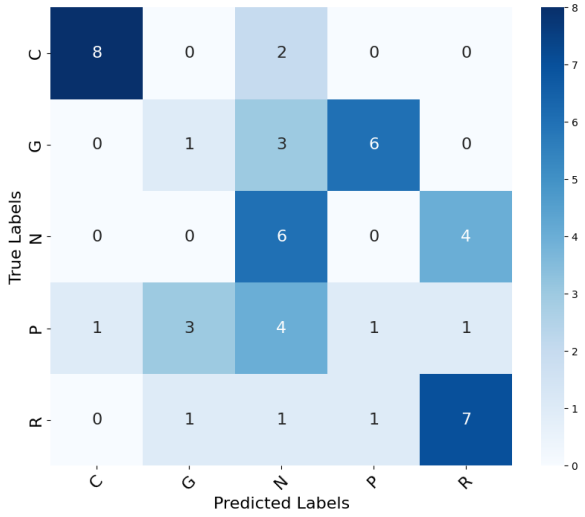

Figure 2: Confusion matrix of the best performing model.

resulting in a misclassification. This shortcoming is mostly due to the model is trained on a very limited dataset, which limits its capacity to accurately recognize hate speech.

Transcript	AL	PL
ఎవరు మాత్రం SC కులంలో పుట్టాలని కోరుకుంటారు (Who wants to be born in SC caste)	Religious	Religious
ఎస్సీలుగా పుట్టాలని ఎవరు కోరుకుంటారు (Who wants to be born as SCs?)	Religious	Gender
నీ బండారం బయటపెట్టి పొలం మధ్య నిలబెట్టి గుడ్డలు ఊడదీస్తా (You will take out your barn and stand it in the middle of the field and blow the rags)	Personal Defamation	Personal Defamation
హిందువులెవ్వరు ముస్లింలకు వ్యతిరేకం కాదని కూడా ఆయన తన అభిప్రాయంగా చెప్పారు (He also said in his opinion that no Hindu is against Muslims	Non-Hate	Non-Hate

Figure 3: Few randomly selected samples with actual (AL) and predicted labels (PL).

7 Conclusion

This study investigated several transformers and DL techniques in both audio and text modality with cross-modal attention for detecting hate speech in Telugu. Among unimodal (Audio) models, Hubert surpasses all others with an F-1 score of 22.94%. Among various bimodal (audio + text) combinations, Whisper-small + TeluguBERT-3 achieved the highest F1 score of 43.68%. However, we found that mBERT achieves a higher F1 score of 49.68% using text only. This study demonstrates that the textual unimodal approach gives us a superior performance. Further improvements can be made by increasing the dataset and using other multimodal models. Besides, exploring various LLMs may improve results for detecting Telugu hate speech.

Limitations

The current implementations possess some weaknesses, such as (i) The dataset is limited in size, so the model suffers from generalization issues, and (ii) the noise and recording quality of audio data affect performance.

Acknowledgments

We thank the DravidianLangTech 2025 shared task organizers for running this task. This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET).

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Abhishek Anilkumar, Jyothish Lal G, B Premjith, and Bharathi Raja Chakravarthi. 2024. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *Speech and Language Technologies for Low-Resource Languages (SPELLL)*, Communications in Computer and Information Science.

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2018. What does this imply? examining the impact of implicitness on the perception of hate speech. In *Language Technologies for the Challenges of the Digital Age: 27th International Conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings 27*, pages 171–179. Springer.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Atalla Naufal Hakim, Yuliant Sibaroni, and Sri Suryani Prasetyowati. 2024. Detection of hate-speech text on indonesian twitter social media using indobertweet-bilstm-cnn. In *2024 12th International Conference on Information and Communication Technology (ICoICT)*, pages 374–381. IEEE.

Joan L Imbwaga, Nagatatna B Chittaragi, and Shashidhar G Koolagudi. 2024. Automatic hate speech detection in audio using machine learning algorithms. *International Journal of Speech Technology*, 27(2):447–469.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Teo Keipi, Matti Näsi, Atte Oksanen, and Pekka Räsänen. 2016. *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Jyothish Lal G, B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Rajalakshmi Ratnavel. 2025. Overview of the Shared Task on Multimodal Hate Speech Detection in Dravidian Languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, and Sudip Kumar Naskar. 2024. Attentive fusion: A transformer-based approach to multimodal hate speech detection. *arXiv preprint arXiv:2401.10653*.

John T Nockleby. 1994. Hate speech in context: The case of verbal threats. *Buff. L. Rev.*, 42:653.

B Premjith, Bharathi Raja Chakravarthi, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Karnati, Sai Mangamuru, and Chandu Janakiram. 2024a. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 49–55.

B Premjith, G Jyothish, V Sowmya, Bharathi Raja Chakravarthi, K Nandhini, Rajeswari Natarajan, Abirami Murugappan, B Bharathi, Saranya Rajiakodi, Rahul Ponnusamy, et al. 2024b. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@ dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2022. M-bad: A multilabel dataset for detecting aggressive texts and their targets. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 75–85.

K Sreelakshmi, B Premjith, Bharathi Raja Chakravarthi, and KP Soman. 2024. Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. *IEEE Access*.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511.

A Class-wise Distribution of Dataset

Figure A.1 presents the class-wise distribution of the dataset, illustrating the frequency of samples across five distinct categories: Non-hate, Personal Defamation, Gender, Religious, and Political. The Non-hate category constitutes the largest proportion, with 198 instances, followed by Personal Defamation (122), Gender (101), Religious (72), and Political (58). This distribution highlights a class imbalance, with Non-hate being the dominant class, which may influence model training and performance.

Figure A.2 illustrates a few examples of the input and output of the dataset.

B System Requirements

This study was developed using Python 3 (version 3.10.12) and Python-based libraries from the PyTorch 2 framework to implement transformers, including BERT, TBERT, dBERT, and Whisper-small. The implementation required 29GB of RAM, 16GB of VRAM, and 73.1GB of storage space. We utilized an NVIDIA Tesla P100 GPU on Kaggle. For data analysis and preprocessing, we employed pandas (2.1.4) and numpy (1.24.3). For unimodal, ML models were built using scikit-learn (1.2.2), while DL models were trained with Keras

(2.13.1) and TensorFlow (2.13.0). Additionally, PyTorch (2.0.0) and transformers (4.36.2) implement transformer-based bimodal models.

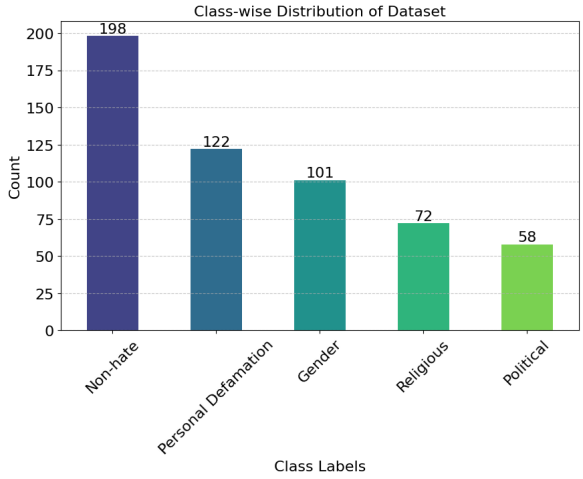

Figure A.1: Class-wise Distribution of the training dataset.

Transcript	Label
ఏసుప్రభు నిజమైన దేవుడు అని చెప్తాకే వచ్చాను నేను. ఛీ తీయ్ బండి తీయ్ (I have come to tell you that Jesus is the true God. Chee Tee Bandi Tee)	Religious
వైయస్సార్ కాంగ్రెస్ అనగానే ఆగుమాటిక్ గ అవినీతి, అది రెండు పార్టీలు మధ్య ఉన్న వేత్తయసం. (YSR Congress is an autocratic corruption, it is a conflict between two parties.	Political
నా మీద బతికి ఉన్నా గోజ్జ లంగాకొడుకులారా (Live on me, you bastards)	Gender
ఒక మనిషిని కదిలించే శక్తి సహితాయ్నకి మతమే ఉంటుంది అక్షరాన్ని మతమే ఉంటుంది (Religion is the power that moves a man, religion is the letter)	Non-Hate

Figure A.2: Task-3 sample with Transcript and label.