Cuet Yet Another Baseline @ DravidianLangTech 2026: Shared Task on Prompt Recovery for LLM in Telugu

Rotna Dipika Debnath, Shahrin Afroz Hoque Ruhi, Ayesha Labiba, Arpita Mallik, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2104002, u2104015}@student.cuet.ac.bd,
{u2104008, u2004023}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Prompt recovery in large language models (LLMs) is the task of inferring the communicative intent and stylistic framing of the original instruction from model-generated output. This task is especially challenging for low-resource Dravidian languages such as Telugu, where agglutinative morphology, register variation, and scarce annotated data complicate stylistic modelling. In this paper, we present our system for the Shared Task on Prompt Recovery for LLM in Telugu at DravidianLangTech @ ACL 2026, which aims to classify Telugu transcript excerpts into nine communicative style categories: Formal, Informal, Optimistic, Pessimistic, Humorous, Serious, Inspiring, Authoritative, and Persuasive.We have implemented a transformer-based approach using ai4bharat/IndicBERTv2-MLM-only, MuRIL-base and Telugu-BERT for Telugu communicative style classification. Our system fine-tunes the pretrained Indic language training samples to capture stylistic patterns in Telugu transcripts. Our approach achieved a macro F1 score of 0.2993 on the evaluation set, demonstrating the potential of Indic-focused pretrained models for stylistic analysis in low-resource language settings.Controlled ablations reveal that label smoothing benefits stronger Indic backbones but degrades weaker ones, and that surface linguistic feature augmentation does not complement rich contextual representations on small datasets.

1 Introduction

Large language models (LLMs) have become the dominant paradigm for natural language processing, with carefully designed prompts being the primary mechanism for controlling model behaviour (Schulhoff et al., 2024).

Recovering the stylistic intent of a prompt from the generated output is critical for interpretability, safety, and controlled generation (Brown et al., 2020; Wei et al., 2022). For Telugu, a Dravidian language with agglutinative morphology and over 80 million native speakers, this problem is further complicated by limited labelled resources and underrepresentation in major multilingual pretraining corpora.

The DravidianLangTech @ ACL 2026 conference(Chakravarthi et al., 2026) has introduced a dataset of 3,601 Telugu transcripts rewritten in one of nine styles by an LLM. Given the rewritten text, a system must predict the style label. We have addressed this task by fine-tuning IndicBERT v2 (Doddapaneni et al., 2023), a multi-stage pretrained encoder covering 23 Indian languages, augmented with label smoothing regularisation. We have compared it against Telugu-BERT(Joshi et al., 2022) and a MuRIL-based hybrid with handcrafted linguistic features(Khanuja et al., 2021).

The main contributions of this work have been:

- Fine-tuning of ai4bharat/IndicBERTv2-MLM-only with label smoothing for Telugu style classification, ranking 1st on the shared-task leaderboard.
- Controlled ablations of label smoothing across three backbones and linguistic feature augmentation on IndicBERT v2, revealing when each technique helps or hurts.
- A reproducible pipeline with fixed random seeding and early stopping for stable convergence.

Further details of the implementation can be accessed through the GitHub repository: [1].

[1] https://github.com/prism-rkive/prompt_recovery_indic

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 191–195

2 Related Work

IndicBERT (Kakwani et al., 2020) introduced an ALBERT-based compact multilingual encoder for 12 Indian languages. IndicBERT v2 (Doddapaneni et al., 2023) extended this with a three-stage pretraining pipeline (MLM, TLM, task-adaptive) using the XLM-RoBERTa architecture and SentencePiece tokenisation across 23 languages, yielding large gains on downstream Indic NLP tasks. MuRIL (Khanuja et al., 2021) targets transliteration and code-mixing phenomena across 17 Indian languages. Telugu-BERT (Joshi et al., 2022) demonstrated that monolingual Telugu pretraining on 6.2 GB of text outperforms multilingual baselines on Telugu classification tasks.

Fine-tuned BERT representations outperform traditional stylometric pipelines for authorship and style attribution (Fabien et al., 2020). Integrating handcrafted linguistic features with Transformer embeddings via early fusion further improves performance in low-resource settings (Zhao et al., 2021; Nitu et al., 2024). Label smoothing (Müller et al., 2019) improves calibration in multi-class settings by preventing overconfident predictions, which is particularly beneficial when class boundaries are subtle and nine-way confusion is likely. Focal Loss (Lin et al., 2017) and SMOTE (Chawla et al., 2002) address class imbalance but provide no benefit when the dataset is already balanced (Section 3).

3 Data

We have utilized the dataset provided under the Shared Task on Prompt Recovery for LLM in Telugu DravidianLangTech @ ACL 2026 (Chakravarthi et al., 2026). The dataset supports style transfer and controlled text generation, where each instance pairs an original Telugu transcript with a stylistically rewritten version of the same content.

The dataset is provided in Excel (.xlsx) format with four columns: ID, ORIGINAL TRANSCRIPTS, CHANGE STYLE, and STYLE. The ORIGINAL TRANSCRIPTS column contains the source text, CHANGE STYLE provides the style-modified input, and STYLE indicates the target label.

The dataset is split into 3,000 training, 300 development, and 301 test instances, as shown in Table 1. The training split is nearly balanced across all nine classes (321–347 samples,10.7%-11.6%), as shown in Table 2, confirming that class imbalance techniques such as Focal Loss and SMOTE provide no structural benefit.

Language	Train	Development	Test	Total
Telugu	3,000	300	301	3,601

Table 1: Dataset statistics for the Prompt Recovery for LLM in Telugu shared task.

Style Class	Train	%	Class Weight
Pessimistic	347	11.6	0.96
Humorous	344	11.5	0.98
Authoritative	338	11.3	0.99
Persuasive	336	11.2	1.02
Inspiring	332	11.1	0.99
Optimistic	331	11.0	0.99
Formal	327	10.9	1.01
Serious	324	10.8	1.01
Informal	321	10.7	1.02
Total	3,000	100.0	0.96–1.02

Table 2: Style class distribution in the training split. Class weights are computed as $w_c = \bar{n}/n_c$ where $\bar{n}$ is the mean class count.

4 Methodology

4.1 Data Preprocessing

In terms of data preprocessing, labels have been normalized by trimming whitespace and converting to title case, and samples with missing text or labels have been discarded. Each label has been mapped to one of nine style indices: Formal (0), Informal (1), Optimistic (2), Pessimistic (3), Humorous (4), Serious (5), Inspiring (6), Authoritative (7), and Persuasive (8). A fixed random seed (42) has been set for Python, NumPy, and PyTorch (including CUDA) with deterministic CuDNN for reproducibility.

4.2 Overview of Experimented Models

4.2.1 IndicBERT v2

In this approach, We have fine-tuned ai4bharat/IndicBERTv2-MLM-only(278M parameters, XLM-RoBERTa architecture) (Doddapaneni et al., 2023) for sequence classification using HuggingFace Transformers. A randomly initialised linear head projects the 768-dimensional

[CLS] token embedding from the final hidden layer to nine output logits. Input texts from the CHANGE STYLE column are tokenised with SentencePiece (max_length= 256, padding and truncation enabled); the 200K-subword vocabulary covers 23 Indian languages, providing robust segmentation for agglutinative Telugu morphology. The overall architecture is illustrated in Figure 1.

The model is optimised with AdamW (lr $= 2\times10^{-5}$, weight_decay $= 0.01$, batch_size $= 16$) under a linear learning rate decay schedule with no warmup over 1,880 total steps (188 batches $\times$ 10 epochs). Gradients are clipped to max_norm $= 1.0$.

The loss function is CrossEntropyLoss with label smoothing $\varepsilon = 0.1$, which replaces the hard one-hot target for the true class y with a soft distribution:

$$\tilde{q}_k = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{K} & k = y, \\ \dfrac{\varepsilon}{K} & k \neq y, \end{cases} \tag{1}$$

where $K = 9$. Following Müller et al. (2019), this prevents overconfident predictions — particularly important here because adjacent style classes (e.g., Formal, Serious, Authoritative) share substantial lexical and syntactic overlap. Dropout $= 0.3$ is applied to the classification head as additional regularisation. The label smoothing coefficient $\varepsilon = 0.1$ and classification-head dropout of 0.3 follow the fine-tuning recommendations of Müller et al. (2019) and Doddapaneni et al. (2023) respectively; a full grid search was infeasible given the shared-task timeline.

Training runs for at most 10 epochs, saving the checkpoint with the highest validation Macro F1. Training stops early if no improvement is seen for three consecutive epochs (patience $= 3$); the best checkpoint was obtained at epoch 4. At inference, predicted labels are produced by $\hat{y} = \arg\max_k \text{logit}_k$ and mapped back to style-name strings for submission.

4.2.2 Other Experimented models

We have explored various other methods such as fine-tuning l3cube-pune/telugu-bert (238M parameters, BERT-large with 24 layers) using AdamW with cosine learning rate decay, where the official training and validation splits were merged and re-partitioned into a 90/10 train–validation split with a three-seed ensemble to improve stability. We have also experimented with augmenting google/muril-base-cased (237M parameters) with a 44-dimensional handcrafted Telugu linguistic feature vector, fused with the [CLS] embedding through a multilayer perceptron (MLP) before the final classification layer (Qi et al., 2020).To ensure a fair comparison, we subsequently applied identical label smoothing regularisation ($\varepsilon = 0.1$) to both Telugu-BERT and MuRIL; even under matched training conditions both were outperformed by IndicBERT v2 (Table 3).

5 Results and Analysis

5.1 Comparative Analysis and Ablation

Table 3 presents test-set results for all systems. Our primary IndicBERT v2 system achieves a test Macro F1 of 0.2993 (Precision: 0.3146, Recall: 0.3043, Accuracy: 0.2890). Expanding max_length from 256 to 512 yields a marginal gain (+0.0038 F1), confirming that truncation is a contributing but not primary bottleneck. Telugu-BERT's F1 decreases with label smoothing (0.2160 $\rightarrow$ 0.1713), confirming the calibration benefit requires a strong backbone to amplify. Adding linguistic features to IndicBERT v2 decreases F1 to 0.2528, with Serious collapsing to F1 $= 0.000$, suggesting the fusion MLP overfits on surface proxies when the backbone already encodes rich contextual representations.

System	LS	Test Macro Average			
		P	R	F1	Acc.
IndicBERT v2 †	✓	0.3146	0.3043	**0.2993**	0.2890
IndicBERT v2 (512 tok.) ‡	✓	0.3046	0.3072	0.3031	0.2924
IndicBERT v2 + Ling. Feat. ‡	✓	0.2448	0.2710	0.2528	0.2558
Telugu-BERT	×	0.2354	0.2634	0.2160	0.2700
Telugu-BERT (w/ LS)	✓	0.1885	0.2289	0.1713	0.2159
MuRIL + Ling. Feat. (w/ LS)	✓	0.2406	0.2356	0.2285	0.2193

Table 3: Test-set and ablation results. † = primary submission; ‡ = post-hoc analysis, not submitted; LS = label smoothing ($\varepsilon = 0.1$). Macro F1 is the official metric.

5.2 Per-class Performance

Table 4 shows per-class results. Optimistic achieves the highest F1 (0.4848) while Serious is the weakest (F1 $= 0.0845$), likely due to its dependence on sustained register consistency across passages truncated at max_length= 256.

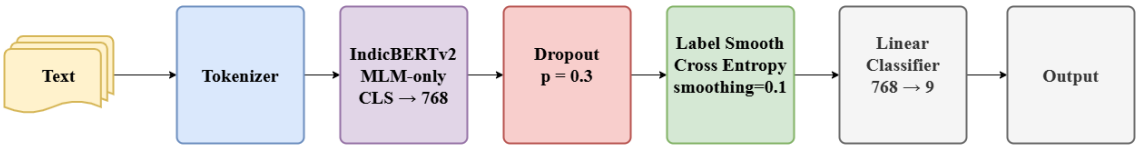

Figure 1: Architecture of the IndicBERT v2 system for Telugu style classification.

Style	P	R	F1	Support
Formal	0.1528	0.2895	0.2000	38
Informal	0.3514	0.3714	0.3611	35
Optimistic	0.4848	0.4848	0.4848	33
Pessimistic	0.4000	0.3636	0.3810	22
Humorous	0.2424	0.2581	0.2500	31
Serious	0.1250	0.0638	0.0845	47
Inspiring	0.5000	0.2432	0.3273	37
Authoritative	0.2500	0.2000	0.2222	30
Persuasive	0.3250	0.4643	0.3824	28
Macro avg	**0.3146**	**0.3043**	**0.2993**	**301**

Table 4: Per-class test-set results for our IndicBERT v2 system.

5.3 Error Analysis

Figure 2 shows the test-set confusion matrix. Formal is over-predicted (67 of 301), acting as a catch-all for neutral-register text. Serious is the weakest class (recall = 0.0638) and Inspiring shows high precision (0.50) but low recall (0.2432), indicating the model is conservative about predicting it. Three dominant confusion patterns, all sharing the same root cause, are detailed in Table 5: stylistic signals that disambiguate adjacent categories are discourse-level and passage-final, not token-level and passage-initial.

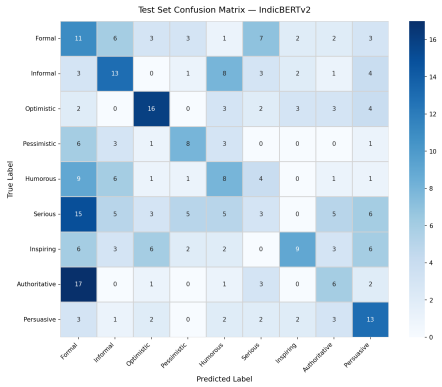

Figure 2: Confusion Matrix of the IndicBERT v2 Model.

Sample	True	Pred.	Reason
ప్రాచీన భారతదేశ చరిత్ర: అధ్యయన విధానం మరియు ఆధారాలు (Ancient Indian History: Methods and Sources of Study)	Serious	Formal	Grave register emerges past truncation boundary
అందరికీ నమస్కారం అండి (Greetings to all)	Inspiring	Optimistic	Aspirational moral framing is in truncated portion
జలియన్ వాలాబాగ్: ఉమర్ అలీషా రచనలో చారిత్రక విశ్లేషణ (Jallianwala Bagh: Historical Analysis in Umar Alisha's Writing)	Authoritative	Formal	Expert voice builds progressively beyond context window

Table 5: Representative misclassification examples (17/47, 11/37, and 12/30 instances respectively).

6 Conclusion

We present a Telugu communicative style classification system for the DravidianLangTech @ ACL 2026 shared task, achieving a Macro F1 of 0.2993, with ablations showing that label smoothing and backbone quality interact such that stronger Indic pretraining amplifies calibration benefits while weaker backbones are destabilised by soft targets, linguistic feature augmentation via MLP fusion consistently underperforms the plain encoder due to overfitting on small datasets, and error analysis reveals the primary failure mode is discourse-level, where passage-final stylistic signals distinguishing adjacent categories are lost to truncation, motivating future work on cross-attention feature fusion, ensemble methods, dataset expansion, and sliding-window attention for long-document truncation.

Limitations

The 3,000-sample training set limits the model's ability to learn fine-grained stylistic distinctions, particularly for categories with subtle register differences such as Formal, Serious, and Authoritative. Expanding max_length from 256 to 512 yields a modest improvement (Macro F1: 0.2993 → 0.3031) but does not resolve the low performance on Serious (F1 = 0.0769), suggesting

that truncation alone is not the primary bottleneck and that discourse-level modelling beyond token-window extension is needed. Our study is restricted to BERT-style encoder models; decoder-only or longer-context architectures remain unexplored due to computational constraints. Finally, we report point estimates without bootstrapped confidence intervals or significance tests; with 301 test samples, such intervals would be a valuable addition in future work.

Ethical Considerations

We have done our research, maintaining all ethical standards in a responsible manner. Our work classifies pre-existing, synthetically generated Telugu text into predefined style categories. The dataset contains no personally identifiable information, and our models do not generate new content. Through our work, we intend to support the development of technologies that elevate the understanding of low-resource languages.

Acknowledgements

The authors used GitHub Copilot for boilerplate code completion during implementation. All scientific contributions, analysis, and writing are the authors' own.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Bharathi Raja Chakravarthi and 1 others. 2026. Overview of the shared task on prompt recovery for LLM in Telugu. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech)*. Association for Computational Linguistics.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Sumanth Doddapaneni, Gowtham Kohli, Aman Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. IndicBERTv2: Multi-stage pre-training for Indic NLP. *arXiv preprint arXiv:2212.05236*.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA: BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*.

Raviraj Joshi, Purva Nankambakkam, Devang Limaye, Pradnya Thakkar, and Chinmay Kanchan. 2022. L3Cube-HindBERT and DevBERT: Fine-tuned BERT language models for Hindi and Devanagari. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, N. Chandra Shekhar Gokul, Avik Iyer, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Bharadwaj, and Partha Talukdar. 2021. MuRIL: Multilingual representations for Indian languages. *arXiv preprint arXiv:2103.10730*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32.

Mihaela Nitu, Cornelia Caragea, and Doina Caragea. 2024. Authorship attribution using stylometric features and transformer models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Sander Schulhoff, Michael Ilie, Nishant Bhatt, and 1 others. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35.

Haining Zhao, Oren Halvani, Lukas Miculivicius, and Benno Stein. 2021. Linguistic features meet transformer encoders for authorship attribution. *arXiv preprint arXiv:2111.14538*.