IndicBERTv2-MLM-only for Fine-Grained Misinformation Analysis in South Indian Languages

N.V. Gokul[1], J. JeswinJoel[1], S. Gautham[1] and J. Rajeswari[1]

[1]*Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India*

Abstract

This work addresses the challenge of fine-grained misinformation detection in Large Language Model (LLM) generated summaries for Indian languages. We focus on Subtask 2: Detect Misinformation in LLM Output, which requires classifying LLM-generated summaries into specific categories of factual incorrectness. The task is particularly difficult for regional languages like Tamil, Telugu, Malayalam, and Kannada due to limited annotated resources. To tackle this, we fine-tune `IndicBERTv2-MLM-only`, a multilingual transformer model pre-trained on Indian languages. Our methodology uses article-summary pairs with stratified sampling and optimizes for macro-F1 score across four distinct classes: misrepresentation, fabrication, false attribution and incorrect quantities. We fine-tune separate `IndicBERTv2-MLM-only` models for each target language using identical architectures and hyperparameters. The models achieve competitive results given the complexity of the fine-grained classification task, with cross-lingual averages of 71.69% accuracy and 46.69% macro-F1, demonstrating better handling of multilingual complexity.

Keywords: Misinformation detection, Indian languages, Transformers, IndicBERTv2-MLM-only, Multilingual NLP.

1. Introduction

Information plays a crucial role in shaping public opinion, especially during sensitive periods such as elections, conflicts, and pandemics. When false or misleading content spreads unchecked, it can amplify misunderstandings, trigger public panic, and drive impulsive decisions that may escalate civil unrest or threaten national stability. Effectively identifying and managing misinformation is therefore essential, both to enable targeted remediation and to protect the credibility of news outlets and online platforms.

This work is part of the PROMID Shared Task at FIRE 2025, which focuses on Misinformation Detection and Prompt Recovery [1, 2, 3]. Specifically, we address Subtask 2: Misinformation Detection in LLM-generated text. This subtask targets the problem of fine-grained misinformation detection in summaries produced by large language models (LLMs). Given a news article and its LLM-generated summary, the goal is to classify each instance into one of four categories: misrepresentation, fabrication, false attribution, or incorrect quantities. We focus on four South Indian languages—Tamil, Telugu, Kannada, and Malayalam—where high-quality resources and tools for misinformation detection remain limited. The model receives the article–summary pair as input and learns to detect contextual mismatches, factual errors, and inconsistencies indicative of misinformation.

Previous approaches have largely relied on traditional machine learning models such as Support Vector Machines and Logistic Regression trained on English datasets, which do not generalize well to multilingual or low-resource settings. More recent multilingual transformer-based models (e.g., BERT variants) [4] typically frame misinformation detection as a binary task (misinformation vs. non-misinformation) [5] and often depend on class-weighting schemes, but they rarely consider the combined effect of original articles and LLM-generated summaries, nor do they adequately cover Indian regional languages.

The key gap we address is the lack of multilingual support for South Indian (Dravidian) languages in fine-grained misinformation detection. In particular, existing systems rarely (i) target regional

Forum for Information Retrieval Evaluation, December 17-20, 2025, India

gokul2470019@ssn.edu.in (N.V. Gokul); jeswinjoel2470029@ssn.edu.in (J. JeswinJoel); gautham2470021@ssn.edu.in (S. Gautham); rajeswarij@ssn.edu.in (J. Rajeswari)

languages such as Tamil, Telugu, Kannada, and Malayalam, and (ii) jointly use both the news article and its LLM-generated summary for multi-class classification across distinct misinformation categories. To bridge this gap, we fine-tune IndicBERTv2-MLM-only [6], a transformer model pre-trained on multiple Indian languages, for four-way misinformation classification. The four misinformation categories are label-encoded for multi-class prediction. To handle class imbalance, we use stratified splits and evaluate primarily with macro-averaged F1, which gives equal importance to all classes. We implement this approach through language-specific fine-tuning, creating separate `IndicBERTv2` model instances for each target language.

Our experimental results show that the proposed language-specific approach achieves promising performance across the four languages: Kannada (F1: 52.75%, Accuracy: 75%), Telugu (F1: 48.00%, Accuracy: 72.25%), Malayalam (F1: 47.00%, Accuracy: 71.50%), and Tamil (F1: 39.00%, Accuracy: 68.00%). On average, the models attain 71.69% accuracy and 46.69% macro-F1 across languages, demonstrating the effectiveness of our language-isolated fine-tuning strategy in handling multilingual complexity and LLM-induced summarization errors in low-resource South Indian language settings.

2. State of the Art

Raja et al. introduce one of the first systematic frameworks for fake news detection in Dravidian languages using transfer learning, fine-tuning multilingual encoders such as mBERT and XLM-R on Tamil, Malayalam, and Kannada news datasets [7]. Their work clearly demonstrates that transformer-based models substantially outperform classical baselines and confirms the value of pre-trained multilingual representations for low-resource Indian languages. However, the task formulation is binary (fake vs. real) and each instance is treated as a single news headline or article. The models do not consider the interaction between an original article and a derived text (e.g., a summary), nor do they distinguish between different types of factual errors such as misrepresentation or incorrect quantities. As a result, they cannot explain *how* a piece of content is wrong, only that it is likely fake.

Subsequent shared-task systems at DravidianLangTech extend this line of work by exploring a broader set of architectures, including monolingual BERT variants (e.g., Malayalam-BERT) [8] and ensembles that combine multiple multilingual transformers in shared-task systems at DravidianLangTech [9] and team reports such as CIC-NLP [10]. These systems achieve strong accuracy on fake-news benchmarks, but they inherit similar conceptual limitations: they focus on binary or coarse multi-class labels, and their inputs are almost always standalone texts (tweets, posts, headlines, or full articles). As a consequence, they are not exposed to paired inputs where one text (an LLM-generated summary) may selectively distort another (the source article), and they are not designed to detect hallucinations or summarisation-specific distortions introduced by LLMs.

Most misinformation detection systems are built for English or high-resource languages, often using machine translation for others [11], which introduces noise and fails to capture language-specific phenomena in South Indian languages [12]. These systems also typically treat misinformation as binary, lacking fine-grained categories. In contrast, we target four South Indian languages—Tamil, Telugu, Kannada, and Malayalam—using article-summary pairs to detect inconsistencies in LLM-generated summaries. We fine-tune separate `IndicBERTv2-MLM-only` models for each language with identical training protocols, performing four-way classification across `misrepresentation`, `fabrication`, `false attribution`, and `incorrect quantities`, evaluated with macro-F1 for consistent multilingual comparison.

3. Methodology

3.1. Task Objective

Given a piece of LLM-generated text containing misinformation, the objective is to categorize each datapoint into one of four specific categories—Misrepresentation, Fabrication, False Attribution, and

Incorrect Quantities—based on the nature and type of factual incorrectness present in the summaries. This task addresses the growing challenge of detecting nuanced factual errors in LLM outputs, where traditional binary classification fails to capture the diverse ways large language models can introduce inaccuracies.

The PROMID dataset reflects the current scenario in LLM-generated misinformation research, where models increasingly produce plausible but factually flawed content across low-resource South Indian languages (Tamil, Telugu, Kannada, Malayalam). With 989 samples per language showing moderate class imbalance, the dataset captures real-world distribution patterns while enabling fine-grained analysis of different error types prevalent in LLM summaries

3.2. Dataset

The dataset for Subtask 2 on misinformation detection in LLM-generated text [13], provided by the task organizers, contains 989 paired samples for each of the four South Indian languages—Tamil, Telugu, Kannada, and Malayalam—where each sample pairs a large language model-generated summary with its corresponding original article, meticulously annotated into four specific misinformation categories: Misrepresentation, Fabrication, False Attribution, and Incorrect Quantities.Table 1 presents the distribution of samples across these categories.

Misinformation Category	Samples	Percentage
Misrepresentation	294	29.7%
Fabrication	250	25.3%
False Attribution	250	25.3%
Incorrect Quantities	195	19.7%
Total	**989**	**100%**

Table 1
Distribution of samples across misinformation categories in the PROMID dataset (consistent across all four South Indian languages)

3.3. Data Preprocessing

Prior to model training, the dataset was loaded and preprocessed to ensure consistency and reliability. Column names were standardized, and samples with missing summaries or misinformation labels were removed. The incorrect summaries and their corresponding misinformation categories were retained for further analysis. Each category—Misrepresentation, False Attribution, Incorrect Quantities, and Fabrication—was encoded into numerical labels using a fixed and reproducible label mapping.

3.4. Model Architecture

For Subtask 2: Misinformation Detection in LLM-Generated Text, our approach fine-tunes separate `IndicBERTv2-MLM-only` models for each target language (Tamil, Telugu, Kannada, Malayalam) using the HuggingFace Transformers library. Each model follows the standard BERT-base configuration with 12 transformer layers, 12 attention heads, and 768-dimensional embeddings. For classification, we use the `[CLS]` token representation passed through a linear projection layer to produce four output units corresponding to the misinformation categories. Input text is formatted as "Article: {*text*} Summary: {*text*}", tokenized with a maximum length of 384 tokens. We implement an 80/20 stratified train-validation split to preserve class distribution and use macro-F1 scoring for model selection to address class imbalance. The training hyperparameters are shown in Table 2.

Table 2
Training Hyperparameters

Parameter	Value
Batch size	8
Learning rate	1×10^{-5} with 10% warmup
Epochs	8 with early stopping based on validation F1
Optimizer	AdamW (weight decay: 0.01)
Checkpoints	Maximum 2 retained
Evaluation metric	Macro-F1 at each epoch
Precision	FP16 when available
Seed	Fixed at 42

3.5. Training Pipeline Architecture

Figure 1 illustrates our language-specific fine-tuning methodology. We initialize from the multilingual IndicBERTv2-MLM-only base model and create four parallel fine-tuning streams—one for each target language (Tamil, Telugu, Malayalam, and Kannada). Each stream processes its respective language dataset independently, producing language-specific models that generate predictions following a standardized 4-category classification schema. All outputs are aggregated and evaluated using macro-F1 scoring, ensuring balanced performance across both languages and misinformation categories.

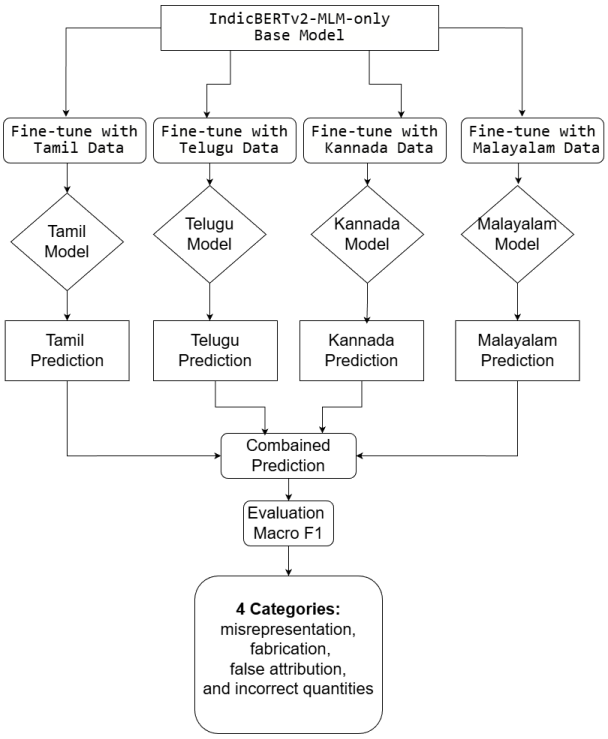

Figure 1: Language-specific fine-tuning pipeline for misinformation detection. The IndicBERTv2 base model is fine-tuned separately on each language dataset, producing four independent models that generate predictions in a common 4-category format. All outputs are evaluated with macro-F1 optimization to handle class imbalance.

4. Results

The IndicBERTv2-MLM-only model shows consistent convergence with training loss dropping steadily from $\sim$1.39 (Epoch 1) to < 0.80 (Epoch 8). Validation F1 peaks mid-training (Epochs 4-6) before stabilizing.

4.1. Training Progression Summary

Table 3 summarizes the best validation accuracy and macro F1 scores achieved across the four Dravidian languages during the 8-epoch training process.

Language	Best Val. Accuracy	Best Val. F1 (Macro)
Tamil	0.5455 (Epoch 5)	**0.5186** (Epoch 5)
Telugu	0.4545 (Epoch 4)	0.4204 (Epoch 5)
Malayalam	**0.5303** (Epoch 6)	0.5070 (Epoch 7)
Kannada	0.5202 (Epoch 5)	0.5101 (Epoch 5)

Table 3
Best validation performance across languages (8 epochs)

4.2. Class-wise Performance Analysis (Kannada)

Table 4 reports detailed per-class and overall metrics on the Kannada validation set. The model performs best on *misrepresentation*, achieving the highest F1-score, while minority classes such as *fabrication* and *incorrect_quantities* show high precision but relatively low recall, indicating that the model is conservative when predicting these categories and misses a portion of true positives.

Category	Precision	Recall	F1	Accuracy	Support	TP	FP	FN
fabrication	0.89	0.25	0.39	0.56	32	8	1	24
false_attribution	0.67	0.62	0.64	0.84	13	8	4	5
incorrect_quantities	1.00	0.20	0.33	0.86	10	2	0	8
misrepresentation	0.65	0.88	0.75	0.74	25	22	12	3
Overall	**0.80**	**0.49**	**0.53**	**0.75**	80	40	17	40

Table 4
Kannada validation set: detailed class-wise and overall metrics

Overall performance (macro-averaged) shows precision of 0.80, recall of 0.49, F1-score of 0.53, and accuracy of 0.75, reflecting a model that is generally precise in its predictions but still tends to under-recall certain under-represented classes.

4.3. Class-wise Performance Analysis (Malayalam)

Table 5 presents detailed per-class and overall metrics for the Malayalam validation set. The model performs best on the *misrepresentation* category, while minority classes such as *fabrication* and *incorrect_quantities* show lower recall, indicating that many true instances of these categories are still missed.

Category	Precision	Recall	F1	Accuracy	Support	TP	FP	FN
fabrication	0.78	0.22	0.34	0.53	32	7	2	25
false_attribution	0.57	0.31	0.40	0.79	13	4	3	9
incorrect_quantities	0.75	0.30	0.43	0.86	10	3	1	7
misrepresentation	0.59	0.88	0.71	0.68	25	22	15	3
Overall	**0.67**	**0.43**	**0.47**	**0.72**	80	36	21	44

Table 5
Malayalam validation set: detailed class-wise and overall metrics

Overall, the model achieves a macro-averaged precision of 0.6725, recall of 0.4275, F1-score of 0.47,

and accuracy of 0.715. These results show that the classifier is relatively precise in its predictions but still under-recognizes several true positive instances, especially for the fabrication class.

4.4. Class-wise Performance Analysis (Tamil)

Table 6 presents detailed per-class and overall metrics for the Tamil validation set. The model performs best on the *misrepresentation* class, whereas minority classes such as *fabrication*, *false_attribution*, and *incorrect_quantities* show lower recall, indicating that many true instances of these categories are still missed.

Category	Precision	Recall	F1	Accuracy	Support	TP	FP	FN
fabrication	0.78	0.22	0.34	0.53	32	7	2	25
false_attribution	0.43	0.23	0.30	0.75	13	3	4	10
incorrect_quantities	0.29	0.20	0.24	0.77	10	2	5	8
misrepresentation	0.59	0.80	0.68	0.67	25	20	14	5
Overall	**0.52**	**0.36**	**0.39**	**0.68**	80	32	25	48

Table 6
Tamil validation set: detailed class-wise and overall metrics

Overall, the model achieves a macro-averaged precision of 0.5225, recall of 0.3625, F1-score of 0.39, and accuracy of 0.68. These results suggest that while predictions are moderately precise, the model still under-detects several true positives, particularly in the less frequent misinformation categories.

4.5. Class-wise Performance Analysis (Telugu)

Table 7 reports detailed per-class and overall metrics for the Telugu validation set. As with other languages, the model performs best on the *misrepresentation* category, while minority classes—particularly *fabrication*—show reduced recall, indicating that a notable portion of true positives remains undetected.

Category	Precision	Recall	F1	Accuracy	Support	TP	FP	FN
fabrication	0.82	0.28	0.42	0.56	32	9	2	23
false_attribution	0.44	0.31	0.36	0.75	13	4	5	9
incorrect_quantities	0.60	0.30	0.40	0.84	10	3	2	7
misrepresentation	0.66	0.84	0.74	0.74	25	21	11	4
Overall	**0.63**	**0.43**	**0.48**	**0.72**	80	37	20	43

Table 7
Telugu validation set: detailed class-wise and overall metrics

Overall, the Telugu model attains a macro-averaged precision of 0.63, recall of 0.4325, F1-score of 0.48, and accuracy of 0.7225. This indicates reasonably precise predictions, but with room for improvement in recall, especially for fabrication and incorrect_quantities, where many true instances are still missed.

5. Conclusion

This research presents a comprehensive framework for fine-grained misinformation detection in South Indian languages. Our approach leverages IndicBERTv2, a transformer model pre-trained on major Indian languages, adapting it through transfer learning to distinguish between four distinct categories of misinformation: fabrication,false attribution,incorrect quantities,misrepresentation.

We address several critical challenges in multilingual misinformation analysis. First, we implement a training strategy with macro-F1 optimization to mitigate class imbalance, ensuring balanced performance across all categories regardless of their frequency in the dataset. Second, we establish an

evaluation framework that prioritizes macro-F1 scores alongside accuracy, providing a more comprehensive assessment of model performance. Third, we design a practical two-stage inference pipeline that filters pre-verified content to optimize computational efficiency during deployment, making the system suitable for real-world applications.

The technical implementation involves fine-tuning for eight epochs with a learning rate of 1×10^{-5}, AdamW optimization with weight decay, linear warmup scheduling, and automatic selection of the best model checkpoint based on validation performance. Our methodology demonstrates that transfer learning from multilingual pre-trained models effectively addresses the linguistic diversity and data scarcity challenges inherent in South Indian language processing.

This work contributes to the field of computational linguistics for social good by providing a scalable, nuanced approach to misinformation detection that moves beyond binary classification. The system offers practical utility for content moderation, fact-checking organizations, and platform operators working with South Indian language content, while establishing methodological foundations for future research in low-resource multilingual NLP applications.

Declaration on Generative AI

During the preparation of this work, we used OpenAI's ChatGPT model for language fluency improvement and technical documentation assistance. All experimental work, data analysis, results, and scientific conclusions are our own. We have reviewed and refined all AI-assisted content and take full responsibility for the published work.

References

[1] A. Hegde, G. K. Shahi, S. Satapara, P. Mehta, S. Modha, D. Ganguly, D. Nandini, H. L. Shasirekha, A. K. Jaiswal, G. Pasi, T. Mandl, Prompt recovery for misinformation detection at fire 2025, in: Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '25, Association for Computing Machinery, 2025.

[2] G. K. Shahi, A. Hegde, S. Satapara, P. Mehta, S. Modha, D. Ganguly, D. Nandini, H. L. Shasirekha, A. K. Jaiswal, G. Pasi, T. Mandl, Overview of the first shared task on prompt recovery for misinformation detection (promid 2025), in: K. Ghosh, T. Mandl, S. Pal, S. Majumdar, A. Chakraborty (Eds.), Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, Varanasi, India. December 17-20, 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[3] S. Satapara, P. Mehta, S. Modha, D. Ganguly, Key takeaways from the second shared task on indian language summarization (ILSUM 2023), in: K. Ghosh, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation (FIRE-WN 2023), Goa, India, December 15-18, 2023, volume 3681 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 724–733. URL: https://ceur-ws.org/Vol-3681/T8-1.pdf.

[4] A. Bala, P. Krishnamurthy, Abhipaw@dravidianlangtech: Fake news detection in dravidian languages using multilingual bert, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech 2023), INCOMA Ltd., 2023, pp. 235–238. URL: https://aclanthology.org/2023.dravidianlangtech-1.34/.

[5] N. Subramanian, B. R. Chakravarthi, et al., Overview of the shared task on fake news detection in dravidian languages — dravidianlangtech@naacl 2025, in: Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech@NAACL 2025), Association for Computational Linguistics, 2025, pp. 759–767. URL: https://aclanthology.org/2025.dravidianlangtech-1.128/.

[6] D. Kakwani, A. Kunchukuttan, S. Golla, N. Nivash, M. Pinnis, R. Kaur, M. M. Khapra, Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020.

[7] E. Raja, B. Soni, S. Bhat, Fake news detection in dravidian languages using transfer learning, Engineering Applications of Artificial Intelligence 121 (2023) 106877.
[8] F. Chakraborty, Others, Malayalam-bert based transformer models for fake news detection, in: Proceedings of DravidianLangTech 2025, Association for Computational Linguistics, 2025.
[9] F. Shanmugavadivel, Others, Overview of the shared task on fake news detection in dravidian languages, in: Proceedings of the Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech 2025), Association for Computational Linguistics, Turin, Italy, 2025.
[10] CIC-NLP, Cic-nlp@dravidianlangtech 2025: Fake news detection in dravidian languages using multilingual transformers, in: Proceedings of DravidianLangTech 2025, Association for Computational Linguistics, 2025.
[11] X. Wang, Others, Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey, arXiv preprint arXiv:2410.18390 (2024).
[12] A. Goyal, A. Basu, D. Sharma, Challenges in neural machine translation for dravidian languages: Morphology, script variation, and data noise, in: Proceedings of the Workshop on Indian Language Data: Resources and Evaluation (WILDRE), Association for Computational Linguistics, European Language Resources Association (ELRA), Dublin, Ireland, 2022. URL: https://aclanthology.org/2022.wildre-1.8.
[13] S. Satapara, P. Mehta, D. Ganguly, S. Modha, Fighting fire with fire: Adversarial prompting to generate a misinformation detection dataset, CoRR abs/2401.04481 (2024). URL: https://doi.org/10.48550/arXiv.2401.04481. doi:10.48550/ARXIV.2401.04481. arXiv:2401.04481.