Misinformation Detection in Multilingual Social Media Texts Using LLM-Based Translation, Augmentation, and DeBERTa Fine-Tuning

Avinash Trivedi[1,*], Chindukuri Mallikarjuna[1]

[1] SRM University-AP, Amaravati, Andhra Pradesh 522240, India

Abstract

Misinformation circulating on social platforms often distorts public understanding and can escalate real-world consequences, particularly in volatile geopolitical contexts. In this study, we describe the system constructed for subtask 3 of the PROMID 2025 shared task, which focuses on identifying misleading content within tweets pertaining to the Russo–Ukrainian conflict. The submission from our team (*Sarang*) secured 2nd place, supported by strong evaluation metrics: a precision of 0.90, a recall of 0.91, and a weighted F1-score of 0.90. Because the dataset contained instances in multiple languages, all non-English posts were rendered into English through a large language model. To mitigate skewed label distributions, we introduced synthetic variants of the minority class, thereby easing class imbalance. The classification pipeline relied on a deberta-v3-small encoder, which proved adept at capturing fine-grained semantic distinctions. The resulting performance underscores the reliability of the proposed approach and provides a competitive reference point for forthcoming work. Overall, the study offers practical insights for advancing misinformation detection in future shared-task settings.

Keywords

Misinformation, LLM, Deberta, PROMID@FIRE-2025

1. Introduction

Social media and online platforms enable information to circulate globally within seconds, a strength that also accelerates the proliferation of misinformation, disinformation, and fake news. False or misleading content can distort public perception, undermine trust in institutions, and trigger social or political instability, particularly during crises or conflicts. Automated detection of such content has therefore become a critically important problem in Natural Language Processing (NLP) and machine learning.

In recent years, deep-learning and transformer-based models have emerged as especially promising for fake-news and misinformation detection. Compared to traditional machine-learning approaches, these models are able to learn richer semantic, syntactic, and contextual patterns, making them more effective at distinguishing truthful from deceptive content[1]. Review studies also confirm the growing trend: comprehensive surveys of deep-learning-based fake-news detection outline how supervised, semi-supervised, and unsupervised methods often combining content, social context, and external knowledge have been explored to improve robustness and generalization[2]. Despite these advances, significant challenges remain, especially when dealing with real-world social media data. Posts are often noisy, informal, multilingual, and may suffer from severe class imbalance (e.g., far fewer misinformation posts than benign ones). Survey analyses highlight that many models, though effective on curated datasets, struggle when confronted with such realistic constraints [3].

Recognizing these gaps, the present study develops a fine-tuned transformer-based misinformation detection system tailored for social media text. Specifically, to handle language diversity, we translate non-English posts into English. To address class imbalance and enrich the minority misinformation class, we use synthetic data generation via an LLM-driven augmentation process. Finally, we train a

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

[*] Corresponding author

avinashtrivedi.2008@gmail.com (A. Trivedi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-02-24

supervised classifier on the balanced dataset, aiming for robust performance under realistic social-media conditions.

To provide a comprehensive view of our approach, the remainder of this paper is structured as follows: Section 2 reviews the relevant literature on misinformation detection in social media. Section 3 describes the dataset used in this study along with preprocessing steps. Section 4 outlines the proposed methodology, including translation-based normalization, LLM-driven augmentation, and model fine-tuning. Section 5 details the experimental setup and discusses the results, followed by Section 6, which concludes the paper and highlights potential directions for future research.

2. Related work

Deep learning has become the foundation for modern misinformation detection. A comprehensive analysis in [4] highlights the evolution from traditional supervised learning to advanced neural architectures, emphasizing the importance of adapting models to dynamic, crisis-related misinformation trends on social media platforms. The work also stresses challenges such as noisy text, limited context in short messages, and domain shifts driven by emerging events.

With the rapid rise of large language models (LLMs), misinformation detection is now entering a new phase. [5] provides an extensive overview of how LLM-based reasoning, semantic understanding, and prompt-driven inference contribute to improved detection accuracy. The paper also notes several emerging issues, including hallucination, trustworthiness, and vulnerability to user-altered misinformation.

To improve generalization across different online platforms and domains, transfer-learning-based misinformation detection has gained attention. [6] proposes adaptable transfer learning techniques that allow pretrained models to better handle topic drift, temporal variation, and diverse misinformation narratives, a key challenges in real-world deployment. A significant portion of misinformation spreads in multiple languages, making multilingual detection essential. [7] surveys progress across high and low resource languages and shows that performance remains heavily skewed toward English and a few major languages. To this end, [8] provides a comprehensive review of detection approaches specifically targeting low-resource linguistic environments, highlighting difficulties such as lack of annotated data and cultural context variations. To tackle this, [9] proposes unified LLM-based architectures that can operate effectively in multilingual or low-resource settings, demonstrating clear improvements compared to traditional fine-tuned transformers. However, the authors note that such models still require adaptation mechanisms for highly informal and noisy user-generated content.

Collectively, these findings reveal clear research gaps: existing models struggle to maintain strong recall under multilingual noise, scarcity of labeled misinformation data, and severe class imbalance. To address these challenges, our approach translates all non-English content into English to reduce cross-lingual variation, employs LLM-driven data augmentation to strengthen minority misinformation class representation, and fine-tunes a deberta-based classifier for robust performance under real-world crisis-driven social-media conditions.

3. Dataset

We obtained the dataset from subtask-3 of the PROMID 2025 shared task [10, 11], consisting of tweets related to the Russo-Ukrainian conflict. Each tweet is manually annotated as either misinformation or non-misinformation, enabling supervised learning for misinformation classification. This dataset [12] was collected using the AMUSED framework [13] and includes various fields such as text, retweet, hashtag and user related meta data such as user ID, user_screen_name, user_description,user_location etc. The overall dataset distribution is summarized in Table 1.

Class	Training Set Count
Misinfo	364
Nonmisinfo	34174
Total	**34538**

Table 1
Data distribution of development set

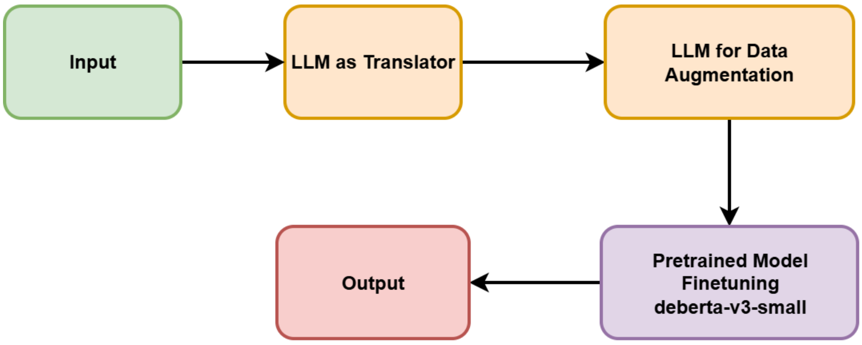
Input
LLM as Translator
LLM for Data Augmentation
Pretrained Model Finetuning deberta-v3-small
Output

Figure 1: Model architecture

4. Methodology

4.1. Zero-shot and few-shot prompting on LLMs

We initially evaluated the models in zero-shot and few-shot settings across several LLM families, including Gemma, Qwen, and Llama. In addition, we applied prompt-optimization techniques using the DSPy framework [14, 15] . Despite these efforts, the overall performance remained consistently low. Table 2 shows summary of the findings.

Model	Test Score		
	Precision	Recall	F1
gemma3:4b	0.73	0.55	0.61
gemma3:12b	**0.88**	**0.87**	**0.88**
qwen3:8b	0.82	0.75	0.77
llama3.1:8b	0.82	0.72	0.75

Table 2
Baseline score: Few-shot with prompt optimization

4.2. Best performing system

We ultimately opted to fine-tune a pretrained classification model. The dataset exhibited substantial class imbalance: one class contained only 364 samples, whereas the other exceeded 34k. Because the corpus spanned multiple languages, our initial experiments used mdeberta-v3-base, but its performance was unsatisfactory. This led us to evaluate its english-only counterpart. We selected a deberta-based architecture due to its strong results across a range of text-classification benchmarks. Consequently, all non-English instances were first translated into English using gemma-3-12b with the prompt shown in Figure 2. We then employed the same model for data augmentation (prompt in Figure 3), generating four synthetic variants for each of the 364 minority-class samples. To create a balanced training set for fine-tuning deberta-v3-small, we randomly drew an equal number of instances from the majority class. The architecture of our best-performing system is illustrated in Fig 1.

```
Prompt

{"role": "system",
"content": "You are a precise translation assistant."},
{"role": "user",
"content": """
Translate the following tweet into clear, natural English.
Keep the meaning and tone the same.
Do not add explanations, only output the translated English text.

Tweet : {text}
Output:
"""}
```

Figure 2: Prompt for tweet translation

```
Prompt

{"role": "system",
"content": "You are a multilingual social media writer."},
{"role": "user",
"content": """Rewrite the following tweet into n_variants new variations in the same language ("lang").
Preserve its meaning and intent.

1. If it's misinformation, keep the same misleading claim but change tone, punctuation, or slang
2. If it's non-misinformation, keep it factual but vary the style.
Do NOT translate into English.
Do not add explanations, only output the new variations.

Tweet (label): {text}
Output:
"""}
```

Figure 3: Prompt for data augmentation

Once the dataset was preprocessed, we utilized it to fine-tune the deberta model, a state-of-the-art transformer-based language model.

5. Experimental Results

Our pipeline, illustrated in Figure 1, integrates several components: an LLM-driven module for translating non-English (*1179 instances in the raw dataset*) tweets, a synthetic oversampling procedure to bolster the minority misinformation class, and a fine-tuning stage in which a deberta-v3-small model is trained on the rebalanced corpus. We construct a single balanced dataset by augmenting misinformation tweets and downsampling non-misinformation instances. The resulting dataset is split into training and validation sets using an 80:20 ratio, where the validation set is used exclusively for internal evaluation. No explicit filtering is applied to remove potentially malformed translations, under the assumption that any translation noise affects training, validation and test data uniformly. For both translation and data augmentation, we employ an LLM with deterministic decoding settings, fixing the temperature to 0 and top-p to 1. The training configuration followed the hyperparameter settings listed in Table 4, selected to promote stable optimization and reliable downstream performance.

To assess system efficacy, we employed the PROMID 2025 subtask-3 benchmark comprising 2,414 test instances and evaluated the model with precision, recall, and weighted F1. The official leaderboard scores are summarized in Table 3, the classifier reached 0.90 precision, 0.91 recall, and a weighted F1-score of 0.90. Taken together, these outcomes indicate that the combination of multilingual normalization and targeted minority-class augmentation provides measurable gains in robustness when handling misinformation embedded in noisy social-media text. Consolidated experimental results appear in Table 2 and Table 3, while Table 4 documents the hyperparameters used in the top-performing configuration.

Model	Test Score		
	Precision	Recall	F1
deberta-v3-small	0.90	0.91	0.90

Table 3
Best system submission

Hyperparameter	Value
max token length	512
per device train batch size	6
per device eval batch size	6
num train epochs	8
weight decay	0.01
warmup steps	500
optim	adamw_torch
learning rate	5e-05
lr scheduler type	linear
seed	42
load best model at end	True

Table 4
Hyperparameter values

6. Conclusions and Future Work

After evaluating multiple inference strategies, ranging from zero-shot and few-shot paradigms to several prompt-tuning configurations, we found that the *deberta-v3-small* architecture, once fine-tuned on our processed dataset, yielded the most favorable outcomes, particularly in terms of weighted performance metrics. The pipeline in which multilingual tweets were first translated and subsequently supplemented with synthetic samples to correct class imbalance substantially improved the model's ability to capture language patterns. This revised data regime produced performance gains that surpassed those observed in earlier experimental baselines.

Our findings highlight the central role of translation-driven preprocessing and data augmentation in strengthening model generalization and resilience. Future work will aim to assess the benefits of scaling to larger LLMs and examining additional deberta variants.

Declaration on Generative AI

During the preparation of this work, the author(s) used chatGPT-5.1 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] L. Hu, S. Wei, Z. Zhao, B. Wu, Deep learning for fake news detection: A comprehensive survey, AI open 3 (2022) 133–155.

[2] J. Li, M. Lei, A brief survey for fake news detection via deep learning models, Procedia Computer Science 214 (2022) 1339–1344.

[3] S. K. Hamed, M. J. Ab Aziz, M. R. Yaakub, A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion, Heliyon 9 (2023).

[4] A. Khraisat, Manisha, L. Chang, J. Abawajy, Survey on deep learning for misinformation detection: Adapting to recent events, multilingual challenges, and future visions, Social Science Computer Review (2025) 08944393251315910.

[5] P. P. Mathai, S. M. Louis, Misinformation detection in the era of large language models: Challenges, advances, and future directions, in: 2025 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), IEEE, 2025, pp. 1–8.

[6] L. Tian, Adaptable Transfer Learning Models for Online Misinformation Detection, Ph.D. thesis, RMIT University, 2024.

[7] S. Rananga, A. Modupe, A. Isong, V. Marivate, Misinformation detection: a review for high and low resource languages (2024).

[8] X. Wang, W. Zhang, S. Rajtmajer, Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey, arXiv preprint arXiv:2410.18390 (2024).

[9] M. Islam, J. A. Khan, M. Abaker, A. Daud, A. Irshad, Unified large language models for misinformation detection in low-resource linguistic settings, arXiv preprint arXiv:2506.01587 (2025).

[10] 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.

[11] 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.

[12] G. K. Shahi, Y. Mejova, Too little, too late: Moderation of misinformation around the russo-ukrainian conflict, Websci ’25, 2025. doi:10.1145/3717867.3717876.

[13] G. K. Shahi, T. A. Majchrzak, Amused: An annotation framework of multimodal social media data, in: F. Sanfilippo, O.-C. Granmo, S. Y. Yayilgan, I. S. Bajwa (Eds.), Intelligent Technologies and Applications, Springer International Publishing, Cham, 2022, pp. 287–299.

[14] O. Khattab, K. Santhanam, X. L. Li, D. Hall, P. Liang, C. Potts, M. Zaharia, Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP, arXiv preprint arXiv:2212.14024 (2022).

[15] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, C. Potts, Dspy: Compiling declarative language model calls into self-improving pipelines, 2024.