ByteBreaker@DravidianLangTech 2026: XLM-RoBERTa Large with Sliding-Window Chunking and Top-K Mean Pooling for Writing Style Classification

Chava Srinivasa Sai
Boston University
Massachusetts, USA
srinivassaichava@gmail.com

Rangoori Vinay Kumar
IIIT Dharwad
Karnataka, India
vinaykumarrangoori33@gmail.com

Jigeesha Sai Surapaneni
Dhanekula Institute of Engineering & Technology
Andhra Pradesh, India
jigeeshasai31@gmail.com

Chava Shanmukha Sai
IIITDM Kancheepuram
Tamil Nadu, India
shanmukhasaichava@gmail.com

Abstract

Identifying writing styles in long documents is challenging because stylistic signals are distributed unevenly and differ in subtle ways. We describe ByteBreaker, our system for the Prompt Recovery for LLM Shared Task at *DravidianLangTech@ACL-2026*, which classifies the writing style of LLM-rewritten Telugu documents into nine categories. To handle documents exceeding the 512-token transformer limit, we apply sliding-window chunking (stride = 256) and fine-tune XLM-RoBERTa Large on the style-rewritten text. At inference, Top-K Mean Pooling emphasises the most confident chunks. We train with five random seeds and submit three runs: a weighted ensemble (Run 1), a mean-guided single model (Run 2), and a Top-K-guided single model (Run 3). Run 3 achieves the best macro F1 of 0.3306; Run 1 achieves the best accuracy of 0.3256 with macro F1 of 0.3290. Our code is available on GitHub.

1 Introduction

Writing style is a critical factor in how a message is received. Beyond content, tone, vocabulary, sentence structure, and formality shape communication, and automatic style classification has become significant in NLP with the growing use of LLMs for text rewriting.

The Prompt Recovery task presents a novel variation: the goal is to recover the intended writing style from LLM-rewritten documents. Distinguishing closely related styles such as *Formal* and *Authoritative* remains challenging because their signals overlap, and many documents exceed the 512-token transformer limit, requiring segmented processing with effective evidence aggregation.

We propose ByteBreaker, which fine-tunes XLM-RoBERTa Large on overlapping 512-token chunks and applies Top-K Mean Pooling at inference to focus on the most stylistically informative segments. Multi-seed training and a weighted ensemble further improve robustness.

2 Related Work

Early style classification used handcrafted lexical and syntactic features (Argamon et al., 2007; Koppel et al., 2009; Sai et al., 2024) and CNN-based methods (Shrestha et al., 2017). Transformer models—BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020)— substantially advanced text classification via Hugging Face Transformers (Wolf et al., 2020). The 512-token limit has been addressed by hierarchical attention (Yang et al., 2016), sliding-window fine-tuning (Sun et al., 2019), and long-context models such as Longformer (Beltagy et al., 2020) and BigBird (Zaheer et al., 2020). Ensemble and multi-seed strategies further boost robustness in multilingual shared tasks (Saumya et al., 2022; Fharook et al., 2022; Kavatagi et al., 2023; Chakravarthi et al., 2020; Hande et al., 2020; Biradar et al., 2021; Premjith et al., 2026).

3 Task and Dataset

3.1 Task Description

The Prompt Recovery for LLM Shared Task at *DravidianLangTech@ACL 2026* (B et al., 2026) requires assigning one of nine style labels—*Authoritative*, *Formal*, *Humorous*, *Informal*, *Inspiring*, *Optimistic*, *Persuasive*, *Pessimistic*, *Serious*—to each LLM-rewritten Telugu document. Performance is evaluated by macro-averaged F1, with

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 158–162

accuracy secondary.

3.2 Dataset

The dataset (B et al., 2026) contains 3,000 training, 300 validation, and 301 test examples. Each entry provides the original transcript and its LLM rewrite in the "CHANGE STYLE" column, which we use exclusively for training as it contains stronger stylistic signals than the original. The test set is slightly imbalanced (Table 1): Serious is most frequent (15.61%) and Pessimistic least (7.31%).

Style	Support	% of Test
Authoritative	30	9.97
Formal	38	12.62
Humorous	31	10.30
Informal	35	11.63
Inspiring	37	12.29
Optimistic	33	10.96
Persuasive	28	9.30
Pessimistic	22	7.31
Serious	47	15.61
Total	**301**	**100**

Table 1: Label distribution in the test set ($N = 301$).

4 System Design and Methodology

Figure 1 illustrates the end-to-end ByteBreaker pipeline.

4.1 Model Architecture

We fine-tune XLM-RoBERTa Large[1] (Wolf et al., 2020)—a 24-layer transformer with 1024-dim hidden states and ≈560M parameters—with a linear classification head mapping the `[CLS]` token to 9 logits. The Large variant captures subtler inter-style distinctions than the base model.

4.2 Sliding-Window Chunking

We split each document into overlapping 512-token chunks (stride 256, 50% overlap) using `return_overflowing_tokens=True`. Each chunk inherits its parent's label and trains independently; the overlap ensures boundary tokens appear in two consecutive chunks, preserving local context.

4.3 Top-K Mean Pooling

At inference, a document's N chunk logits in $\mathbb{R}^9$ are aggregated by selecting the K highest-confidence chunks and averaging:

$$S_d = \text{top-}K \text{ chunks by } \max_c L_{i,c},$$
$$\hat{y}_d = \arg\max_j \frac{1}{|S_d|} \sum_{i \in S_d} L_{i,j}. \quad (1)$$

We set $K = \min(3, N)$; this consistently outperformed mean and LogSumExp pooling on the validation set by focusing on the most stylistically informative chunks.

4.4 Training Configuration

Table 2 lists hyperparameters. Model selection uses document-level macro F1 via a custom Top-K callback; label smoothing (ε=0.10) and early stopping (patience = 2) reduce overconfidence and overfitting.

Hyperparameter	Value
Base model	xlm-roberta-large
Max sequence length	512 tokens
Chunk stride	256 tokens
Learning rate	1×10^{-5}
Effective batch size	16 (2×8 grad. accum.)
Max epochs	12
Warmup ratio	0.10
Weight decay	0.01
Gradient clipping	1.0
Label smoothing (ε)	0.10
Best model metric	Doc-level macro F1
Early stopping patience	2 epochs
Mixed precision	BF16
Seeds	{42, 7, 123, 2025, 99}

Table 2: Training hyperparameters.

4.5 Multi-Seed Training, Ensemble, and Runs

We train across five seeds $\{42, 7, 123, 2025, 99\}$, exporting the best validation macro-F1 checkpoint per seed. Run 1 ensembles four checkpoints—best multi-seed (A), single-seed Top-K-guided (B), `ckpt-4062` Top-K-guided (C; Run 3), and `ckpt-2022` mean-guided (D; Run 2)—combining logits as $L_{\text{ens}} = \sum_m w_m L_m$ with weights from two-stage grid search (coarse step 0.10, fine step 0.02). All three runs use Top-K Mean Pooling at inference (Table 3).

Run	Model	Note
Run 1	Weighted ensemble (A–D)	Best accuracy
Run 2	Model D (mean-guided)	Single model
Run 3	Model C (top-K-guided)	Best macro F1

Table 3: Submission run configurations.

[1] https://huggingface.co/FacebookAI/xlm-roberta-large

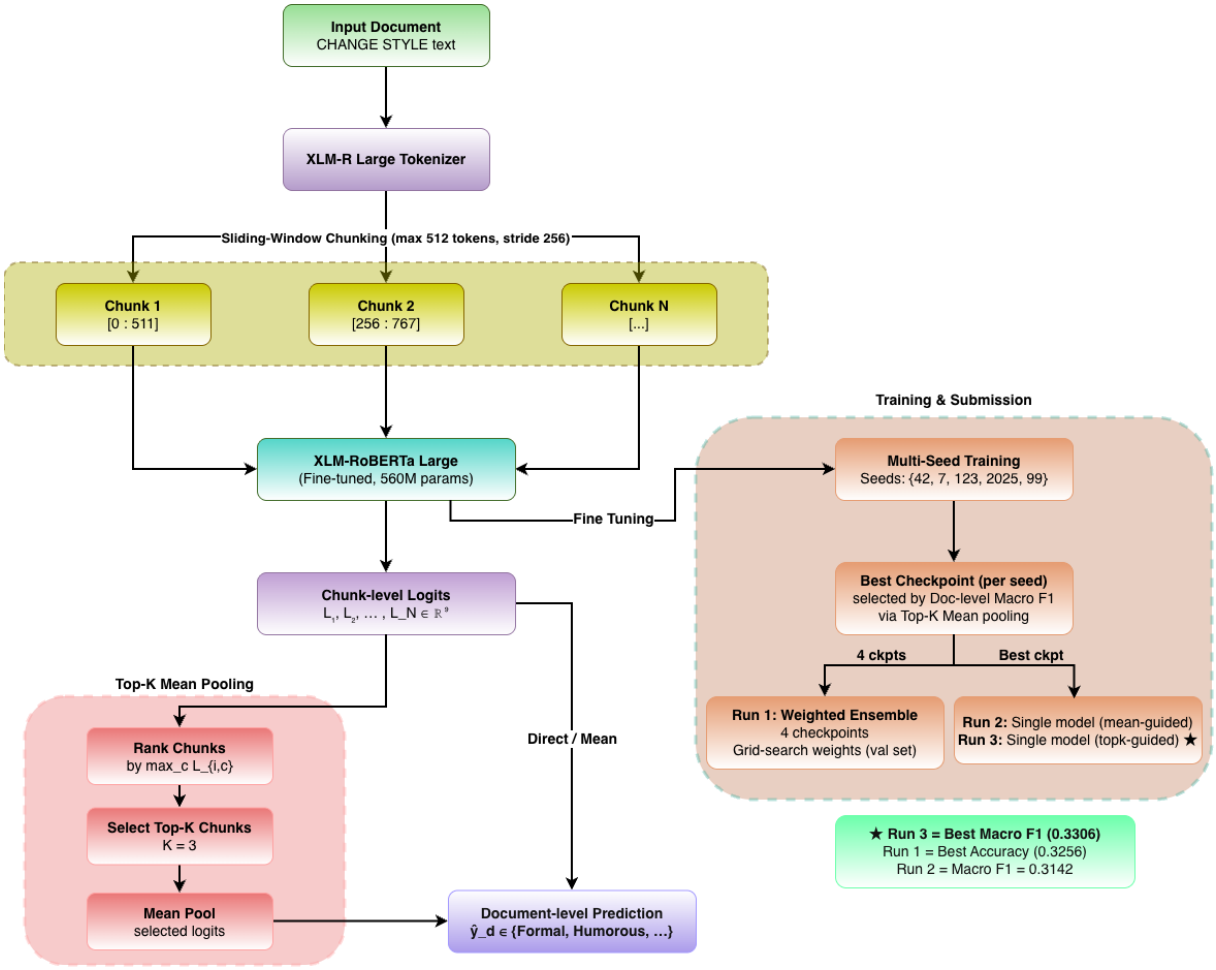

Figure 1: End-to-end pipeline of the ByteBreaker system. Documents exceeding 512 tokens are split into overlapping chunks (stride = 256), encoded by fine-tuned XLM-RoBERTa Large, and aggregated via Top-K Mean Pooling into a document-level prediction. Five seeded models feed either a weighted ensemble (Run 1) or single-model submissions (Runs 2–3).

5 Results and Discussion

Tables 4 and 5 show official results. Run 3 achieves the best macro F1 (**0.3306**); Run 1 the best accuracy (**0.3256**), as accuracy favours Serious (N=47) whereas macro F1 weights all classes equally.

Run	Acc.	P	R	F1
Run 1 (Ensemble)	**0.3256**	0.3670	0.3400	0.3290
Run 2 (mean-guided)	0.3189	0.3572	0.3342	0.3142
Run 3 (topk-guided)†	0.3123	0.3524	0.3284	**0.3306**

Table 4: Official results (macro-averaged). † Best macro F1.

Easiest classes. Optimistic and Humorous reach F1 ≈0.39–0.47 owing to distinctive lexical patterns in the LLM rewrites. **Hardest class.** Serious is the weakest (F1 ≈0.10–0.14), sharing a formal register with Authoritative and Formal. **Run 3 vs. ensemble.** Model D in the ensemble dilutes Pessimistic precision from 0.50 to 0.30; a Top-K-aligned checkpoint beats a heterogeneous ensemble on minority classes.

6 Conclusion

We introduced ByteBreaker for the Prompt Recovery task at *DravidianLangTech@ACL 2026*, combining sliding-window chunking with Top-K Mean Pooling for long-document style classification. Run 3 achieves the best macro F1 of 0.3306; Run 1 the best accuracy of 0.3256 (macro F1: 0.3290). Aligning the training criterion with the inference aggregation strategy is crucial, particularly for minority classes.

Limitations

ByteBreaker achieves moderate overall performance (macro F1 ≈0.33); nine closely related styles are difficult to separate from LLM-rewritten

		Run 1			Run 2			Run 3		
Style	**Supp.**	P	R	F1	P	R	F1	P	R	F1
Authoritative	30	0.33	0.20	0.25	0.28	0.17	0.21	0.18	0.23	0.20
Formal	38	0.28	0.24	0.26	0.32	0.24	0.27	0.28	0.26	0.27
Humorous	31	0.60	0.29	0.39	0.56	0.29	0.38	0.50	0.32	0.39
Informal	35	0.26	0.60	0.37	0.25	0.66	0.36	0.48	0.31	0.38
Inspiring	37	0.53	0.22	0.31	0.55	0.16	0.25	0.44	0.30	0.35
Optimistic	33	0.44	0.52	0.47	0.39	0.52	0.44	0.41	0.55	0.47
Persuasive	28	0.41	0.46	0.43	0.37	0.46	0.41	0.29	0.46	0.36
Pessimistic	22	0.30	0.41	0.35	0.33	0.41	0.37	**0.50**	0.41	**0.45**
Serious	47	0.15	0.13	0.14	0.17	0.11	0.13	0.10	0.11	0.10
Macro	301	0.37	0.34	0.33	0.36	0.33	0.31	0.35	0.33	0.33

Table 5: Per-class P, R, and F1 for all three runs.

text alone, with Serious consistently confused with Authoritative and Formal. The $K=3$ and stride = 256 choices were validated on the provided split and may not generalise without re-tuning; comparisons with long-context models such as Longformer and BigBird remain future work.

Ethical Considerations

This work uses data from the shared task organisers and public pre-trained models; no personal data is collected. Writing style classification risks misuse in automated content profiling; we encourage responsible deployment and provide per-class breakdowns to expose model limitations.

Acknowledgements

We thank the DravidianLangTech organizers for providing the dataset and organizing the shared task. We also acknowledge the creators of XLM-RoBERTa and other open-source resources that made this work possible. No generative AI tools were used for research, experiments, or content generation in this manuscript.

References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Premjith B, Jyothish Lal G, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Durairaj Thenmozhi, Ratnavel Rajalakshmi 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.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. mBERT based model for identification of offensive content in south indian languages. In *Working Notes of FIRE 2021 – Forum for Information Retrieval Evaluation*.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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.

Shaik Fharook, Syed Sufyan Ahmed, Gurram Rithika, Sumith Sai Budde, Sunil Saumya, and Shankar Biradar. 2022. Are you a hero or a villain? A semantic role labelling approach for detecting harmful memes. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 19–23.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada codemixed dataset for sentiment analysis and offensive language

detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Sanjana M Kavatagi, Rashmi R Rachh, and Shankar S Biradar. 2023. VTUBGM@LT-EDI-2023: Hope speech identification using layered differential training of ULMFiT. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 209–213.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

B Premjith, Bharathi Raja, and Prasanna Kumar Kumaresan. 2026. Findings of the shared task on prompt recovery for LLM in Telugu – DravidianLangTech@ACL 2026. In *Proceedings of the Sixth Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech-2026)*, San Diego, California, USA. Association for Computational Linguistics. To appear.

Chava Sai, Rangoori Kumar, Sunil Saumya, and Shankar Biradar. 2024. IIITDWD_SVC@DravidianLangTech-2024: Breaking language barriers; hate speech detection in Telugu-English code-mixed text. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 119–123. Association for Computational Linguistics.

Sunil Saumya, Vanshita Jha, and Shankar Biradar. 2022. Sentiment and homophobia detection on YouTube using ensemble machine learning techniques. In *Working Notes of FIRE 2022 – Forum for Information Retrieval Evaluation (Hybrid)*. CEUR.

Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 669–674.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? In *Proceedings of the 18th Chinese National Conference on Computational Linguistics (CCL)*, pages 194–206.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.