Detecting 2022 Russo–Ukrainian Conflict Misinformation Using a Hybrid Transformer Approach

Pratik Priyanshu

SRH University Heidelberg, Germany

Abstract

Social media misinformation during conflicts spreads faster than fact-checkers can respond. We address this challenge through the PROMID shared task at FIRE 2025 (Subtask 3), which focuses on detecting misinformation in tweets related to the Russo–Ukrainian conflict. Our approach combines multilingual transformer-based representations with engineered linguistic features capturing stylistic patterns characteristic of misleading content. The dataset exhibits extreme class imbalance, with genuine content outnumbering misinformation by a ratio of 94:1. We address this challenge using class-weighted loss functions, decision threshold tuning, and stratified cross-validation. Experimental results demonstrate that the proposed hybrid system achieves a weighted F1-score of 0.918, with a recall of 0.94 and precision of 0.87 for misinformation detection.

Keywords

2022 Russo-Ukrainian Conflict, LLMs, Misinformation, Classification model

1. Introduction

Social media rapidly becomes an information battlefield when armed conflicts intensify. During the Russo–Ukrainian conflict, false narratives have spread across languages and geographic borders within hours, making it increasingly difficult to distinguish fact from fiction. Social media platforms such as Twitter facilitate near real-time communication across continents; however, this speed also introduces serious vulnerabilities to manipulation, coordinated disinformation campaigns, and the large-scale amplification of misleading content.

The stakes in such environments are exceptionally high. When misinformation spreads unchecked during periods of armed conflict, it can significantly influence public opinion, shape political and policy decisions, and, in extreme cases, even impact military operations. Emotionally charged and sensational false content often propagates more rapidly and reaches broader audiences than factual information or subsequent corrections. This imbalance creates a pressing need for automated misinformation detection systems capable of operating at scale, supporting human moderators, journalists, and fact-checkers in identifying potentially harmful content before it becomes widely disseminated.

Automatic misinformation detection, however, remains a challenging task. Social media text is inherently noisy, typically short, grammatically inconsistent, and heavily dependent on implicit context. Authors of deceptive content frequently employ sophisticated rhetorical strategies such as sensationalism, fear-mongering, identity-based appeals, and selective framing to attract attention and evade detection. These challenges are further amplified in conflict scenarios, where discussions span multiple languages and often involve code-switching, requiring systems that can generalize effectively across linguistic and cultural boundaries.

Furthermore, misinformation detection is adversarial by nature. As detection systems improve, malicious actors adapt their tactics by crafting narratives that blend factual elements with falsehoods, exploit ambiguity, or manipulate contextual interpretation. Techniques such as strategic omission, distorted reinterpretation of genuine information, and deliberate imitation of authentic journalistic style render simple keyword-based or surface-level approaches ineffective for robust detection.

Previous research in this domain has explored a wide range of methods, from manually engineered feature-based approaches to complex neural architectures. Early studies emphasized linguistic style,

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

pratikpriyanshu12345@gmail.com (P. Priyanshu)

source credibility indicators, and information propagation patterns within social networks [1, 2]. More recent transformer-based models have demonstrated strong performance by capturing deep semantic and contextual relationships through attention mechanisms [3, 4]. Despite their success, these models often struggle under conditions of extreme class imbalance, where genuine content vastly outnumbers misinformation, and their substantial computational requirements limit practical deployment in resource-constrained settings [5].

In this work, we present a misinformation detection system developed for the PROMID shared task at FIRE 2025 [6, 7]. The task provides annotated Twitter data related to the Russo–Ukrainian conflict, collected using the AMUSED framework [8]. The primary objective of this research is to develop a system capable of effectively detecting misinformation under a severe class imbalance of approximately 94:1, while maintaining high recall without sacrificing precision. To achieve this, we combine multilingual transformer-based semantic representations with carefully engineered linguistic features that capture stylistic cues associated with misleading content. At the same time, we emphasize computational efficiency to ensure that the proposed approach remains suitable for deployment in resource-constrained environments. Through this design, we aim to achieve strong weighted F1 performance across both majority and minority classes, demonstrating that practical, interpretable, and efficient misinformation detection systems can be realized without reliance on excessively large models.

2. Related Work

Research on misinformation detection has evolved substantially over the past decade, moving from manually engineered features to advanced neural architectures. Early approaches relied on handcrafted linguistic features such as punctuation usage, writing style, readability scores, and sentiment patterns to identify deceptive content [1, 2]. In addition, source credibility indicators, including author reputation, account metadata, and information propagation characteristics such as retweet cascades and temporal dynamics, were explored to assess content reliability [9]. While these feature-based methods offered strong interpretability and low computational cost, they struggled to generalize across domains and required frequent manual updates as misinformation tactics evolved.

Subsequent advances in deep learning introduced neural models such as convolutional neural networks and recurrent neural networks, which improved representation learning by capturing local and sequential patterns in text [10, 11]. Although these models outperformed traditional approaches on several benchmarks, they required large labeled datasets and provided limited transparency into their decision-making processes. More recently, transformer-based architectures, including BERT and multilingual variants such as XLM-RoBERTa, achieved state-of-the-art performance by learning contextualized representations through self-attention mechanisms [3, 4]. Despite their effectiveness, transformer models face practical challenges in real-world misinformation detection, particularly severe class imbalance [5], mismatch between pretraining data and informal social media language [12], and substantial computational requirements that hinder deployment in resource-constrained environments.

To address these limitations, recent work has explored hybrid approaches that combine transformer-based semantic representations with explicit linguistic and stylistic features, leveraging the strengths of both paradigms [13, 14]. Such models have demonstrated improved robustness, interpretability, and adaptability, especially in multilingual and low-resource settings where labeled data is scarce and linguistic variation is high [15, 16, 17]. These developments motivate the hybrid methodology adopted in this work, which aims to balance performance, efficiency, and interpretability for conflict-related misinformation detection.

3. Dataset and Problem Formulation

The PROMID (Prompt Recovery and Misinformation Detection) shared task at FIRE 2025 addresses the problem of detecting propaganda and misinformation in social media during military conflicts [6]. We

focus on Subtask 3, which formulates misinformation detection as a binary classification problem on tweets related to the Russo–Ukrainian war.

Given a tweet, the task is to determine whether it contains misinformation. The organizers adopt a generalized definition of misinformation, encompassing false or misleading assertions of fact, fabricated or manipulated media, distortion of genuine information through contextualization, and propaganda narratives aimed at influencing public perception rather than informing. This definition reflects the inherent complexity of information manipulation during conflicts, where the boundary between opinion, deception, and deliberate falsehood is often ambiguous [18].

The dataset consists of 34,538 tweets collected through systematic monitoring of relevant hashtags, keywords, and accounts. Annotation and data collection were performed using the AMUSED framework [8]. Each tweet is assigned a binary label indicating misinformation or non-misinformation. The dataset exhibits extreme class imbalance, with only 364 tweets (approximately 1.05%) labeled as misinformation, corresponding to a ratio of roughly 94:1. Under such conditions, naive classifiers can achieve high accuracy by predicting the majority class while completely failing to identify misinformation, rendering accuracy an unsuitable evaluation metric.

The tweets display typical social media characteristics, including brevity, informal language, heavy use of hashtags, user mentions, and URLs, multilingual content in English, Russian, and Ukrainian, frequent code-switching, and emotionally charged language. These properties pose challenges for traditional NLP techniques and necessitate careful preprocessing strategies.

Annotation was performed by trained annotators following task-specific guidelines. Inter-annotator agreement was measured using Cohen's kappa, and disagreements were resolved through adjudication. Despite these measures, some degree of subjectivity remains due to borderline cases and partially misleading content, introducing unavoidable label noise.

Evaluation follows the official task protocol, using weighted F1-score to account for class imbalance. Weighted F1 combines class-wise F1-scores according to their support, preventing dominance by the majority class while reflecting overall system performance. Precision and recall are also reported to provide detailed insights into model behavior under different deployment scenarios.

4. Methodology

Our misinformation detection framework integrates multiple components designed to work synergistically: text preprocessing to normalize noisy social media input, hybrid feature extraction combining learned and engineered representations, and training strategies explicitly designed to address extreme class imbalance.

4.1. Text Preprocessing

Raw tweets are subjected to a normalization pipeline that reduces noise while preserving meaningful stylistic and semantic signals relevant to misinformation detection. The preprocessing strategy balances two competing objectives: standardizing input to improve generalization and reduce vocabulary sparsity, while retaining linguistic patterns characteristic of deceptive content.

All URLs are replaced with a special `[URL]` token, preserving the presence of external references without inflating vocabulary size. User mentions are similarly normalized to a `[USER]` token to reduce sparsity and protect privacy. Emojis are converted to their textual descriptions to retain emotional and semantic information. Excessive character repetitions used for emphasis are truncated to a maximum of two consecutive characters. Original casing is preserved for transformer inputs to retain semantic and stylistic cues, while lowercased text is used for engineered feature computation. Finally, redundant whitespace is collapsed into single spaces. Together, these steps produce cleaner and more consistent textual representations without removing signals indicative of misinformation.

4.2. Hybrid Feature Extraction

Feature extraction combines two complementary sources of information: contextual semantic representations learned by transformer-based language models and explicit linguistic features derived from domain knowledge. This hybrid design is motivated by the observation that while transformers excel at capturing deep semantic and contextual relationships, they may underemphasize surface-level stylistic cues that frequently characterize misleading social media content.

We employ the multilingual transformer `xlm-roberta-base`, pretrained on over 100 languages including English, Russian, and Ukrainian. The model enables cross-lingual generalization without language-specific architectures and offers a favorable trade-off between representational power and computational efficiency with approximately 270 million parameters. Given an input tweet token sequence $\{t_1, t_2, \ldots, t_n\}$, the transformer produces contextualized token embeddings using self-attention. The embedding of the special `[CLS]` token is extracted as a sentence-level representation:

$$h_{\text{cls}} = \text{Transformer}(x)_{\text{cls}}. \tag{1}$$

This 768-dimensional vector encodes semantic content, syntactic structure, and contextual dependencies present in the tweet. The transformer is fine-tuned during training using a small learning rate to adapt representations to the misinformation detection task while mitigating catastrophic forgetting.

To complement transformer embeddings, we extract a set of interpretable linguistic features associated with misinformation in prior studies and journalistic practice. These features capture aspects of writing style and form, including text length and complexity, punctuation usage, capitalization patterns, frequency of hashtags, URLs, and user mentions, numerical expressions used to convey spurious credibility, and repetitive structures indicative of automated or manipulative text. The resulting feature vector $f \in \mathbb{R}^d$ (with $d = 15$) is standardized using statistics computed on the training data.

4.3. Model Architecture

The hybrid model fuses transformer embeddings and engineered features via concatenation:

$$z = [h_{\text{cls}} \oplus f] \in \mathbb{R}^{768+d}. \tag{2}$$

The combined representation is passed to a sigmoid-based classifier:

$$p = \sigma(Wz + b), \tag{3}$$

where $W \in \mathbb{R}^{1\times(768+d)}$ and $b \in \mathbb{R}$ are learned parameters. The output $p \in [0, 1]$ represents the probability that a tweet contains misinformation.

More complex fusion strategies, including attention-based mechanisms and deeper classification heads, were explored but did not yield meaningful performance gains over simple concatenation while increasing computational cost and overfitting risk.

4.4. Handling Class Imbalance

Severe class imbalance poses a major challenge for misinformation detection. Without explicit mitigation, models achieve deceptively high accuracy by predicting only the majority class.

4.4.1. Class-Weighted Loss

We employ class-weighted binary cross-entropy:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[w_1 y_i \log(p_i) + w_0(1 - y_i)\log(1 - p_i)\right], \tag{4}$$

where class weights are defined inversely proportional to class frequencies:

$$w_1 = \frac{N}{2n_1}, \quad w_0 = \frac{N}{2n_0}. \tag{5}$$

Here, n_1 and n_0 denote the number of misinformation and non-misinformation samples, respectively.

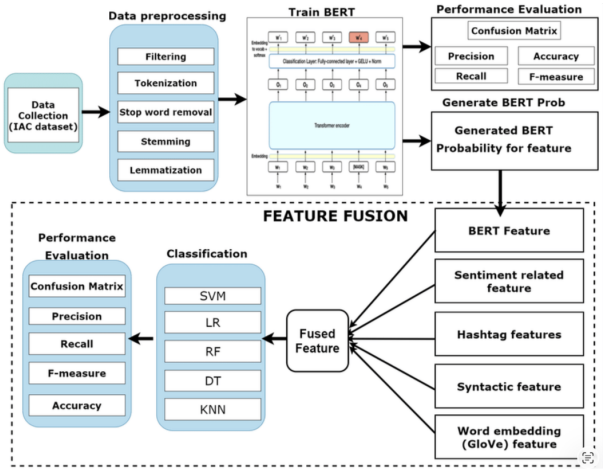

Figure 1: Overview of the hybrid feature fusion architecture. Transformer embeddings from XLM- RoBERTa combine with engineered linguistic features for misinformation classification

4.4.2. Threshold Tuning

Rather than using a fixed decision threshold of 0.5, we tune the threshold τ on validation data to maximize weighted F1-score:

$$\hat{y}_i = \begin{cases} 1 & \text{if } p_i \geq \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Thresholds in the range $[0.1, 0.9]$ are evaluated with a step size of 0.01.

4.4.3. Stratified Cross-Validation

We employ stratified three-fold cross-validation to ensure representative class distributions across training and validation splits.

4.5. Training Configuration

The model is trained using the AdamW optimizer with a learning rate of 2×10^{-5}. A linear learning rate schedule with a 10% warmup phase followed by linear decay is applied. Mixed-precision (FP16) training with automatic loss scaling is used to improve efficiency and reduce memory usage. Training proceeds for three epochs with a batch size of 16 per GPU, corresponding to approximately 6,500 optimization steps per fold. Early stopping is applied based on validation weighted F1-score. Dropout with probability 0.1 is applied to transformer layers, and a dropout rate of 0.3 is applied to the concatenated feature vector.

5. Experimental Results

All experiments are conducted using PyTorch 2.0 and the Hugging Face Transformers library (v4.30). Training is performed on a single NVIDIA RTX 3090 GPU with 24 GB memory, though the model can be deployed on GPUs with 16 GB memory or less. The full stratified cross-validation procedure, including hyperparameter tuning, completes in approximately four hours, demonstrating the computational efficiency of the proposed approach.

Results show consistent performance improvements across model variants. A transformer-only baseline significantly outperforms feature-based models, highlighting the importance of contextual semantic representations. Incorporating engineered linguistic features further improves performance, indicating that stylistic cues provide complementary information. The final hybrid model with threshold tuning achieves a weighted F1-score of 0.918, with recall of 0.94 and precision of 0.87 for misinformation

detection, alongside near-perfect performance on the majority class. Low variance across folds indicates strong generalization despite the limited number of positive samples.

Error analysis reveals that most failures involve ambiguous or context-dependent tweets, well-written narratives closely resembling legitimate reporting, sarcasm, and code-switching. Ablation studies confirm that class weighting, threshold tuning, and engineered features each contribute meaningfully to overall performance.

6. Discussion

The experimental results highlight several practical implications for misinformation detection systems. The proposed hybrid architecture demonstrates that strong performance can be achieved without extremely large models, enabling deployment on standard hardware with reasonable training time. The inclusion of engineered linguistic features improves interpretability, which is particularly important in content moderation contexts where transparency and accountability are essential. Decision threshold tuning further provides operational flexibility, allowing practitioners to balance precision and recall based on specific deployment requirements.

At the same time, several limitations remain. The model is trained on Twitter data related to a specific conflict and may not generalize seamlessly to other domains, platforms, or evolving misinformation strategies. Performance may vary across languages, particularly in low-resource and code-switched settings. Additionally, analyzing tweets in isolation ignores conversational context, user behavior, and multimodal signals that often play a critical role in misinformation spread.

Ethical considerations are central to deployment. Automated systems must balance the risk of suppressing legitimate speech against the harm caused by unchecked misinformation, particularly during conflicts. Human-in-the-loop oversight remains essential for handling ambiguous cases and mitigating bias. Overall, while no single system can fully solve the misinformation problem, well-designed hybrid approaches offer an effective, efficient, and responsible foundation for practical deployment and future extensions.

Acknowledgments

We thank the PROMID shared task organizers and the FIRE 2025 workshop team for curating the dataset and establishing the evaluation framework that made this research possible. We are grateful for the computational resources provided that enabled our experiments. We also acknowledge valuable feedback from reviewers that improved this work.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys 53 (2020) 1–40.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD Explorations Newsletter 19 (2017) 22–36.

[3] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[4] A. Conneau, et al., Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[5] J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance, Journal of Big Data 6 (2019) 1–54.
[6] 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.
[7] 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.
[8] 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.
[9] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 675–684.
[10] L. Hu, S. Wei, Z. Zhao, B. Wu, Deep learning for fake news detection: A comprehensive survey, AI Open 3 (2022) 133–155.
[11] J. Li, M. Lei, A brief survey for fake news detection via deep learning models, Procedia Computer Science 214 (2022) 1339–1344.
[12] S. Rananga, A. Modupe, A. Isong, V. Marivate, Misinformation detection: A review for high and low resource languages, 2024.
[13] 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).
[14] 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).
[15] X. Wang, W. Zhang, S. Rajtmajer, Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey, 2024. arXiv:2410.18390.
[16] M. Islam, J. A. Khan, M. Abaker, A. Daud, A. Irshad, Unified large language models for misinformation detection in low-resource linguistic settings, 2025. arXiv:2506.01587.
[17] 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.
[18] 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.