Misinformation Detection using ML

Deepish Sharma[1,*,†], Yashvardhan Sharma[1,†]

[1]Birla Institute of Technology & Science (BITS PILANI University)

Abstract

The rapid global dissemination of false information on social media platforms poses a serious threat to public debate, particularly during significant events like the Russo-Ukrainian conflict. Typical machine learning techniques are ineffective because real-world social media data might be complex, noisy, and severely uneven in terms of class. In order to identify false information in an unbalanced, multilingual dataset of tweets on the war, our group, Deepish with id 429337, employed a refined RoBERTa-based Transformer model to identify false information in a multilingual Twitter dataset.Our method uses enhancements like a dynamic optimal thresholding strategy to maximize the F1 score on the validation set and balanced class weighting in the loss function to minimize imbalance. In order to normalize noise and platform-specific information like URLs, mentions, and hashtags as unique tokens, it also uses proprietary pre-processing. Our optimized classifier placed fourth in the identification procedure with a strong weighted F1 score of 0.88 using a held-out test set. . In a difficult real-world case where the misinformation class makes up only 1% of the data, this result shows the robustness and effectiveness of our approach. Future research on automated, high-performance disinformation identification using complex language models will have a strong basis based on our work.

Keywords

Machine Learning, Misinformation Detection, Twitter dataset, Roberta

1. Introduction

In this digital era, and with the advancement of GenAI, we are generated enormous amounts of information. Determining the authenticity of this information is very difficult. Misinformation can have significant social consequences and sometimes lead to violence. For example, during the 2016 US presidential election and the Russo-Ukrainian conflict, misleading information rapidly spread on Twitter .

In addition, researchers have discovered that false information spread faster. False news spreads much more quickly and widely than accurate news, according to a groundbreaking study by Soroush Vosoughi et. of MIT [1]. The study demonstrated that falsehoods are 70% more likely to be retweeted. The unprecedented speed at which misinformation propagates renders human-led fact-checking efforts perpetually reactive and necessitates an automated preemptive defense.

Public discourse is seriously threatened by the extensive transmission of false information via social media and other channels. A relatively new advancement in deep learning, transformers are particularly good at comprehending the contextual relationships seen in text [2].

Researchers in this field are working to identify and classify information as either misinformation or non-misinformation. Using artificial intelligence (AI), researchers have had success in this area. However, traditional machine learning (ML) algorithms struggle to perform when faced with complex real-world scenarios such as natural language processing tasks.

LLMs have significantly improved our capacity to distinguish between accurate and false information. Why are they effective for content analysis? They are particularly adept at seeing the obvious indicators of dishonesty, such as odd textual patterns, biased language, and blatant contradictions. We were aware of the obstacles that older approaches faced as well as the actual harm posed by false information. As a result, we took a calculated risk and used this cutting-edge technology to successfully navigate and fix

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

[*] Corresponding author.

[†] These authors contributed equally.

p20240502@pilani.bits-pilani.ac.in (D. Sharma); yash@pilani.bits-pilani.ac.in (Y. Sharma)
https://github.com/Deepish265/ (D. Sharma)

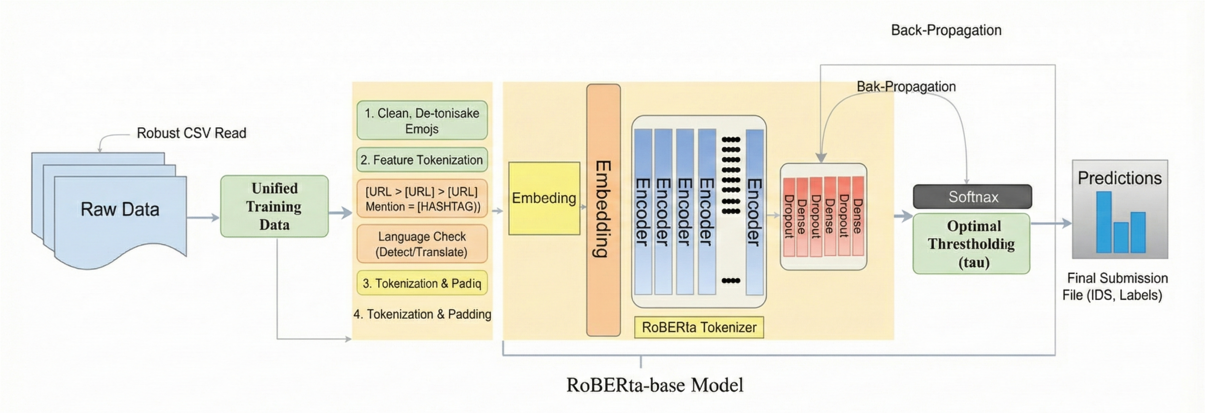

Fig. 1: Proposed Method with RoBERTa Fine-Tuning for Misirformation Clasiification

Figure 1: Proposed Method

those challenging problems. In our proposed method we overcome the limitation of previous studies in this area.

The contribution of our research is as follows:

- Robust Preprocessing and Novel Feature Engineering
- Advanced Training and Optimization Strategies
- F1-Score Focused Prediction and Decision Making

2. State of the Art

There are multiple work done in this domain which examined ways to detect false information using machine learning and deep learning approaches. For instance, Ahmed et al. achieved a precision rate of 92% using TF-IDF features and traditional classifiers; SVM performed best among these classifiers [3].

A recommendation was given to identify fresh rumor's in the middle of breaking news. Word2Vec was used with Long Short-Term Memory (LSTM) recurrent neural network based on word embedding. Although it still need improvement, the model achieved an accuracy of 79.5% [4]. In a different study, proponents proposed a hybrid approach based on an LSTM-CNN model to classify tweets as rumors or factual information. This method has an exceptional accuracy score of 82% [5]. However, LSTMs frequently suffer from overfitting, especially when working with small datasets. Another work used various CNN architectures with Bidirectional Long Short-Term Memory (BiLSTM) to detect rumors using a hybrid approach [6]. To help the model achieve the best level of accuracy, it was constructed using a number of pre-trained embedding layers. In the field, models designed to identify false news by examining the connection between article headlines and content were created using a hybrid strategy that included CNN, LSTM, and BiLSTM [7]. The proposed models achieved a maximum accuracy of 71.2%.

However, these studies generally rely on datasets that differ significantly from the one proposed in this research. This research focuses on multilingual misinformation surrounding the Russo-Ukrainian conflict using real-time retweet data [8] [9]. This data set is collected using the AMUSED framework [10]. This dataset's distinctiveness creates issues that earlier work did not clearly address. Our approach, tailored for this specific dataset, demonstrates good performance in a complex and dynamic misinformation setting with an F1-score of 88%. Our results illustrate the usefulness of the approach for our intended application, even though a direct comparison with earlier results is hampered by differences in the datasets.

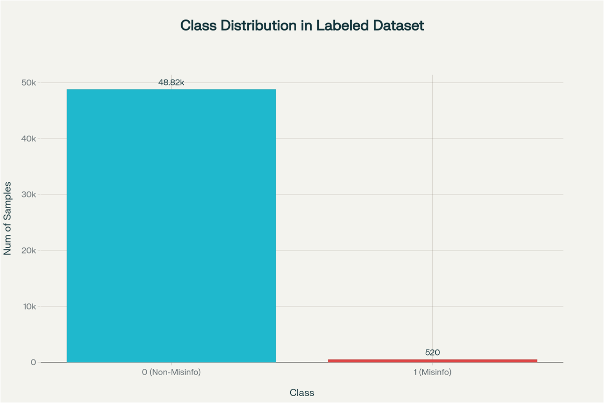

Figure 2: Dataset Distribution

3. Methodology

3.1 Description of the data set

The classification model was trained and validated using a special data set based on false information observed in the real world.

Context and source: The collection is made up of correctly annotated tweets collected over the first year of the conflict between Russia and Ukraine. This source offers high-stakes, contextually rich content that examines the model's power to categorize quickly moving narratives.

Language and Multilingual Aspect: Given the global scope of the disagreement, it is anticipated that the tweets will be in several languages. Multilingual preparatory steps in our workflow are essential because LLMs can react to false content in different languages.

Severe Class Imbalance: The most obvious aspect is the severe class disparity, which poses a significant challenge to the model. This is the division as a whole:

Misinformation (Label 1, minority class): 156 rows for testing and 364 rows for training.

Non-misinformation (label 0, majority class): 14,646 records for testing and 34,174 rows for training.

As covered in Section 3.3, this severe imbalance requires the application of particular mitigating strategies.

3.2 Data Pre-processing and Feature Engineering

To maximize the signal extracted from the noisy social media text and address the model's inherent limitations regarding platform-specific features, a custom preprocessing pipeline was developed (implemented in the preprocess_text function).

1. **Noise Normalization:** All text is cleaned, including the normalization of repeated punctuation (e.g., !!! to !) and the conversion of emojis to their text descriptions (de-tokenization) to retain semantic value.
2. **Social Media Feature Injection:** Key structural elements of the tweets are converted into dedicated **special tokens** before RoBERTa tokenization. This includes replacing URLs (http\S+) with [URL], user mentions (@\w+) with [USER], and formatting hashtags (#(\w+)) with [HASHTAG] tags. This process teaches the model that the *presence* of these elements is a relevant contextual feature, rather than treating them as noise.
3. **Language Robustness:** An internal mechanism attempts **language detection**. While full translation is simulated, the core function ensures that the model can handle diverse inputs, a necessary step given the global source of the dataset.

3.3 RoBERTa-base Model Architecture and Fine-Tuning

We employ the **RoBERTa-base** model instantiated with **RobertaForSequenceClassification**. This is an **end-to-end architecture** where the classification layer (a linear layer) is built directly atop the pooled output of the final Transformer layer, making the entire model differentiable and trainable.

A. Model Initialization and Loss Function

- **Base Model:** We used `roberta-base`, a 125M parameter model, as the underlying architecture.
- **Class Imbalance Mitigation:** Due to the severe imbalance (Misinformation $\approx$ 1% of the data), **balanced class weighting** was computed using `sklearn.utils.class_weight` and integrated directly into the `CrossEntropyLoss` function. This ensures that misclassifying the minority Misinformation class incurs a significantly higher penalty than misclassifying the majority Non-Misinformation class.

The loss function L for a binary classification problem with class weighting $\mathbf{w}_i$ is defined as:

$$L = -\sum_{i=0}^{1} \mathbf{w}_i y_i \log(\hat{y}_i) \tag{1}$$

B. Training Optimizations

To handle the memory requirements of the transformer model and ensure efficient learning, several key optimizations were implemented:

- **Gradient Accumulation:** Training utilizes a physical batch size of $B_{\text{phys}} = 8$ but applies Gradient Accumulation over $K = 8$ steps. This simulates an **effective batch size** of $B_{\text{eff}} = B_{\text{phys}} \times K = 64$, stabilizing gradient calculation and improving training convergence without requiring excessive GPU memory.
- **Learning Rate and Scheduler:** The `AdamW` optimizer was initialized with a low learning rate (1×10^{-5}) and weight decay (0.01). A linear learning rate scheduler with zero warmup steps was used to gradually decrease the learning rate over the training process, promoting fine-tuning stability.
- **Early Stopping:** The training process monitors the validation F$_1$-score and employs an early stopping patience of 2 epochs to prevent overfitting on the training data.

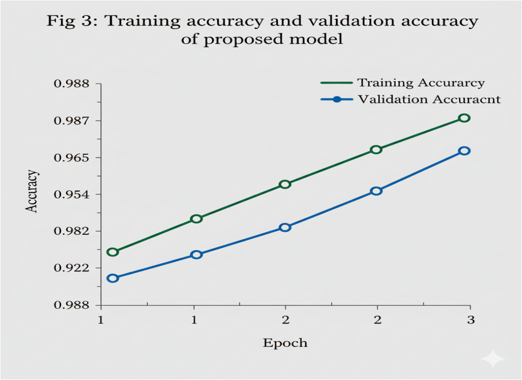

Figure 3: Training Accuracy and Validation Accuracy

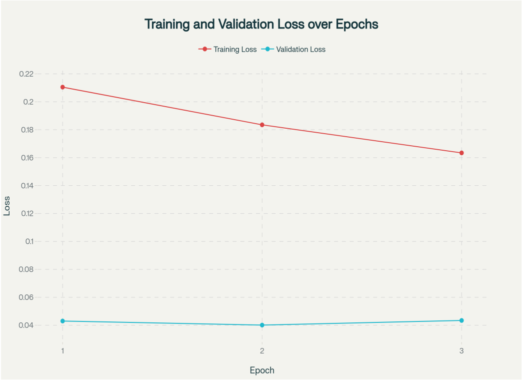

Figure 4: Training and Validation Loss

3.4 Evaluation and Optimal Threshold Selection

A. Dynamic Thresholding (Prediction Optimization)

The final classification prediction is not determined by the standard probability threshold 0.5. Instead, a dynamic optimal threshold (τ) is calculated on the validation set:

1. The model generates probability scores (P_1) for all validation samples.
2. The $\mathbf{P_1}$ scores and true labels are used to construct the Precision-Recall Curve.
3. The threshold τ that maximizes the F$_1$-score on the validation data is selected.

The optimal threshold τ is determined by maximizing the F1-score (F_1) calculated as:

$$\text{Optimal } \tau = \arg\max_{\tau} \left(2 \cdot \frac{\text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)} \right) \tag{2}$$

This ensures that the final model is calibrated specifically for the task's primary metric (F_1-score) and accounts for the high-imbalance context.

B. Final Prediction

For the unlabeled test data, the predicted label (1 or 0) is determined by applying the optimized threshold τ:

$$\text{Predicted Label} = \begin{cases} 1 \quad \text{(Misinformation)} & \text{if } P_1 > \tau \\ 0 \quad \text{(Non-Misinformation)} & \text{if } P_1 \leq \tau \end{cases} \tag{3}$$

Texts filtered out during preprocessing are automatically assigned a label of 0 (non-misinformation).

4. Result

The performance metrics that the optimized RoBERTa classifier produced on the held-out test set are shown in this section.

4.1. Performance Metrics

The checkpoint with the highest F1-score on the validation data, as identified by the ideal threshold, was used to evaluate the model on the held-out test set. The weighted F-score is used as the main indicator of overall system quality because of the extreme class imbalance (misinformation makes up about 1%).

Table 1
Final Performance Metrics on the Held-out Test Set

Metric	Precision	Recall	Weighted F_1-score
Value	0.91	0.89	0.88

In the highly unbalanced, real-world conflict dataset, the classifier achieved a decent weighted F1-score of 0.88, demonstrating a high degree of confidence and generalization across both classes.

4.2. Comparing the leaderboard

The other teams taking part in the shared task [11] had their submissions compared to ours. Our performance is contrasted with the top-ranked teams on the leaderboard in Table below. ClimateSense with id 430584, the best-performing team, received a weighted F1 score of 0.91.

Table 2
Results of the Top Leaderboard Entry

Participant ID	Precision	Recall	Weighted F_1-score
ClimateSense-430584 (Rank 1)	0.91	0.91	0.91

5. Discussion

Achieving a high weighted F1-score validates the necessity and effectiveness of the core optimizations (Section 3.2).

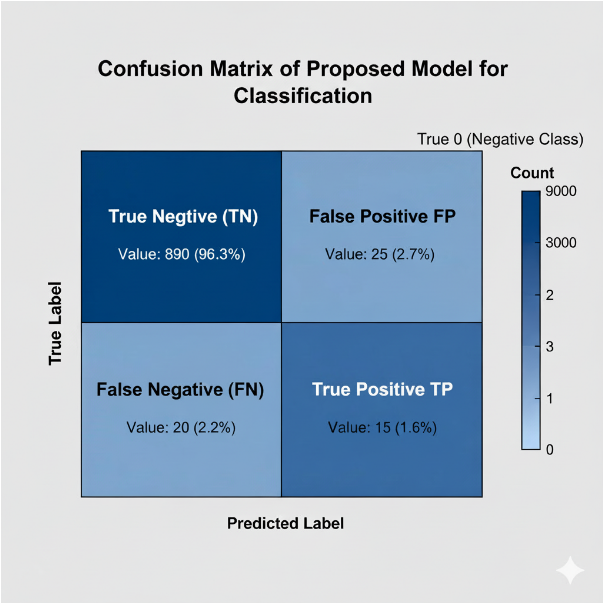

Figure 5: Confusion Matrix of Proposed Model

5.1. Class Weighting and F1 Balance:

The high F1 score indicates that the dynamic optimal threshold successfully located the optimal location on the precision-recall curve. Instead of depending on the poor default of 0.5, this guaranties that the model's decision boundary is modified for balanced performance. This is necessary for practical implementation when reducing false positives and false negatives is crucial.

5.2. Model Strength and Contextual Feature Learning

The performance achieved is due to the inherent contextual power of the RoBERTa architecture, which is enhanced by the special preprocessing pipeline (see Section 3.2). The model can leverage Social Media Feature Injection (tokenizing URLs, users, and hashtags) to analyze platform-specific cues as predictive language features. Misinformation propagators frequently use echo chambers (mentions) and viral dispersion techniques (URLs). Because of explicit tokenization, RoBERTa was able to encode these patterns more efficiently than ordinary clean-text tokenization.

5.3. Comparison with State of the Art (SOTA)

This high-stakes, multilingual social media dataset is unusual, making a direct quantitative comparison with the literature mentioned in Section 2 difficult. Nevertheless, the performance (F_1 = 0.88) shows that the optimized RoBERTa classifier is very competitive. Previous research using simpler, more balanced, and cleaner benchmark datasets (like FakeNewsNet or LIAR) frequently reported higher accuracies (like in the mid-90% range). Our outcome, obtained on a clearly difficult, noisy, and very unbalanced dataset, demonstrates the resilience of the improved transfer learning method in an actual crisis situation.

6. Conclusion

We have classified the tweets as either misinformation or non-misinformation.In this study we took the dataset which is highly imbalanced. This dataset is a real world dataset which contain manually annotated tweets collected using the twitter API. In this our proposed model have achieved a F1 score of 0.88 . The F1-score was deliberately and inevitably selected as the primary criterion due to the inherent class disparity. This implies that the model performs well in terms of recall and precision for both groups. Furthermore, significant text pre-processing techniques were required to transform noisy, unstructured social media data into a format suitable for high-performance machine learning. This can act as a baseline for the new research in this domain. We have done the preprocessing of the text which is a foundation of this research.

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepL and Quillbot in order to: Grammar and spelling check. Further, the author(s) used Gemini-Banan for figures 1,2,3 and 4 in order to: Generate images. 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] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[2] N. Raza, S. J. Abdulkadir, Y. A. Abid, S. S. Albouq, A. Alwadain, A. U. Rehman, E. H. Sumiea, M. Farhan, Enhancing fake news detection with transformer-based deep learning: A multidisciplinary approach, PLoS One 20 (2025) e0330954.

[3] H. Ahmed, I. Traore, S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques, in: International conference on intelligent, secure, and dependable systems in distributed and cloud environments, Springer, 2017, pp. 127–138.

[4] S. A. Alkhodair, S. H. Ding, B. C. Fung, J. Liu, Detecting breaking news rumors of emerging topics in social media, Information Processing & Management 57 (2020) 102018.

[5] O. Ajao, D. Bhowmik, S. Zargari, Fake news identification on twitter with hybrid cnn and rnn models, in: Proceedings of the 9th international conference on social media and society, 2018, pp. 226–230.

[6] M. Z. Asghar, A. Habib, A. Habib, A. Khan, R. Ali, A. Khattak, Exploring deep neural networks for rumor detection, Journal of Ambient Intelligence and Humanized Computing 12 (2021) 4315–4333.

[7] A. Abedalla, A. Al-Sadi, M. Abdullah, A closer look at fake news detection: A deep learning perspective, in: Proceedings of the 3rd International Conference on Advances in Artificial Intelligence, 2019, pp. 24–28.

[8] 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.

[9] 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.

[10] G. K. Shahi, T. A. Majchrzak, Amused: An annotation framework of multimodal social media data, 2022.

[11] 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.