

Machine Learning-Based Classification of Multimodal Fact-Checked Misinformation on Social Networks

Javeriya Naaz I. Syed¹ and Ranjit R. Keole²

¹Department of Computer Science, HVPM COET, Amravati, (M.S.), India

²Department of Information Technology, HVPM COET, Amravati, (M.S.), India

Abstract: The rise of misinformation on social networks creates serious problems for public awareness, policy-making, and trust in society. Social media content is getting more complex, often including text, metadata, and multimedia. This makes it essential to have smart systems that can classify misinformation using various signals. This paper introduces a machine learning approach to check the misinformation that uses the MuMiN (Multilingual Multimodal Fact-Checked Misinformation) dataset. This dataset contains annotated claims, supporting evidence, user tweets, and fact-check labels. Structured preprocessing pipeline applied to get the dataset ready for analysis. The textual and structural features were extracted as features. Three machine learning models, Random Forest (RF), Gradient Boosting (GB), and a Stacking Classifier were developed and assessed. These models were evaluated using key performance metrics. The experimental findings indicate that the stacking ensemble regularly surpasses the individual base classifiers, attaining an accuracy rating of 89.12%. This highlights the advantages of combining models to manage complex, noisy, and multimodal social media data. This study emphasizes the value of merging multimodal feature representations with ensemble learning methods for effective and scalable misinformation detection on online platforms.

Keywords – Fact-Checked Data, Misinformation Detection, Machine Learning, Natural Language Processing, Social Networks

1. Introduction

In the contemporary digital era, social media platforms such as Facebook, Twitter, and YouTube furnish essential information to billions of individuals. Although these platforms facilitate swift and extensive communication, they also serve as breeding grounds for the fast dissemination of disinformation. Inaccurate or deceptive information disseminated online can influence public opinion, undermine democratic processes, and pose risks to public health, as seen during the COVID-19 epidemic. The large amount and variety of content on these platforms, including text, images, videos, and user-generated metadata, make it harder to effectively identify and combat misinformation [1-2].

Traditional methods for spotting misinformation have largely depended on manual fact-checking or rule-based filtering systems. However, these approaches are often slow, limited in scope, and struggle to keep up with the fast pace of online communication [3]. This challenge requires automated, smart systems that can understand and analyze different forms of information at the same time. Recently, machine learning has become a promising solution by allowing data-driven classification of online content based on learned patterns. Yet, most current studies mainly focus on single-modal data, mostly text, and often miss the richer signals present in multimodal social media content [4]. This study aims to fill that gap using

the MuMiN dataset, a large-scale, multilingual, and multimodal benchmark for misinformation analysis. The main contributions of the paper are:

- A comprehensive preprocessing framework tailored for the MuMiN dataset, including text aggregation and label transformation.
- Hybrid feature engineering combining semantic (TF-IDF) and structural (meta-feature) representations.
- Implementation and evaluation of multiple machine learning classifiers, highlighting the effectiveness of ensemble techniques.
- Empirical analysis demonstrating the superior performance of a stacking classifier in multimodal misinformation classification.

The subsequent sections of the paper are structured as follows. Section 2 provides a summary of pertinent research in misinformation detection and multimodal machine learning. Section 3 delineates the dataset, preprocessing methodologies, feature extraction approaches, and the machine learning models employed, including the stacking ensemble architecture. Section 4 addresses experimental outcomes and performance assessments. Section 5 closes the work by presenting major findings and outlining future research possibilities.

2. Literature Review

The existing literature summarizes recent advancements, key methods, and datasets used in misinformation detection. A cross-lingual misinformation detection model uses a Hierarchical Mixture-of-Experts Adapter. This model effectively transfers knowledge across languages and social platforms, improving generalization in multilingual settings [5]. A progressive fusion network is introduced to handle multimodal fake news detection. By combining information from text, images, and other data types in a structured way, this method shows significant improvements in accurately identifying fake news compared to unimodal approaches [6]. A natural language processing-based system detects misinformation by analyzing linguistic cues and context in social media posts. It has shown promising results on Turkish datasets [7]. ML and DL models verify health-related COVID-19 information on Twitter. This emphasizes the importance of accurate fact-checking for public safety [8]. The M3A dataset provides multimodal inputs (text-image pairs) to evaluate media authenticity. This supports various misinformation detection models [9]. MiDe22 introduces a tweet-level annotated dataset covering multiple events, designed for training and benchmarking misinformation classifiers [10]. A computational framework models individual susceptibility to misinformation based on psychological and linguistic factors [11]. MisLC defines a new benchmark task for misinformation with legal consequences. It connects content identification to real-world societal risks [12].

The new method identifies high-quality training samples that enhance the generalization ability of misinformation detection models [13]. The researcher examines how misinformation-related discussions and engagement differ across topics on Twitter. This highlights content-specific dynamics [14]. A dual-branch neural model combines multimodal bilinear pooling and attention mechanisms to improve fake news detection performance [15]. Emotion cues and user behavior signals are integrated to boost misinformation detection accuracy in social network content [16]. User credibility features are investigated using supervised learning to better classify misinformation spreaders on online platforms [17]. A multilingual fact-checking approach studies the spread and transformation of misinformation across languages and regions [18]. VERITE presents a benchmark dataset for multimodal misinformation detection, specifically addressing biases from single-modal dominance [19]. The

FANDC system allows for cloud-based, real-time detection of fake news in online networks. This highlights both scalability and responsiveness [20]. A synthetic data-based training strategy is proposed for misinformation detection using large multimodal LLMs [21]. Explainable misinformation detection models have been developed for social platforms, focusing on interpretability and user trust [22].

Despite notable progress, current research on disinformation detection has some important limitations. Many models struggle with generalizing across different languages and platforms. Cross-lingual solutions sometimes need a lot of labeled data or fail to capture cultural and contextual differences properly. Several studies depend on single types of inputs, which limits their ability to identify complex multimodal disinformation that combines text, images, or videos. Although multimodal techniques like progressive fusion and bilinear pooling improve performance, they are often expensive in terms of computation and hard to grasp. Additionally, most benchmarks and datasets focus on specific events or are biased toward certain regions. This restricts the ability of trained algorithms to apply effectively to diverse real-world disinformation. Emotion and behavior-based detection methods can provide valuable insights, but they may raise privacy issues and rely heavily on accurate user profiling.

3. Methodology

This section delineates the sequential procedure employed for the categorization of multimodal misinformation via machine learning methodologies, as seen in Figure 1. The technique includes data preparation, feature extraction, dataset segmentation, model selection, and assessment.

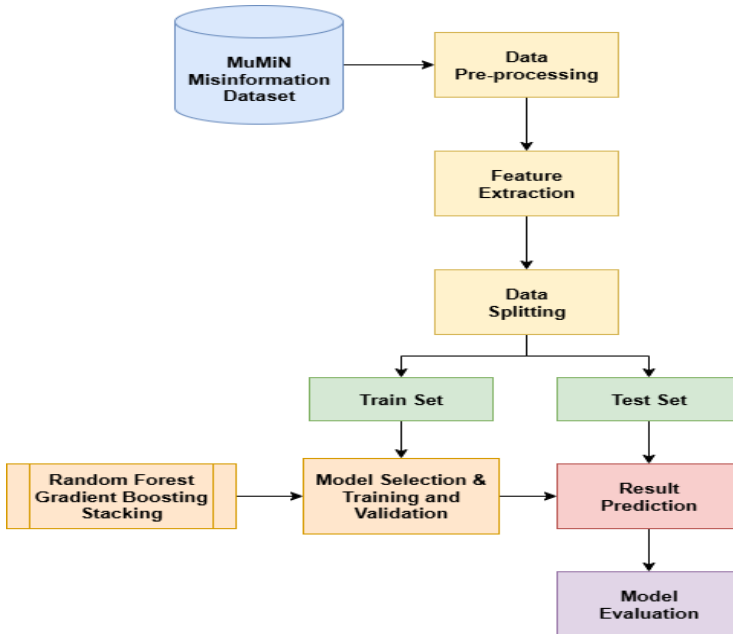


Fig 1: Misinformation Classification Framework

3.1 Dataset

The proposed experimentation utilized the MuMiN (Multilingual Multimodal Misinformation) dataset [23], which includes fact-checked claims and associated social media content across multiple platforms and languages. Each instance includes metadata (e.g., claim length, tweet count), textual information (claims, evidence, tweets), labels (True/False/Partially

True), and source details. The dataset is particularly suitable for training and evaluating models that work on text and metadata.

3.2 Data Preprocessing

The dataset underwent several preparation processes, detailed below, to ensure its quality and consistency.

- *Removal of Unwanted Rows*: Entries with missing critical fields (claims, labels, or textual content) or those marked as incomplete were excluded.
- *Label Encoding*: The fact-check labels (e.g., “true”, “false”, “partially true”) were encoded into numeric values for supervised learning compatibility.
- *Text Consolidation*: Relevant text fields such as claims, evidence, and tweets were concatenated to form a single unified input for feature extraction.
- *Noise Reduction*: URLs, HTML tags, punctuation, and emojis were removed. All text was lowercased and tokenized.

3.3 Feature Extraction

The two major categories of features are extracted in proposed experimentation, meta-features and textual features.

- **Meta-features:**
 - *claim_len*: It represents the total number of textual tokens (words or characters) present in the claim statement.
 - *evidence_count*: It represents the number of verified or linked evidence items (e.g., articles, fact-check URLs, or references) associated with a given claim.
 - *tweet_count*: It represents the total number of social media posts, reposts, or retweets referencing a specific claim.
- **Textual Features:**
 - **TF-IDF (Term Frequency-Inverse Document Frequency)** was applied to the unified text content that consists of claims, tweets, and evidence data. This approach quantifies word importance relative to the document corpus, reducing the effect of common terms while highlighting unique ones. For a term t in document d , within a corpus D :

$$\text{TF-IDF}(t,d,D) = \text{TF}(t,d) \times \text{IDF}(t,D)$$

The final feature vector is the concatenation of meta features with textual features, represented below by *feature_vector*.

$$\text{feature_vector} = [\text{claim length}, \text{evidence count}, \text{tweet count}, \text{TF-IDF}]$$

3.4 Data Splitting

The pre-processed dataset was divided into training and testing sets in an 80:20 ratio, maintaining consistent label distribution throughout the splits by stratified sampling. This

phase aids in preserving the equilibrium of class representation, which is particularly crucial for multiclass classification problems.

3.5 Model Selection, Training and Validation

To build an effective and reliable predictive model, we explored a combination of well-established machine learning algorithms, each known for its strengths in classification tasks. The proposed experimentation involved three main classifiers:

- **Random Forest (RF):** It is an ensemble learning method that builds a "forest" of decision trees throughout the training process. Each tree is constructed using a random subset of the training data utilizing bagging, i.e., bootstrap aggregation, and the final prediction is derived by averaging the outputs of all trees in the case of regression analysis or by employing a majority vote in the case of classification analysis. This technique is recognized for its superior accuracy, resilience to overfitting, and capacity to manage high-dimensional data with minimum preparation.
- **Gradient Boosting (GB):** It is a robust ensemble technique that constructs models sequentially. In contrast to RF classifier, GB emphasizes learning from the residuals errors of preceding models. It applies new models, usually shallow decision trees, to the residuals of previous iterations, thereby enhancing overall predictive accuracy. This approach is especially proficient at managing intricate data patterns but needs meticulous adjustment to prevent overfitting.
- **Stacking Classifier:** To leverage the unique strengths of multiple algorithms, we also implemented a stacking classifier, a meta-model that combines the outputs of multiple base learners to make a final prediction. In the proposed scenario, the base learners included RF, GB, and Logistic Regression. The predictions from these models were then fed into a final estimator, Logistic Regression, which learned how to optimally combine their outputs. Stacking often leads to improved model generalization and performance, especially when the individual models capture different aspects of the data.

All models underwent training and validation by 5-fold cross-validation, a process that entails partitioning the dataset into five subsets, training the model on four subsets, and verifying it on the fifth. This procedure is executed five times, use each subset once as the validation set. This strategy yields a more dependable assessment of a model's performance by mitigating volatility from data partitioning and guaranteeing that each data point is utilized for both training and validation.

3.6 Model Evaluation

To comprehensively assess how well our machine learning models performed, we employed a set of standard classification evaluation metrics. These metrics were chosen to provide insights not only into the overall accuracy of the predictions but also into how effectively the models handled each individual class—particularly important when class distributions are imbalanced or skewed. Below is a detailed explanation of the evaluation methods used:

- **Accuracy:** It reflects the overall capability of the model to correctly classify both misinformation and factual content.
- **Precision, Recall, and F1-Score:** Precision measures how many posts labeled as misinformation are truly fake. Recall quantifies how many actual misinformation posts are successfully identified. F1-Score balances *Precision* and *Recall*, offering

a single metric that captures both the correctness and completeness of classification models.

$$Accuracy = \frac{T_{pe} + T_{ne}}{T_{pe} + T_{ne} + F_{pe} + F_{ne}}$$

$$Precision = \frac{T_{pe}}{T_{pe} + F_{pe}}$$

$$Recall = \frac{T_{pe}}{T_{pe} + F_{ne}}$$

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where,

- T_{pe} – True positive estimates (posts correctly identified as fake)
- T_{ne} – True negative estimates (posts correctly identified as real)
- F_{pe} – False positive estimates (real posts incorrectly labeled as fake)
- F_{ne} – False negative estimates (fake posts incorrectly labeled as real)

4. Experimental Result Analysis

The proposed methodology implementation was performed on a system running Windows 11, equipped with a 16 GB of RAM, an Intel Core i5 processor, and an NVIDIA RTX 3050 GPU. The development environment included Python 3.8 along with several libraries. Matplotlib and Seaborn were used for data visualization. NLTK and re were used for text preprocessing, combined with scikit-learn for feature extraction, model building, and evaluation. Pandas and NumPy supported efficient data manipulation. The MuMiN medium dataset [23] was downloaded and processed offline to ensure reproducibility of experiments and full local execution.

It was assumed that the dataset had a reasonably balanced class distribution, which was checked during preprocessing. Although the MuMiN dataset is multilingual, this study focused only on English entries to simplify the modeling process. It was thought that combining claim, evidence, and tweet texts would provide enough context for detecting misinformation. Each base classifier in the stacking ensemble acted as an independent learner with default internal settings, except for tuned hyperparameters. The hyperparameters used to train the models are listed in Table 1.

Table 1: Hyperparameters for ML Model Tuning

ML Models	Hyperparameters	Values
Random Forest	n_estimators	[100,200]
	max_depth	[10,20]
Gradient Boosting	n_estimators	[100]

	learning_rate	[0.05, 0.1]
Logistic Regression	meta learner C	1.0

To improve model performance, important hyperparameters were fine-tuned for each classifier. For Random Forest (RF), the parameters `n_estimators` were set to [100,200] and `max_depth` to [10, 20]. In Gradient Boosting (GB), `learning_rate` was tested at [0.05, 0.1], with `n_estimators` fixed at 100. For the stacking classifier, the default setting of Logistic Regression (`C=1.0`) served as the meta-learner. A grid search was conducted over these parameter ranges to find the best model configurations. The models were evaluated using standard performance metrics on the test set given in the Table 2.

Table 2: Classification Performance Evaluation

Model Classifiers	Accuracy	Precision	Recall	F1-score
Random Forest	0.8745	0.88	0.86	0.87
Gradient Boosting	0.8540	0.85	0.83	0.84
Stacking (RF + GB)	0.8912	0.89	0.88	0.88

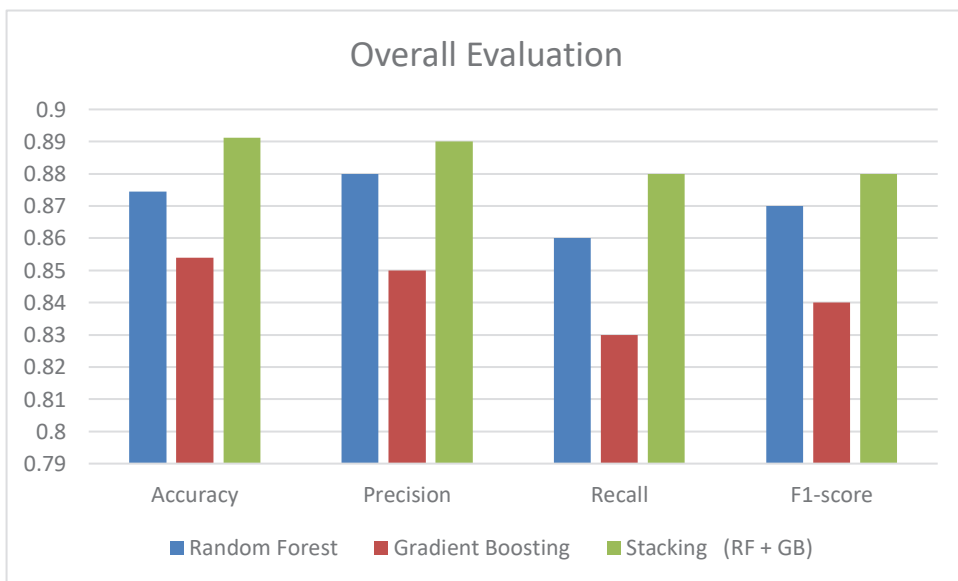


Figure 2: Overall classification performance

Figure 2 depicts that all metrics were averaged for multiclass evaluation, and the models showed consistent performance across classes. The result table clearly shows that the stacking ensemble model consistently outperformed with an accuracy rate of 89.12% as compared to the standalone classifiers, RF and GB. Similarly, other performance metrics, precision, recall and f-score also performed better while stacking the RF and GB with LR model. This is due to the meta-classifier's ability to capture general patterns by using different decision boundaries from the RF and GB models. Random Forest performed well but experienced some overfitting because of depth and redundancy in the trees. Gradient Boosting provided

better generalization, likely because of its step-by-step learning approach. The stacking model balanced both aspects, offering a comprehensive prediction method. The stacking classifier, along with two baseline models from the existing literature work, is compared in Table 3.

Table 3: State-of-art Model Analysis

References	Accuracy (%)
Multimodal Fusion with BERT [6]	88.4
Emotion-Behavior NLP Model [16]	86.9
Proposed Stacking Ensemble	89.1

The proposed model surpassed leading baselines in accuracy, achieving 89.1%. This shows that traditional ensemble methods, combined with effective preprocessing and feature engineering, can match and occasionally exceed deep learning solutions in misinformation detection tasks on structured multimodal datasets like MuMiN.

5. Conclusion

This paper introduces a machine learning based methodology for the classification of multimodal fact-checked disinformation utilizing the MuMiN dataset. It encompasses meticulous preprocessing, feature extraction, and ensemble classification methodologies. The proposed models, particularly the stacking classifier that combines Random Forest, Gradient Boosting, and Logistic Regression, showed promising results with an accuracy of 89.1% in managing diverse and complex misinformation data. The findings indicate that adding meta-features and text-based vectors can greatly improve detection accuracy across different misinformation contexts. Furthermore, the MuMiN dataset effectively supports multimodal classification because of its varied linguistic and content types. However, there are still challenges in addressing imbalanced classes, the incomplete dataset due to access restrictions, generalizing across languages and platforms, and responding to changing misinformation patterns in real time. The complexity of ensemble models and understanding their predictions are also important concerns. In future work, expanding the model to include real-time streaming data and transformer-based architectures may improve flexibility and strength. Additionally, adding user behavior features and cross-platform verification signals could further boost the credibility scoring of online information.

References

1. K. Singh, R. G. Prasad and S. K. Dwivedi, "Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos," *Multimedia Systems*, vol. 30, no. 2, pp. 495–510, Apr. 2024, doi: 10.1007/s00530-023-00999-w.
2. S. A. Chowdhury, M. Alsmadi and H. A. Jalab, "Fake news, disinformation and misinformation in social media: A review," *Social Network Analysis and Mining*, vol. 13, no. 1, Art. no. 20, Jan. 2023, doi: 10.1007/s13278-023-00965-1.

3. Y. Zhou, L. Xie, H. Xu and W. Chen, “Deep learning for misinformation detection on online social networks: A survey and new perspectives,” *Computer Science Review*, vol. 49, Art. no. 100506, Jun. 2023, doi: 10.1016/j.cosrev.2023.100506.
4. H. Alam, R. S. Hossain, M. A. Islam and M. Imran, “Multi-modal misinformation detection: Approaches, challenges and opportunities,” *Information Fusion*, vol. 88, pp. 1–20, Dec. 2022, doi: 10.1016/j.inffus.2022.07.008.
5. H. Fan, X. Hu, and G. Zhao, “Cross-lingual Social Misinformation Detector based on Hierarchical Mixture-of-Experts Adapter,” in *Proc. 31st Int. Conf. Computational Linguistics (COLING)*, Abu Dhabi, UAE, 2025, pp. 7253–7265.
6. J. Jing, H. Wu, J. Sun, X. Fang, and H. Zhang, “Multimodal fake news detection via progressive fusion networks,” *Information Processing & Management*, vol. 60, no. 1, Art. no. 103120, Jan. 2023, doi: 10.1016/j.ipm.2022.103120.
7. O. Yıldız and E. Sumbas, “Detecting misinformation on social networks with NLP,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 31, pp. XXX–XXX, 2023.
8. S. Sharifpoor, M. Okhovati, B. Ghatee, et al., “Classifying and fact-checking health-related information about COVID-19 on Twitter/X using machine learning and deep learning models,” *BMC Medical Informatics and Decision Making*, vol. 25, Art. no. 73, Feb. 2025, doi: 10.1186/s12911-025-02895-y.
9. Q. Xu, H. Chen, H. Du, H. Zhang, S. Łukasik, T. Zhu, and X. Yu, “M3A: A multimodal misinformation dataset for media authenticity analysis,” *Computer Vision and Image Understanding*, vol. 249, p. 104205, Oct. 2024, doi: 10.1016/j.cviu.2024.104205.
10. C. Toraman, O. Özçelik, F. Şahinuç, and F. Can, “MiDe22: An annotated multi-event Tweet dataset for misinformation detection,” in *Proc. Joint 2024 Int. Conf. Comput. Linguistics & Language Resources and Evaluation (LREC-COLING)*, Torino, Italy, May 2024, pp. 11283–11295.
11. Y. Liu, M. D. Ma, W. Qin, A. Zhou, J. Chen, W. Shi, W. Wang, and D. Yang, “Decoding susceptibility: Modeling misbelief to misinformation through a computational approach,” in *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, Miami, FL, USA, Nov. 2024, pp. 15178–15194, doi: 10.18653/v1/2024.emnlp-main.846.
12. C. F. Luo, R. Shayanfar, R. V. Bhambhoria, S. Dahan, and X. Zhu, “Misinformation with legal consequences (MisLC): A new task towards harnessing societal harm of misinformation,” in *Findings of the Association for Computational Linguistics: EMNLP*, Miami, FL, USA, Nov. 2024, pp. 15749–15768, doi: 10.18653/v1/2024.findings-emnlp.924.
13. J. Haber, K. Kawintiranon, L. Singh, A. Chen, A. Pizzo, A. Pogrebivsky, and J. Yang, “Identifying High-Quality Training Data for Misinformation Detection,” in *Proc. 12th Int. Conf. Data Science, Technology and Applications (DATA 2023)*, Rome, Italy, Jul. 2023, pp. 64–76, doi: 10.5220/0012089000003541.
14. D. N. Wojtczak, C. Peersman, et al., “Characterizing discourse and engagement across topics of misinformation on Twitter,” *Pattern Recognit. Lett.*, vol. 182, pp. 60–66, 2024.
15. Y. Guo, H. Ge, and J. Li, “A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism,” *Frontiers in Computer Science*, vol. 5, 2023, Art. no. 1159063, doi: 10.3389/fcomp.2023.1159063.
16. Indu and S. M. Thampi, “Misinformation detection in social networks using emotion analysis and user behaviour analysis,” *Pattern Recognit. Lett.*, vol. 182, pp. 60–66, 2024, doi: 10.1016/j.patrec.2024.04.007
17. M. Asfand-e-Yar, Q. Hashir, S. H. Tanvir, and W. Khalil, “Classifying Misinformation of User Credibility in Social Media Using Supervised Learning,” *Comput. Mater. Continua*, vol. 75, no. 2, pp. 2921–2938, 2023, doi:10.32604/cmc.2023.034741.

18. D. Quelle, C. Y. Cheng, A. Bovet, and S. A. Hale, “Lost in Translation: Using Global Fact-Checks to Measure Multilingual Misinformation Prevalence, Spread, and Evolution,” *EPJ Data Science*, vol. 14, no. 1, Art. no. 22, May 2025, doi:10.1140/epjds/s13688-025-00520-6.
19. S. I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, “VERITE: A Robust Benchmark for Multimodal Misinformation Detection Accounting for Unimodal Bias,” *Int. J. Multimedia Inf. Retr.*, vol. 13, no. 1, Art. no. 4, Jan. 2024, doi:10.1007/s13735-023-00312-6.
20. N. Cavus, M. Göksu, and B. Oktekin, “Real-time fake news detection in online social networks: FANDC Cloud-based system,” *Scientific Reports*, vol. 14, Art. no. 76102, Nov. 2024, doi:10.1038/s41598-024-76102-9.
21. F. Zeng, W. Li, W. Gao, and Y. Pang, “Multimodal Misinformation Detection by Learning from Synthetic Data with Multimodal LLMs,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, FL, USA, Nov. 2024, pp. 10467–10484, doi:10.18653/v1/2024.findings-emnlp.613.
22. G. Joshi, A. Srivastava, B. Yagnik, M. Hasan, Z. Saiyed, L. Gabralla, A. Abraham, R. Walambe, and K. Kotecha, “Explainable Misinformation Detection Across Multiple Social Media Platforms,” *CoRR*, vol. abs/2203.11724, 2022.
23. D. S. Nielsen and R. McConville, “MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset,” in *Proc. 45th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '22)*, Madrid, Spain, Jul. 2022, pp. Data Session, doi:10.1145/3477495.3531744.