Misinformation Detection in Russo-Ukrainian Conflict Tweets

Thibault Ehrhart[1], Raphaël Troncy[1], Grégoire Burel[2] and Harith Alani[2]

[1]*EURECOM, 450 Route des Chappes, 06410 Biot, France*
[2]*The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom*

Abstract

Misinformation on social media poses a significant challenge during major geopolitical events, where rapid dissemination of misleading content can distort public understanding. The PROMID 2025 Subtask 3 focuses on identifying misinformation in tweets related to the 2022 Russo-Ukrainian conflict, a task complicated by extreme class imbalance, multilingual content, and heterogeneous metadata. In the provided dataset, misinformation accounts for only 1.05% of all tweets, making it difficult for transformer-based models to learn generalizable patterns. To address this challenge, we evaluate two approaches that both rely on the RoBERTa-large transformers based model: a baseline model trained solely on the original PROMID dataset, and an augmented model that incorporates an additional 5,022 Ukraine-related misinformation tweets coming from the Fact-checking Observatory (FCO). Our results show that while the baseline model achieves high precision, it performs poorly in recall due to overfitting on the limited misinformation examples. In contrast, the augmented model substantially improves misinformation detection, increasing F1-score from 0.4682 to 0.6967 and weighted F1 from 0.8516 to 0.9059. Our findings demonstrate that targeted data augmentation is an effective strategy for mitigating severe class imbalance and enhancing generalization in misinformation detection tasks. This constitutes the ClimateSense approach in the public leaderboard that was ranked 1st on the final test set of the PROMID 2025 Subtask 3[1]. Our approach is fully reproducible using the code at https://github.com/climatesense-project/promid2025-task3.

Keywords

Misinformation detection, Data augmentation, RoBERTa, Fact checking

1. Introduction

The proliferation of misinformation on social media platforms has become a critical challenge, particularly during major geopolitical events such as the 2022 Russo-Ukrainian conflict. The ability to automatically detect misinformation is essential for maintaining information integrity and public trust. This paper presents our approach to PROMID Subtask 3, which focuses on binary classification of tweets related to the 2022 Russo-Ukrainian conflict as either misinformation or genuine content. The task is part of the PROMID (Prompt Recovery for MisInformation Detection) shared task, which aims to explore methods for identifying misinformation in human and LLM-generated texts [1, 2, 3].

The task presents several significant challenges that make it particularly difficult for machine learning approaches. First, the dataset shows a severe class imbalance, with misinformation tweets representing only 1.05% of the training data (364 misinformation versus 34,174 non-misinformation tweets). Second, misinformation content appears in multiple languages and varies significantly in linguistic characteristics. Third, the social media context provided (both textual and metadata features) differs between misinformation and non-misinformation instances.

[1] https://www.codabench.org/competitions/10869/#/pages-tab

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

*Corresponding author.

†These authors contributed equally.

thibault.ehrhart@eurecom.fr (T. Ehrhart); raphael.troncy@eurecom.fr (R. Troncy); gregoire.burel@open.ac.uk (G. Burel); harith.alani@open.ac.uk (H. Alani)

https://www.eurecom.fr (T. Ehrhart); https://www.eurecom.fr (R. Troncy); https://www.open.ac.uk (G. Burel); https://www.open.ac.uk (H. Alani)

0000-0003-1377-8279 (T. Ehrhart); 0000-0003-0457-1436 (R. Troncy); 0000-0003-0029-5219 (G. Burel); 0000-0003-2784-349X (H. Alani)

We hypothesize that the extreme class imbalance in the PROMID dataset may lead models to overfit to the limited misinformation examples, learning superficial patterns rather than generalizable features of misinformation. To test this hypothesis, we develop and compare two approaches: (1) Baseline Approach, fine-tuning RoBERTa-large directly on the original PROMID dataset with class weighting and oversampling to address imbalance, and (2) Augmented Approach, incorporating external Ukraine-related misinformation data to increase minority class representation before fine-tuning.

Our comparative analysis reveals that while the baseline model performs well on the majority class, it struggles to correctly identify misinformation due to overfitting. By incorporating additional Ukraine-related misinformation data, our augmented approach improves the detection of misinformation and achieves better overall performance while demonstrating enhanced generalization capabilities. On the official test set, our system achieved a weighted F1-score of 0.91, ranking 1st out of 11 participating teams. We release the source code of our approach at https://github.com/climatesense-project/promid2025-task3.

The remainder of this paper is organized as follows. Section 2 reviews related work on misinformation detection using transformer-based models. Section 3 describes our methodology, including dataset analysis, preprocessing, and model architecture. Section 4 presents our experimental results and error analysis. Finally, Section 5 concludes with a discussion of limitations and future works.

2. Related Work

Research on tweet-level misinformation detection has increasingly focused on large pretrained language models, which consistently outperform traditional machine-learning approaches such as Support Vector Machines (SVMs) and Long Short-Term Memory networks (LSTMs). Transformer-based architectures, including BERT [4], RoBERTa [5], DeBERTa [6], and XLM-R [7], capture richer contextual and semantic cues in short social media posts, which makes them particularly effective for this task [8].

Toraman et al. demonstrate that transformers models significantly outperform classical baselines on MiDe22, a multilingual (English/Turkish) misinformation dataset covering multiple events, including the 2022 Russo-Ukrainian conflict [9]. Their results show that DeBERTa achieves the highest F1-score of 83.95% on English misinformation detection, while XLM-R performs best on Turkish (82.82% F1). Similarly, Weinzierl and Harabagiu report strong performance on the VaccineLies corpus [10], while Hossain et al. observes lower performance on COVID-19 misinformation [11], highlighting that dataset characteristics such as linguistic heterogeneity, topic complexity, and class distribution strongly influence model effectiveness. These findings reinforce that misinformation detection performance is highly dataset-specific and does not transfer reliably across domains.

Transformer-based models have also been successful in related tasks such as rumor detection and propaganda identification. Anggrainingsih et al. demonstrated that BERT-based sentence embeddings improve accuracy over non-transformer approaches by capturing nuanced linguistic and discourse cues in short, noisy texts [12]. In the propaganda detection domain, the SemEval shared tasks on propaganda technique classification [13] have shown that fine-tuned transformers can identify specific manipulation techniques in news articles with high accuracy [14].

Our work builds on these foundations by applying transformer-based classification to Ukraine-related misinformation while specifically addressing the extreme class imbalance through external data augmentation [15, 16].

3. Methodology

3.1. Task Definition and Dataset Overview

PROMID Subtask 3 focuses on misinformation detection in social media texts. The objective is to classify tweets related to the 2022 Russo-Ukrainian conflict as either misinformation (positive class) or non-misinformation (negative class). The dataset comprises manually annotated tweets collected using the Twitter API during the first year of the 2022 Russo-Ukrainian conflict. A key challenge of this subtask is the highly imbalanced class distribution, which tests how well models perform under these conditions. The dataset contains misinformation tweets in multiple languages, with additional metadata (e.g. account age, bot account indicators). Performance is evaluated using Precision, Recall, and weighted-average F1-score.

3.1.1. PROMID Training Dataset

The PROMID training dataset is derived from the work of Shahi and Mejova [1], who analyzed moderation of misinformation around the 2022 Russo-Ukrainian conflict. The dataset was collected using the AMUSED framework [17], a systematic approach for annotating multimodal social media data. It contains 34,538 tweets collected during the first year of the Russo-Ukrainian conflict. The dataset exhibits a severe class imbalance, with only 364 misinformation tweets (1.05%) and 34,174 non-misinformation tweets (98.95%). Each tweet in the dataset includes 32 features capturing tweet content and author metadata, including tweet engagement metrics (favorite count, retweet count), user profile information (followers count, friends count, account creation date, verification status), bot detection scores (Botometer score with calculation date, manual bot check labels), temporal features (account age relative to conflict start), and content metadata (language, hashtags, geolocation).

Notably, misinformation tweets have a substantially richer metadata profile, with an average feature fill rate of 80.8% compared to 52.8% for non-misinformation tweets.

To further characterize the dataset, we compute additional descriptive statistics for the textual content, metadata coverage, and language distribution of tweets. Table 1 summarizes these statistics.

Table 1
Statistical Analysis of the PROMID Dataset

Statistic	Misinformation	Non-Misinformation	Combined
Number of tweets	364	34,174	34,538
Percentage	1.1%	98.9%	100%
Average tweet length (chars)	211.1	180.5	180.8
Standard deviation tweet length (chars)	77.5	87.4	87.4
Average tweet length (tokens)	30.9	22.9	23.0
Standard deviation tweet length (tokens)	12.7	13.2	13.2
Average hashtags per tweet	1.2	4.2	4.2
Average mentions per tweet	0.2	0.7	0.7
Average URLs per tweet	0.9	0.8	0.8
Average metadata fill rate (%)	77.1	50.2	50.4
Unique words	4,708	139,914	140,280
Tweets in English (%)	70.3%	49.1%	49.3%

Several patterns emerge from Table 1. Misinformation tweets tend to be longer, averaging 211.1 characters (30.9 tokens) compared to 180.5 characters (22.9 tokens) for non-misinformation content. Interestingly, genuine tweets contain significantly more hashtags (4.2 vs. 1.2) and mentions (0.7 vs. 0.2). The proportion of English content in misinformation tweets is also much higher (70.3% vs. 49.1%).

3.1.2. FCO Ukraine Dataset

To address the extreme class imbalance, we incorporate additional misinformation about the 2022 Russo-Ukrainian conflict coming from the Fact-checking Observatory (FCO)[1] [18]. The unfiltered FCO data consists of more than 6 million Ukraine-related misinformation tweets collected between November 2021 and June 2023.

The FCO website was initially created in 2020 as an effort to track misinformation and the impact of fact-checking during the COVID-19 pandemic by automatically generating human-readable weekly

[1] Fact-checking Observatory, https://fcobservatory.org/.

reports about misinformation and fact-checks spread on 𝕏. The website was extended in late 2021 to include a section dedicated to the Russian invasion of Ukraine. To date, the website has generated 156 weekly COVID-19 misinformation reports and 83 Russo-Ukrainian war reports. The analysis of the social media posts led to insights about the spread of fact-checks misinformation on social media and its impact on misinformation [19, 20].

The FCO reports are generated through an automated pipeline that automatically collects relevant URLs from organizations that have been vetted by the Poynter Institute's International Fact-checking Network (IFCN)[2] and then tracks their mention on 𝕏 before generating visual reports every week using predefined templates. During the data collection process, each misinforming URL is given a normalised score between $+1$ (completely true claim) and -1 (completely false claim) based on the fact-cheking organisation rating. As a result, the collected data consists of posts with mentions of misinformation URLs or mentions of fact-checking URLs. For our approach to PROMID Subtask 3, we used this data to enrich the provided PROMID training dataset to alleviate the extreme class imbalance of the provided dataset.

The unfiltered dataset consists of more than 6 million 𝕏 posts obtained from more than 6k fact-checks. We only select posts that have a rating of -1 and contain misinformation (i.e., we remove posts that share fact-checking URLs). This results in 30,447 tweets from 23,366 distinct users. We then filter out non-English content and remove duplicates. The final set contains 5,022 unique tweets. This augmentation increases the number of misinformation samples from 364 to 5,386, which represents a 14.75× increase.

The combined dataset contains 39,560 tweets, comprising 5,386 misinformation (13.6%) and 34,174 non-misinformation (86.4%) samples. We split the data into training (85%, 33,626 tweets) and validation (15%, 5,934 tweets) sets using stratified sampling to preserve the class distribution across splits.

3.2. Text Preprocessing

Our preprocessing pipeline follows a minimal approach. We begin by converting all text entries to string format and removing any empty tweets that may have resulted from data collection errors. We deliberately retain URLs, hashtags, and mentions, as these may carry semantic information relevant to the misinformation detection.

The preprocessed text is then tokenized using the RoBERTa tokenizer with a maximum sequence length of 256 tokens. We apply padding to shorter sequences and truncation to longer ones to standardize the length across the dataset.

3.3. Model Architecture

We use RoBERTa-large [5] as our base model, which contains 355 million parameters. We add a classification head consisting of a dropout layer (p=0.1) followed by a linear layer that maps the 1024-dimensional [CLS] token representation to 2 output classes. The model is fine-tuned end-to-end using the following hyperparameters:

- Batch size: 16
- Learning rate: 2e-5 with linear decay
- Optimizer: AdamW with weight decay of 0.01
- Warmup: 10% of total training steps
- Gradient clipping: max norm of 1.0
- Training epochs: 4

These hyperparameters were selected based on preliminary experiments on a held-out portion of the training data.

[2] The International Fact-Checking Network (IFCN), https://www.poynter.org/ifcn.

3.3.1. Handling Class Imbalance

Given the severe class imbalance, we implement two complementary strategies to prevent the model from defaulting to majority class prediction:

Weighted Cross-Entropy Loss. We compute class weights inversely proportional to class frequencies using the formula:

$$w_c = \frac{N}{k \cdot n_c} \tag{1}$$

where N is the total number of samples, k is the number of classes, and n_c is the number of samples in class c. For the augmented dataset, this yields weights of 0.5788 for non-misinformation and 3.6726 for misinformation.

Weighted Random Sampling. During training, we apply weighted random sampling to construct each batch, with sampling probabilities inversely proportional to class frequencies.

3.3.2. Training Details

All experiments are conducted on a single NVIDIA L40S GPU with 48GB memory. Training the augmented model for 4 epochs takes approximately 10 minutes. We use the Hugging Face Transformers library [21] version 4.56.1 with PyTorch 2.8.0+cu126.

Table 2 shows the training progression for the augmented model across 4 epochs.

Table 2
Training Progression (Augmented Model)

Epoch	Train Loss	Train Acc	Val F1 (Misinfo)
1	0.1355	0.9301	0.9263
2	0.0521	0.9838	0.9063
3	0.0264	0.9898	0.9278
4	0.0152	0.9926	0.9352

The model achieves rapid convergence, with training accuracy reaching 93.01% after the first epoch and 99.26% by the fourth epoch. The slight dip in validation F1-score at epoch 2 suggests some initial instability, but performance recovers and peaks at 0.9352 in epoch 4. We select the epoch 4 checkpoint for final evaluation based on the validation performance.

4. Results

4.1. Main Results

Table 3 presents a comparison between the baseline trained on the original PROMID data, and the augmented model trained with external misinformation data.

The results reveal a clear precision-recall tradeoff between the two approaches. The baseline model achieves exceptional precision (97.64%) but critically poor recall (30.77%), detecting only approximately one-third of actual misinformation instances. This pattern strongly suggests that the model overfitting to the 364 training examples.

The augmented model demonstrates substantially improved generalization, achieving 90.44% precision and 56.33% recall. While precision decreases by 7.2 percentage points, this is outweighed by the 83.1% relative improvement in recall. The misinformation F1-score increases from 0.4682 to 0.6967 (48.8% improvement), and the weighted F1-score improves from 0.8516 to 0.9059 (6.4% gain), indicating better overall classification performance.

Table 3
Performance Comparison: Baseline vs. Augmented

Metric	Baseline	Augmented	Δ Abs	Δ Rel
Training Data				
Misinfo samples	364	5,386	+5,022	+1379%
Class balance	1.05%	13.6%	+12.55%	+1195%
Performance				
Precision	0.9764	0.9044	-0.0720	-7.4%
Recall	0.3077	0.5633	+0.2556	+83.1%
F1 (misinfo)	0.4682	0.6967	+0.2285	+48.8%
Weighted F1	0.8516	0.9059	+0.0543	+6.4%

4.2. Error Analysis

To better understand the precision-recall tradeoff, we conduct a detailed error analysis by examining specific examples where the baseline and augmented models diverge in their predictions.

Precision Loss Examples. The augmented model shows increased false positive rates on several categories of genuine content:

1. **Political commentary and opinion:** Tweets expressing strong political opinions about Ukraine policy are frequently misclassified. For example, statements criticizing government aid decisions or energy policy are flagged as misinformation, suggesting that the model mixes partisan rhetoric with false content.
2. **Sarcasm and rhetorical questions:** Sarcastic tweets such as *"It never ends. Why? Because Ukraine has already won"* are incorrectly flagged.
3. **Meta-commentary:** Tweets discussing social media content itself (e.g. *"Do you have a screenshot? The tweet has been deleted"*) trigger false positives, suggesting the model associates discussion of content manipulation with misinformation ecosystems.
4. **Non-English content:** Tweets in German, Dutch, Spanish, and other non-English languages show elevated false positive rates, which may result from the English-language filtering applied to the external augmentation data.

Recall Gain Examples. The augmented model successfully identifies diverse misinformation patterns that the baseline model misses:

1. **Conspiracy theories:** Claims about secret bioweapon laboratories, fabricated relationships between public figures, and distorted casualty figures are correctly detected.
2. **Propaganda techniques:** Tweets that selectively frame events to promote specific narratives or those containing verifiable false claims about military operations are identified, including fabricated claims about NATO aircraft being shot down, manufactured quotes from officials, and manipulated humanitarian contexts.
3. **Coordinates narratives:** The model detects recurring false narratives that appear across multiple tweets with slight variations.

5. Conclusion and Future Works

We presented a comparative study of misinformation detection approaches for tweets about the Russo-Ukrainian conflict, developed for PROMID Subtask 3. Our baseline model, trained solely on the original PROMID dataset with 364 misinformation examples, achieved high precision (0.9764) but poor recall (0.3077), consistent with our hypothesis that extreme class imbalance leads to overfitting.

By augmenting the training data with 5,022 external Ukraine-related misinformation examples, we achieved substantially improved generalization with 0.9044 precision, 0.5633 recall, and 0.9059 weighted F1-score. This represents a 48.8% improvement in misinformation F1-score over the baseline approach.

Our results demonstrate that data augmentation is a viable and effective strategy for addressing severe class imbalance in misinformation detection. The high precision maintained by our augmented model makes it suitable for deployment in scenarios where false positives carry significant costs.

Several directions could extend this work. First, incorporating user metadata (follower counts, account age, verification status) alongside textual features. Second, multilingual modeling using transformers such as XLM-R could better handle the diverse languages present in the dataset. Third, combining the high-precision baseline with the high-recall augmented model could achieve better precision-recall balance. Finally, using methods such as attention visualization and feature attribution could help identify which linguistic and contextual features most strongly indicate misinformation, which could be used for improving the model.

Acknowledgments

This work was supported by the European CHIST-ERA program within the ClimateSense project (Grant ID ANR-24-CHR4-0002, EPSRC EP/Z003504/1). We thank the PROMID shared task organizers for providing the dataset and evaluation framework.

Declaration on Generative AI

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

References

[1] G. K. Shahi, Y. Mejova, Too Little, Too Late: Moderation of Misinformation around the Russo-Ukrainian Conflict, in: 17th ACM Web Science Conference (WebSci), 2025. doi:10.1145/3717867.3717876.

[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), in: K. Ghosh, T. Mandl, S. Pal, S. Majumdar, A. Chakraborty (Eds.), Working Notes of FIRE: Forum for Information Retrieval Evaluation, CEUR-WS.org, 2025.

[3] 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: 17th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE), Association for Computing Machinery, 2025.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[6] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. arXiv:2006.03654.

[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.

[8] Y. Peskine, G. Alfarano, I. Harrando, P. Papotti, R. Troncy, Detecting COVID-19-related conspiracy theories in tweets, in: CEUR (Ed.), MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop, 2021.

[9] C. Toraman, O. Ozcelik, F. Şahinuç, F. Can, Mide22: An annotated multi-event tweet dataset for misinformation detection, 2024. arXiv:2210.05401.

[10] M. Weinzierl, S. Harabagiu, Vaccinelies: A natural language resource for learning to recognize misinformation about the covid-19 and hpv vaccines, 2022. arXiv:2202.09449.

[11] T. Hossain, R. L. Logan IV, A. Ugarte, Y. Matsubara, S. Young, S. Singh, COVIDLies: Detecting COVID-19 misinformation on social media, in: 1st Workshop on NLP for COVID-19, Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.nlpcovid19-2.11.

[12] R. Anggrainingsih, G. M. Hassan, A. Datta, Bert based classification system for detecting rumours on twitter, 2021. arXiv:2109.02975.

[13] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: 14th Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona, 2020, pp. 1377–1414. doi:10.18653/v1/2020.semeval-1.186.

[14] Y. Peskine, R. Tronc, P. Papotti, EURECOM at SemEval-2024 Task 4: Hierarchical Loss and Model Ensembling in Detecting Persuasion Techniques, in: 18th International Workshop on Semantic Evaluation (SemEval), Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.semeval-1.172.

[15] Y. Peskine, D. Korencic, I. Grubisic, P. Papotti, R. Troncy, P. Rosso, Definitions Matter: Guiding GPT for Multi-label Classification, in: International Conference on Empirical Methods for Natural Language Processing (EMNLP), Association for Computational Linguistics, 2023, pp. 4054–4063. doi:10.18653/V1/2023.FINDINGS-EMNLP.267.

[16] G. Burel, M. Mensio, Y. Peskine, R. Troncy, P. Papotti, H. Alani, CimpleKG: A continuously updated knowledge graph on misinformation, factors and fact-checks, in: 23rd International Semantic Web Conference (ISWC), Baltimore, USA, 2024.

[17] G. K. Shahi, T. A. Majchrzak, Amused: An annotation framework of multimodal social media data, in: International Conference on Intelligent Technologies and Applications, 2022.

[18] G. Burel, H. Alani, The fact-checking observatory: Reporting the co-spread of misinformation and fact-checks on social media, in: 34th ACM Conference on Hypertext and Social Media (HT), Association for Computing Machinery, 2023.

[19] G. Burel, T. Farrell, M. Mensio, P. Khare, H. Alani, Co-spread of misinformation and fact-checking content during the covid-19 pandemic, in: S. Aref, K. Bontcheva, M. Braghieri, F. Dignum, F. Giannotti, F. Grisolia, D. Pedreschi (Eds.), Social Informatics, Springer International Publishing, 2020, pp. 28–42.

[20] G. Burel, T. Farrell, H. Alani, Demographics and topics impact on the co-spread of covid-19 misinformation and fact-checks on twitter, Information Processing & Management 58 (2021).

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: International Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.