JerinWarriors@DravidianLangTech 2026: A Two-Stream Cross-Attention Approach for Prompt Recovery in Telugu

Savith Arivudainambi
IISER Kolkata, West Bengal, India
sa24ms140@iiserkol.ac.in

Wordson Robert
IISER Kolkata, West Bengal, India
wr24ms190@iiserkol.ac.in

C. Jerin Mahibha
Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India
jerinmahibha@msec.edu.in

Shrey Patnaik
IISER Kolkata, West Bengal, India
sp24ms174@iiserkol.ac.in

Abstract

Identifying the structure of detailed sentences which show glimpses of various annotation cues, in a low resource language that is morphological rich like Telugu is a challenge. Standard baseline architectures like Multi Layer Perceptrons (MLP) struggle with low resource languages. This paper details our proposed solution for the Telugu Prompt-Style Recovery Shared Task at DravidianLangTech @ ACL 2026. We propose a Two-Stream Cross-Attention architecture that uses a shared MuRIL encoder to calculate the relationship between an original transcript and its style-shifted counterpart, helping the MLP to distinguish the styles better and catch the differences better. Through experimentation we have found out that this proposed model handles the signal dilution of the individual labels better than the rest. Our best-performing system achieved a Macro F1-score of 0.2588 on the test set, securing 2nd place out of 13 teams. We have concluded that the local transformation is the main driver for the style recovery in this task. For reproducibility, we release our implementation and experimental setup on GitHub.[1]

1 Introduction

Recent advances in Natural Language Processing (NLP) have significantly improved the ability to analyze social media text in low-resource languages. In the context of Dravidian languages, several studies have explored tasks such as sarcasm detection, sentiment analysis, and offensive language identification using machine learning and transformer-based approaches. For example, prior work has investigated sarcasm detection in Dravidian languages using transformer architectures, demonstrating the effectiveness of contextual language models in capturing subtle linguistic cues in social media text (Madhumitha et al., 2023). Similarly, studies on offensive language identification have applied both traditional machine learning and deep learning techniques to detect abusive or harmful content in regional language datasets (Mahibha et al., 2021b).

Other research has explored sentiment analysis in multilingual and cross-lingual settings, showing that cross-lingual word embedding models can effectively transfer semantic knowledge across languages with limited annotated data (Mahibha et al., 2021a). Additionally, sarcasm detection from Dravidian language text has been examined as an important task for understanding nuanced expressions in online communication (Mahibha et al., 2024). These studies highlight the growing interest in applying advanced NLP techniques to Dravidian languages and demonstrate the need for models that can effectively capture linguistic complexity in these morphologically rich languages.

Recognizing the intent and the style used in low resource Indian languages is a very under-focused area in NLP research. The regular use of sentiment analysis fails these languages because they are not trained on Dravidian languages. These context-rich sentences from morphologically diverse languages, especially with context, prove to be tough to classify into categories, namely Formal, Informal, Optimistic, Pessimistic, Humorous, Serious, Inspiring, Authoritative, and Persuasive.

We ran several models, with several different parameters, and we tried a myriad of different combinations of techniques. We found that almost 65 percent of stylistic changes in the paragraphs happen within the first 256 words. We found out that the usual techniques, such as the hierarchical models, fail spectacularly at aggregating their signals, resulting in an F1 score of 2 to 5 percent. This proves that extensive signal dilution happens over these long sentences. The model learned best when it was learning from the semantic difference between the original text and the rewritten version, indicating that the sentences are not captured well

[1] https://github.com/WordsonRobert/telugu-prompt-recovery-llm-acl-2026-shared-task

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 279–283

by the tone of the original sentences alone, meaning the styling is not absolute.

Keeping all these in mind, we implemented a Two-Stream Cross-Attention Classifier built on the MuRIL (Multilingual Representations for Indian Languages) backbone. By allowing the style-changed text to compare with the original transcript, our model captures the delta in intent and framing as shown by the results that we have achieved in the shared task (Premjith et al., 2026).

2 Related Work

The field of Telugu Natural Language Processing has evolved significantly with the advent of transformer-based architectures. Dowlagar and Mamidi (2021) established foundational benchmarks by evaluating BERT and Multilingual BERT (mBERT) for hate speech detection in Telugu. Their research highlighted that while mBERT provides a robust starting point, the language's agglutinative nature requires specialized fine-tuning to capture subtle morphological shifts. Building upon this, Kakarla and Venkata (2025) focused on the complexities of code-mixed Telugu-English content. By utilizing models like TeluguHateBERT and Hindi-Abusive-MuRIL, they demonstrated that pre-training on linguistically similar Indian languages significantly improves the detection of aggressive intent compared to generic multilingual models.

In the broader context of Dravidian languages, Dave et al. (2021) presented a hybrid approach combining MuRIL embeddings with character N-grams. Their work in the DravidianLangTech workshop proved that MuRIL is superior for low-resource scripts like Telugu because it retains regional semantic nuances that are often "washed out" in globally trained models. The effectiveness of multilingual transformer representations for cross-lingual NLP tasks was earlier demonstrated by Devlin et al. (2019), whose Bidirectional Encoder Representations from Transformers (BERT) architecture laid the foundation for contextualized language modelling. Subsequent work on multilingual contextual models such as mBERT showed that shared multilingual representations can transfer knowledge across languages, including low-resource ones (Pires et al., 2019).

Furthermore, Arunachalam and Maheswari (2024) introduced ensemble transformer methods to address social media cleanup. Their study emphasized that custom preprocessing, such as aggressive stemming, is vital for Dravidian scripts to reduce vocabulary sparsity. The development of MuRIL, a multilingual transformer specifically designed for Indian languages, further improved performance on Indic language tasks by incorporating transliteration and parallel corpora during training (Khanuja et al., 2021). Finally, the theoretical shift toward dual-representation learning is supported by Duan et al. (2022), whose work on two-stream transformers showed that cross-attention fusion between original and modified text segments allows a model to adaptively extract "difference features" that single-stream encoders often ignore. These studies collectively justify the use of a MuRIL-based two-stream cross-attention model for stylistic and offensive speech classification in Telugu.

3 Data Preparation and Style Definitions

The dataset for this task consists of Telugu transcript excerpts paired with their style-shifted versions. To ensure model robustness and prevent label leakage, we performed several preprocessing steps prior to training.

3.1 Data Labeling

The task requires classification into nine distinct categories. We mapped these labels to a numerical index as follows:

ID	Stylistic Category
0	Authoritative
1	Formal
2	Humorous
3	Informal
4	Inspiring
5	Optimistic
6	Persuasive
7	Pessimistic
8	Serious

Table 1: Label mapping for the 9-way Telugu style classification.

3.2 Preprocessing

We utilized the `google/muril-base-cased` tokenizer to process the Telugu Unicode script. Based on our locality analysis, which indicated that over 60% of stylistic cues are concentrated at the beginning of the text, we applied a truncation strategy with a **maximum sequence length of 256 tokens**. In the test phase, where original transcripts were unavailable, the architecture transitioned to a

self-attention configuration by supplying the style-shifted text to both input branches. Although explicit semantic comparison between original and transformed transcripts could not be performed during inference, this design preserved architectural consistency across the feature extraction layers. The representations learned during cross-attention training continued to provide stylistically discriminative features, enabling effective generalization even in the absence of reference transcripts.

4 Methodology

Our proposed architecture, the **Two-Stream Cross-Attention Classifier**, is designed to capture the stylistic "delta" between an original Telugu transcript and its rewritten version. Unlike single-stream models that treat style as an absolute property of a single text, our approach models style as a relational transformation.

4.1 Formalism

Let $\mathcal{S}_{orig}$ and $\mathcal{S}_{chg}$ denote the input sequences for the original transcript and the style-shifted text, respectively. Both sequences are passed through a shared MuRIL encoder $\mathcal{E}$ to obtain dense hidden representations:

$$H_{orig} = \mathcal{E}(\mathcal{S}_{orig}), \quad H_{chg} = \mathcal{E}(\mathcal{S}_{chg}) \tag{1}$$

where $H \in \mathbb{R}^{L \times d}$ (with $L = 256$ and $d = 768$). To capture the specific stylistic shifts, we implement a Cross-Attention mechanism where the style-changed hidden states act as the Query (Q), while the original hidden states provide the Key (K) and Value (V):

$$Q = H_{chg}W^Q, \quad K = H_{orig}W^K, \quad V = H_{orig}W^V \tag{2}$$

To prevent signal dilution from padding tokens, we apply a Key Padding Mask $M \in \{0, -\infty\}^L$. The attended representation Z is computed as:

$$Z = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{3}$$

Following the attention operation, we apply a residual connection and Layer Normalization (LN) to stabilize the enriched features:

$$H_{final} = \text{LN}(H_{chg} + \text{Dropout}(Z)) \tag{4}$$

The final stylistic classification $\hat{y}$ is derived from the pooled representation of the `[CLS]` token at index 0. Figure 1 shows our complete architecture.

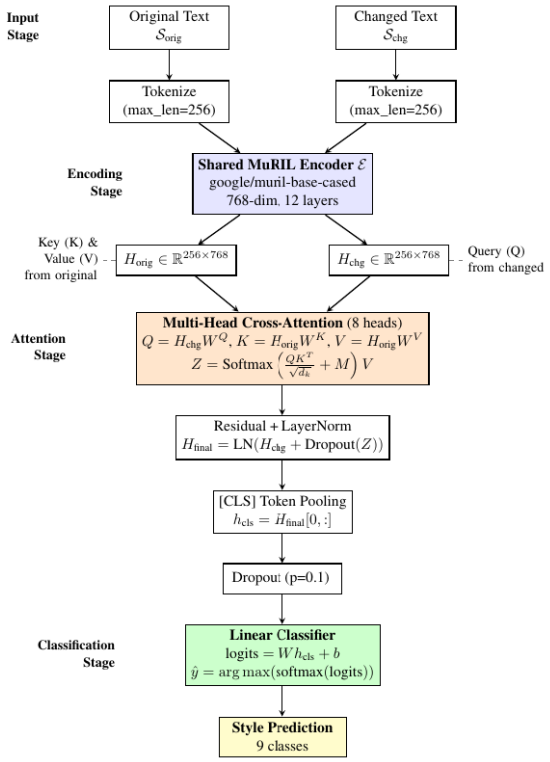

Figure 1: Two-Stream Cross-Attention architecture.

5 Results and Discussion

5.1 Performance Metrics

We evaluated the model over 5 epochs using the AdamW optimizer with a learning rate of 2×10^{-5}. The performance reached its peak at Epoch 3, as shown in Table 2.

Metric	Value
Total Samples	301
Correct	81 (26.9%)
Errors	220 (73.1%)
Accuracy	0.2691
F1 (Macro)	0.2588

Table 2: Test set performance.

5.2 Analysis of Findings

The Overfitting Pivot: We observed that while Training Loss continued to decline from 2.14 to 1.20, the Validation Loss began to rise after Epoch 3. This indicates that the model began to memorize specific vocabulary markers for the nine style labels rather than learning generalizable stylistic shifts.

Locality vs. Hierarchy: Our experiments with 256-token truncation were significantly more successful than hierarchical "chunking" methods. This

True → Predicted	Count	%
Serious → Formal	17	7.7%
Authoritative → Formal	12	5.5%
Humorous → Informal	11	5.0%
Inspiring → Persuasive	9	4.1%
Serious → Informal	9	4.1%
Optimistic → Persuasive	9	4.1%
Humorous → Formal	8	3.6%
Inspiring → Optimistic	8	3.6%
Other	167	75.9%
Total	**220**	**100.0%**

Table 3: Top misclassifications.

Style	Prec	Rec	F1
Authoritative	0.33	0.10	0.15
Formal	0.20	0.39	0.27
Humorous	0.35	0.26	0.30
Informal	0.27	0.43	0.33
Inspiring	0.33	0.11	0.16
Optimistic	0.34	0.36	0.35
Persuasive	0.24	0.43	0.31
Pessimistic	0.30	0.36	0.33
Serious	0.27	0.09	0.13
Macro	**0.29**	**0.28**	**0.26**

Table 4: Per-class metrics.

suggests that stylistic markers in Telugu are heavily localized; the dilution of signals in longer sequences (as seen in the 2-5% F1 scores of hierarchical models) is a major bottleneck for style recovery in low-resource languages.

5.3 Error Analysis

The confusion matrix represented in Figure 2 reveals three semantic clusters: Professional styles (Formal–Authoritative–Serious, 41 confusions), Casual styles (Informal–Humorous, 14 confusions), and Motivational styles (Inspiring–Optimistic–Persuasive, 26 confusions). The model systematically overpredicted Formal (75 predictions vs. 38 true) while underpredicting Serious (15 vs. 47 true). The count and percentage of misclassifications are represented in Table 3. The performance metrics such as precision, recall, and F1 score, associated with the different style categories are represented in Table 4. Best performers were Optimistic (F1=0.35) and Informal (F1=0.33) with clear lexical markers, while Serious (F1=0.13) and Authoritative (F1=0.15) struggled due to subtle tonal differences the 256-token window failed to capture.

6 Conclusion

This paper presented a Two-Stream Cross-Attention approach for the Telugu Prompt-Style Recovery task. Through cross-attention-based representation learning between original and style-shifted texts during training, the proposed system effectively captures stylistically discriminative semantic patterns in morphologically rich sentences. Experimental findings indicate that localized semantic transformations, particularly those occurring in the initial segments of the document, contribute significantly to style identification. Furthermore, the learned representations demonstrate strong generalization capability during inference even when the original reference transcript is unavailable. Future work will explore the use of RAG-based systems to recover missing transcripts in the test sets.

References

V Arunachalam and N Maheswari. 2024. Enhanced detection of hate speech in dravidian languages in social media using ensemble transformers. *Interdisciplinary Journal of Information, Knowledge, and Management*, 19:036.

Bhargav Dave, Shripad Bhat, and Prasenjit Majumder. 2021. Irnlp_daiict@ dravidianlangtech-eacl2021: offensive language identification in dravidian languages using tf-idf char n-grams and muril. In *Proceedings of the first workshop on speech and language technologies for Dravidian Languages*, pages 266–269.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Suman Dowlagar and Radhika Mamidi. 2021. Hasocone@ fire-hasoc2020: Using bert and multilingual bert models for hate speech detection. *arXiv preprint arXiv:2101.09007*.

Lihua Duan, Qi You, Xinke Wu, and Jun Sun. 2022. Multilabel text classification algorithm based on fusion of two-stream transformer. *Electronics*, 11(14):2138.

Santhosh Kakarla and Gautama Shastry Bulusu Venkata. 2025. Code-mixed telugu-english hate speech detection. *arXiv preprint arXiv:2502.10632*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and Sebastian Ruder. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

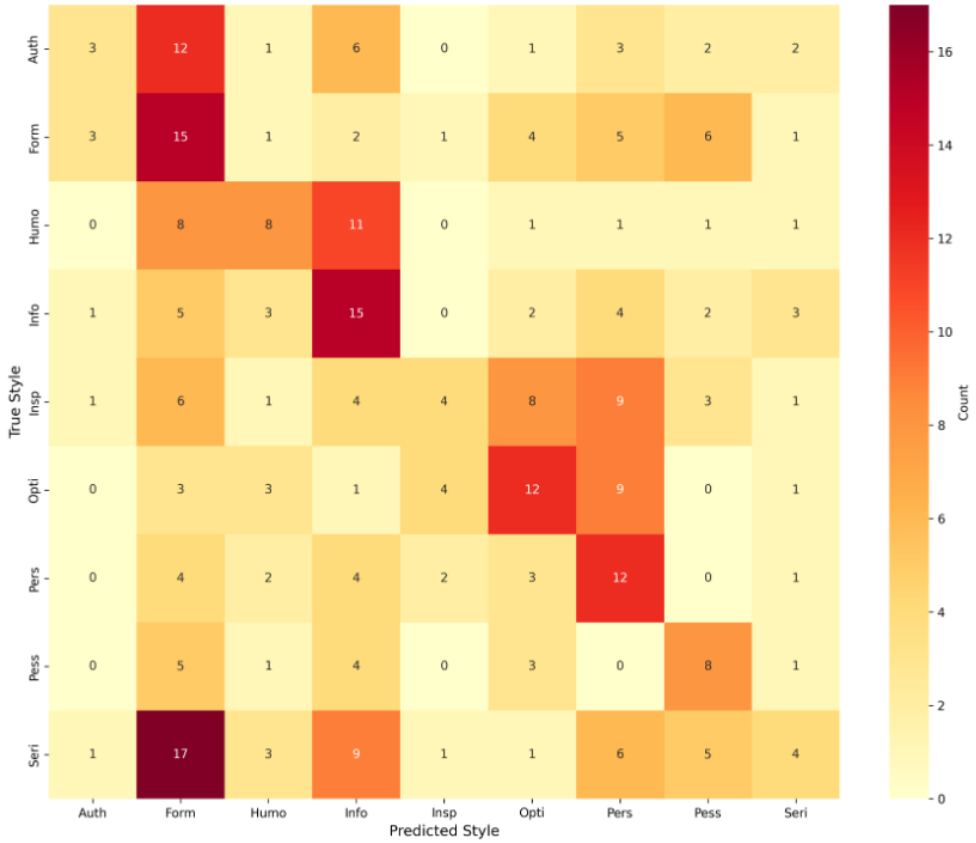

Figure 2: Confusion Matrix

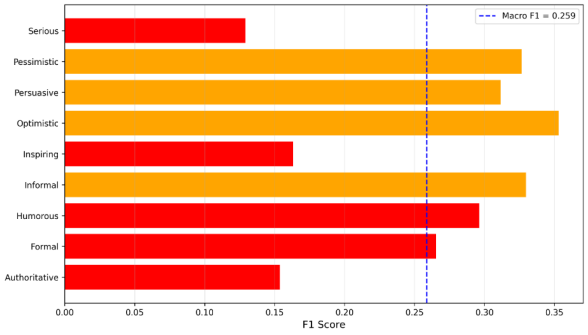

Figure 3: Per class F1-Score

M Madhumitha, Kunguma Akshatra M, J Tejashri, Mahibha C Jerin, and Durairaj Thenmozhi. 2023. Sarcasm detection in dravidian languages using transformer models. In *FIRE (Working Notes)*, pages 306–318.

C Jerin Mahibha, Kayalvizhi Sampath, and Durairaj Thenmozhi. 2021a. Sentiment analysis using cross lingual word embedding model. In *FIRE (Working Notes)*, pages 1094–1100.

C Jerin Mahibha, Kayalvizhi Sampath, Durairaj Thenmozhi, and S Arunima. 2021b. Offensive language identification using machine learning and deep learning techniques. In *FIRE (Working Notes)*, pages 705–713.

C Jerin Mahibha, Gersome Shimi, and Durairaj Thenmozhi. 2024. Sarcasm detection from dravidian language text. In *Forum of Information Retrieval and Evaluation FIRE-2024*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4996–5001.

B Premjith, G Jyothish Lal, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Thenmozhi Durairaj, Ratnavel Rajalakshmi, Rahul Ponnusamy, and Chinthala Bhuvanesh. 2026. Shared Task on Prompt Recovery for LLM in Telugu. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.