A Social Media Misinformation Detection Model Integrating Semantic and Twitter Features

Junwei Peng, Zijie Lin and Zhongyuan Han*

Foshan University, Guangdong, China

Abstract

The proliferation of misinformation on social media platforms has become a critical challenge in the digital age. This study presents a hybrid deep learning approach for detecting misinformation by combining ModernBERT with comprehensive feature engineering. We participated in the FIRE-2025 Task 3 shared task on misinformation detection. Our methodology integrates transformer-based language understanding with hand-crafted features extracted from text, user profiles, and social engagement patterns. To address the severe class imbalance problem, we employ Focal Loss with strategic resampling techniques. The experimental results demonstrate that our hybrid model achieves weighted F1-scores of 0.97 and 0.82 on the two official test datasets, respectively.

Keywords

Misinformation Detection, Semantic Feature, Social Media Feature

1. Introduction

Social media platforms have become the principal source of information for the general public. However, due to limited content oversight and promotion based on popularity rather than accuracy, misinformation spreads rapidly. This can have serious consequences including influencing public opinion, undermining trust in institutions, and creating social unrest [1]. Such content often focuses on polarizing topics and garners massive popularity, amplifying its reach. Therefore, detecting and stopping misinformation at early stages is urgently needed.

We observe that misinformation exhibits not only subtle semantic differences but also distinctive patterns in Twitter-specific features. To leverage these complementary signals, we propose a hybrid model that integrates semantic understanding from ModernBERT with hand-crafted Twitter features for misinformation detection. Our approach achieves weighted F1-scores of 0.97 and 0.82 on the two test sets of FIRE-2025 Task 3, respectively [2, 3].

2. Related Work

Early approaches to misinformation detection primarily relied on hand-crafted features extracted from news content combined with traditional machine learning classifiers. These methods exploited the hypothesis that deceptive content exhibits distinctive patterns in writing style, enabling automatic detection through statistical analysis. Zhou and Zafarani [1] concluded that, within traditional machine learning frameworks, hand-crafted features for detecting fake news are typically extracted from four linguistic levels of text: lexical, syntactic, semantic, and discourse. Building upon these feature categories, representative early work includes: Feng et al. proposed a syntactic stylometry approach using CFG parse tree features for deception detection [4], and Pérez-Rosas et al. [5] proposed an integration of lexical, syntactic, and semantic features with SVM classifiers for fake news identification across multiple datasets. These traditional approaches established important foundations by revealing linguistic patterns distinguishing deceptive content.

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

*Corresponding author.

pengjw8611@gmail.com (J. Peng); lamlovezz113@gmail.com (Z. Lin); hanzhongyuan@gmail.com (Z. Han)

0009-0006-3925-8158 (J. Peng); 0009-0009-1492-809X (Z. Lin); 0000-0001-8960-9872 (Z. Han)

The advent of deep learning opened new research directions for misinformation detection. Ajao et al. [6] proposed hybrid CNN-RNN architectures that combine convolutional layers for local pattern extraction with recurrent layers for sequential modeling, enabling the capture of both spatial and temporal dynamics of deceptive language on Twitter. Additionally, Devlin et al. [7] proposed BERT, which leverages masked language modeling to learn contextual representations from large-scale corpora. BERT's bidirectional attention mechanism enables it to capture nuanced semantic relationships crucial for distinguishing subtle differences between genuine and misleading content. Building upon BERT's foundation, subsequent variants such as RoBERTa [8] optimized pre-training procedures with dynamic masking and larger batch sizes, and DeBERTa [9] introduced disentangled attention to separately model content and position information in the attention mechanism. These advances have substantially improved detection performance by better capturing semantic features and contextual nuances in social media discourse.

Beyond content analysis, research has demonstrated that social context features provide complementary signals for misinformation detection. Castillo et al. [10] integrated user-based features alongside content, demonstrating that user-level characteristics such as account age and follower count serve as valuable indicators of source credibility on Twitter.

Recognizing the value of these diverse features for social media misinformation detection, we propose a hybrid model that integrates them to address the FIRE 2025 detection task.

3. Task and Dataset Description

3.1. Task Definition

The FIRE-2025 Task 3 [2, 3] classifies Twitter posts into two categories: (i) **Misinformation** - posts containing false, misleading, or unverified information, and (ii) **Non-misinformation** - posts containing legitimate, accurate information.

3.2. Dataset Description

The dataset provided by the organizers [2, 3] is derived from Twitter posts about the Russo-Ukrainian conflict [11], collected using the AMUSED annotation framework [12]. Each data instance contains the following information: (i) **Text** - the main content of the social media post, (ii) **User metadata** - including follower count and friends count, and (iii) **Engagement metrics** - retweet count and favorite count.

3.3. Dataset Statistics

As shown in Table 1, the dataset exhibits severe class imbalance, with misinformation representing approximately 1.054% of both the training and validation data.

Table 1
Dataset Distribution

Dataset	Misinfo	Non-misinfo	Total
Train	364	34,174	34,538
Val	156	14,646	14,802

3.4. Evaluation Metrics

The organizers provide the following metrics for evaluation: (i) **Precision** - the proportion of correct misinformation predictions among all misinformation predictions, (ii) **Recall** - the proportion of actual misinformation cases correctly identified, and (iii) **Weighted F1-score** - the F1-score weighted by class support, providing overall model performance.

4. Methodology

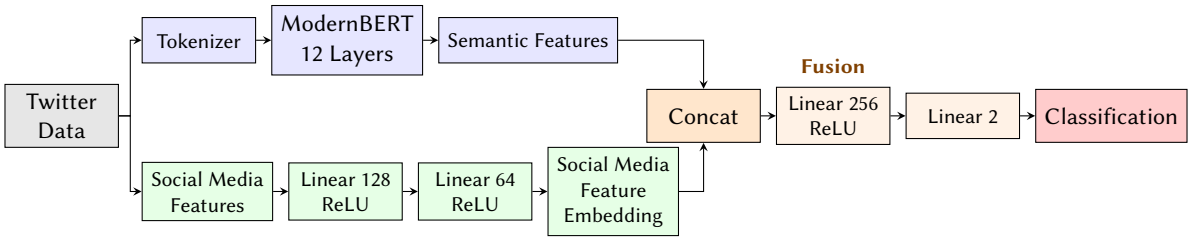

Figure 1: Model architecture integrating semantic features and social media features for misinformation detection.

4.1. Model Architecture

Our hybrid model consists of three main components as shown in Figure 1:

1. Semantic Features (blue blocks in Figure 1): We use ModernBERT-base as the backbone for encoding textual content. The encoder consists of 12 transformer layers, each containing multi-head self-attention and a feed-forward network with residual connections and layer normalization. Input texts are first processed by the tokenizer with a maximum length of 256 tokens, then encoded through ModernBERT. We extract the [CLS] token representation from the final layer as a 768-dimensional text embedding. During training, we fine-tune the ModernBERT weights end-to-end along with the feature fusion layers.

2. Social Media Features (green blocks in Figure 1): We extract 24 hand-crafted features from the datasets and categorize them into three groups: (i) **Text-based features (12)** capturing statistical and stylistic properties including text length, word count, exclamation count, question count, ellipsis count, uppercase/digit ratios, URL/mention/hashtag counts, average word length, and emotion word count; (ii) **User-based features (7)** characterizing the content publisher including verification status, follower count, friends count, follower-to-friends ratio, status count, account age, and profile description length; (iii) **Social engagement features (5)** reflecting post reception including retweet count, favorite count, retweet status, total engagement, and retweet ratio. All features are standardized using StandardScaler for zero mean and unit variance. These features are then processed through a feed-forward network: a 128-dimensional linear layer with ReLU activation and Dropout (0.3), followed by a 64-dimensional linear layer with ReLU activation and Dropout (0.2), producing the final feature representation.

3. Fusion Layer (Final Classification): We concatenate the 768-dimensional BERT embeddings and 64-dimensional feature embeddings using a Concat layer, resulting in an 832-dimensional fused representation. This concatenated representation is processed through: a 256-dimensional linear layer with ReLU activation and Dropout (0.3), followed by a 2-dimensional linear layer, outputting logits for binary classification.

4.2. Focal Loss Function

To handle class imbalance during training, we employ Focal Loss [13]:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

where p_t is the model's estimated probability for the true class, α_t is a class-dependent weighting factor, and γ is the focusing parameter. We set $\alpha = 0.97$ and $\gamma = 2.5$ based on preliminary experiments.

4.3. Threshold Optimization

After model training, we optimize the threshold on the validation set. Instead of using the default threshold of 0.5, we perform a systematic grid search over thresholds ranging from 0.01 to 0.99 in

increments of 0.01. For each threshold candidate, we evaluate the model's performance.

5. Experiment

The organizers released two test sets with different feature availability. The first test set includes comprehensive features, while the second test set only includes text content. To accommodate these differences, we trained two separate models: the first model combines all social media features with semantic features, while the second model combines only text-based features with ModernBERT's semantic features. Additionally, compared to the first model, the second model uses 13 optimized text-based features, replacing emotion word count with two new features: multiple exclamation count and all-caps word count. Moreover, the second model employs a smaller feature extraction network with dimensions of 64 and 32, compared to the first model's 128 and 64 dimensions. Consequently, the fusion layer of the second model concatenates 768-dimensional BERT embeddings with 32-dimensional feature embeddings, resulting in an 800-dimensional representation instead of the 832 dimensions used in the first model.

5.1. Experimental Setup

Table 2 and Table 3 summarize the key training hyperparameters. The first model is trained with early stopping based on the binary F1-score on the validation set, while the second model is trained with early stopping based on the weighted F1-score on the validation set.

Table 2
Training Hyperparameters for the First Model

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	32
Number of Epochs	12
Optimizer	AdamW
Weight Decay	0.01
Maximum Sequence Length	256
Focal Loss Alpha	0.97
Focal Loss Gamma	2.5

Table 3
Training Hyperparameters for the Second Model

Hyperparameter	Value
Learning Rate	8e-6
Batch Size	64
Number of Epochs	8
Optimizer	AdamW
Weight Decay	0.01
Maximum Sequence Length	256
Focal Loss Alpha	0.97
Focal Loss Gamma	2.5

5.2. Data Preprocessing

5.2.1. Text Cleaning

Raw social media text contains noise that can hinder model performance. Our text cleaning pipeline includes: (i) removal of URLs and web links, (ii) removal of special characters while preserving linguistic characters, (iii) normalization of whitespace, and (iv) conversion to lowercase. For missing or empty text fields, we use a placeholder "`empty_text`" to maintain data structure integrity.

5.2.2. Missing Value Handling

Due to field inconsistencies between the misinformation and non-misinformation training sets, we handle missing fields with appropriate default values: for categorical features, we use `False` as the default value when the field is absent; for numerical features, we use `0` as the default value.

5.2.3. Resampling Strategy

To address the severe class imbalance, we implement a resampling approach. The resampling is performed once before training for both models, and the resampled dataset remains fixed throughout all training epochs. For the first model, we upsample misinformation examples by a factor of 10 using random sampling with replacement, while for the second model, we upsample misinformation examples by a factor of 15 using random sampling with replacement.

5.3. Training Progress

Table 4 shows the first model's performance on the validation set across different epochs. The best model is selected at Epoch 2, achieving an F1-score of 0.1834 with precision of 0.1141 and recall of 0.4679.

Table 4
Training Progress on Validation Set for the First Model

Epoch	Precision	Recall	F1(Misinformation)
1	0.0545	0.8333	0.1022
2	0.1141	0.4679	**0.1834**
3	0.1101	0.4615	0.1778
4	0.1050	0.1474	0.1227
5	0.1081	0.2564	0.1521

Table 5 shows the second model's performance on the validation set across different epochs. Although the overall weighted F1-score continued to improve, we observed a decline in recall for the misinformation class on the validation set. After careful consideration, we selected the model trained for two epochs as the optimal model.

5.4. Threshold Optimization Results

For the first model, through systematic grid search on the validation set, we identified the optimal classification threshold as **0.69**. Table 6 presents the model performance on the validation set using this optimized threshold. The threshold optimization improves the F1-score from 0.1834 (at the default 0.5 threshold) to 0.2151.

For the second model, we did not use the threshold optimization strategy.

Table 5
Training Progress on Validation Set for the Second Model

Epoch	Precision	Recall	Weighted-F1
1	0.0464	0.7244	0.9037
2	0.0679	0.5962	0.9435
3	0.0986	0.3077	0.9723
4	0.0900	0.1731	0.9771
5	0.0821	0.1090	0.9793
6	0.0734	0.1026	0.9788
7	0.0833	0.1026	0.9797
8	0.0856	0.1026	0.9799

Table 6
Model Performance on Validation Set with Optimized Threshold

Metric	Value
F1(Misinformation)	0.2151
Precision	0.1435
Recall	0.4295
Optimal Threshold	0.69

6. Results

Table 7 presents the comprehensive evaluation metrics for our first model on the first test set. Our hybrid model achieves a weighted F1-score of 0.97 on the test set with precision of 0.16 and recall of 0.43 for the misinformation class.

Table 7
Final Model Performance on the First Test Set

Metric	Value
Weighted F1-score	0.97
Precision	0.16
Recall	0.43

Table 8 presents the comprehensive evaluation metrics for our second model on the second test set. Our model achieves a weighted F1-score of 0.82 on the test set with precision of 0.82 and recall of 0.84.

Table 8
Final Model Performance on the Second Test Set

Metric	Value
Weighted F1-score	0.82
Precision	0.82
Recall	0.84

Table 9 shows the final ranking for the second test set, displaying the precision, recall, and weighted F1 scores of all participating teams. Our team (whiteby) has an F1 weighted score of 0.82.

7. Conclusion

This paper presents a hybrid deep learning approach for misinformation detection that combines ModernBERT with hand-crafted features derived from textual content, user metadata, and social

Table 9
Performance of all participants on the leaderboard

Participant	Precision	Recall	Weighted-F1
ClimateSense	0.91	0.91	0.91
Sarang	0.90	0.91	0.90
pratikpriyanshu	0.92	0.91	0.89
deepish	0.91	0.89	0.88
priyam_saha17	0.87	0.80	0.82
whiteby(Ours)	**0.82**	**0.84**	**0.82**
sushma03	0.82	0.80	0.85
wangkongqiang	0.78	0.83	0.78
gokul_n_v	0.78	0.69	0.72

engagement metrics. Our methodology addresses the challenges of semantic understanding and extreme class imbalance through the use of Focal Loss and strategic resampling techniques. The experimental results demonstrate that our approach achieves weighted F1-scores of 0.97 and 0.82 on the two official test datasets, respectively.

Acknowledgments

This work is supported by the National Social Science Foundation of China (24BYY080).

Declaration on Generative AI

During the preparation of this work, the authors used Claude in order to assist with English language refinement and improve paper structure and presentation. The authors did not use generative AI for the core research methodology, experimental design, or result analysis. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Comput. Surv. 53 (2021) 109:1–109:40. doi:10.1145/3395046.

[2] 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.

[3] 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.

[4] S. Feng, R. Banerjee, Y. Choi, Syntactic stylometry for deception detection, in: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers, Association for Computational Linguistics, 2012, pp. 171–175.

[5] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, R. Mihalcea, Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, Association for Computational Linguistics, 2018, pp. 3391–3401.

[6] 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, SMSociety 2018, Copenhagen, Denmark, July 18-20, 2018, ACM, 2018, pp. 226–230. doi:10.1145/3217804.3217917.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[8] 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, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[9] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in: Proceedings of the 9th International Conference on Learning Representations (ICLR 2021), OpenReview.net, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[10] C. Castillo, M. Mendoza, B. Poblete, Information Credibility on Twitter, in: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011, ACM, 2011, pp. 675–684. doi:10.1145/1963405.1963500.

[11] 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.

[12] 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.

[13] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324.