© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717327

Early Detection of Fake News Using Multimodal Machine Learning

ISHA PALLAVI BARA[1], DR. SYED SHAHID RAZA[2]
[1]*MBA, Dept. of Business Analytics, CMS Business School, JAIN (Deemed-to-be University), Bengaluru, India*
[2]*Professor, Dept. of Business Analytics, CMS Business School, JAIN (Deemed-to-be University), Bengaluru, India*

Abstract*—The rapid growth of social media platforms has significantly increased the spread of fake news, posing serious threats to public trust, democratic processes and social harmony. Traditional fake news detection methods relying solely on textual analysis are often insufficient as modern misinformation is increasingly multimodal, combining text, images, videos and metadata to enhance credibility and emotional impact. This thesis proposes a multimodal machine learning framework for the early detection of fake news by jointly analysing textual, visual and contextual features. By integrating natural language processing techniques with deep visual feature extraction and multimodal fusion strategies the proposed approach aims to improve detection accuracy at early stages of news dissemination. Experimental evaluation on benchmark datasets demonstrates that multimodal models outperform unimodal approaches highlighting the importance of cross modal correlations in identifying deceptive content.

Keywords*— Fake News Detection, Multimodal Machine Learning, Social Media Analysis, Deep Learning, Misinformation, Early Detection

I. INTRODUCTION

Social media platforms like twitter, facebook, Instagram and countless others have fundamentally reshaped how information travels across society, where news once flowed through gatekept channels with editorial oversight, it now spreads laterally through networks of users who both consume and create content. This transformation has brought undeniable benefits for global connectivity and democratic participation yet it has also opened pathways for misinformation to circulate with unprecedented speed and scale. While this connectivity fosters global communication and community building it has also created a fertile ground for the rapid and unchecked spread of misinformation, commonly known as fake news. Fake news defined as fabricated information that mimics new media content in form but not in organizational process or intent poses a grave danger to societal well being. Its consequences range from manipulating public opinion and eroding trust in journalistic institutions to inciting real world violence and undermining public health initiatives as starkly illustrated during the COVID-19 infodemic (Shu-2020). The sheer volume and velocity of data on social media make manual fact checking infeasible creating an urgent need for automated detection systems.

Initial efforts to combat fake news using machine learning focused predominantly on textual cues. These methods analyse linguistic styles headline sensational , source credibility and propagation patterns to identify deceptive content (Shu- 2019, Zhang &Ghorbani 2020). However contemporary fake news is increasingly multimodal. Purveyors of disinformation pair misleading or fabricated text with emotionally charged or doctored images to enhance persuasiveness and bypass textual filters. A user might be more inclined to believe and share a post with a seemingly authentic, shocking image even if the accompanying text is vague or false. This evolution from unimodal to multimodal deception necessitates a corresponding evolution in detection methodology.

Researchers have increasingly turned to multimodal machine learning as a potential solution. Rather than examining text alone or images alone these approaches attempt to replicate how humans actually process information by synthesizing signals across different channels simultaneously. When a news post pairs a sensational headline with a photograph, a multimodal model can assess not just each element independently but the relationship between them- Does the image actually show what the text claims, has this image appeared before in different contexts, The capacity to ask such cross modal questions

represents a qualitative leap beyond earlier detection methods that capture inconsistencies and patterns invisible to unimodal models. For instant and image may be manipulated or taken out of context a discrepancy a multimodal model can flag by comparing its sematic content with accompanying text (Alam 2021). The growing body of research in this area, such as the works collected in the special issue on "Multimodal Fake News Analysis" (Alam-2022) underscores its critical importance.

II. REVIEW OF LITERATURE

Theme 1: Foundational and Unimodal Approaches to Fake News Detection

Shu, K. Wang & Liu H. (2019) Beyond news contents: The role of social context for fake news detection. Proceeding of the Twelfth ACM International Conference on Web Search and Data Mining, 312-320. (17) – This foundational work highlights the limitation of content only models and introduces the importance work highlights the limitations of content only models and introduces the importance of social context features, such as user profiles and network structures for improving detection accuracy. It provides a benchmark for comparing purely content based models against those using social context, setting the stage for arguments about early detection reliance on content.

Zhang, X & Ghorbani A.A(2020) An overview of online fake news: Characterization, detection and discussion. Information processing & Management, 57(2), 102025(5)- This comprehensive review categorizes fake news detection methods into content based (linguistic, visual) and context based (user, network). It systematically outlines the evolution of the field and underscores the increasing sophistication of fake news, paving the way for multimodal and early early stage approaches.

Shu, k. Sliva , Wang S (2020). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19 (1) (19)- A seminar paper that formalizes the fake news detection problem from a data mining perspective. It proposes a holistic view encompassing news content, social media and temporal dynamics establishing a key theoretical framework for subsequent studies.

Kaliyar, R.K Goswami A & Narang (2021). FNDNet a deep convolutional neural network for fake news detections. Cognitive Systems Research, 61, 32-44(18)- Demonstrates a pure deep learning approach using CNNs on textual data. While unimodal it shows the power of automated feature extraction a principle that extends to multimodal models. This serves as a baseline for comparing the added value of incorporating a visual modality.

Theme 2: The Rise of Multimodal Fake News Detection

Singh, V.K, Ghosh S. & Jose J.M (2021). Toward multimodal fake news detection: A systematic review. Journal of the Association for Information Science and Technology, 72(12)- A critical systematic review that maps the landscape of multimodal fake news detection. It categorizes existing techniques, datasets and challenges providing a structured overview of the field evolution and identifying key research directions including early detection and cross modal coherence.

Boididou, C, Papadopoulos, S Zampoglou, M Apostolidid L papadopoulou & (2018). Detection and verification of misinformation in social media. Multimedia Tool and Application , 77(10) – An early and influential work that tackles misinformation detection in multimedia content. It explores the use of both visual and textual verification techniques, laying the groundwork for integrated multimodal systems. The challenges it identifies such as out of context images, remain central.

Alam, F. Dalvi, Imran M & Ofli F(2022). Multimodal fake news analysis. Entropy , 24 (3)- This special issue introduction brings together cutting edge research in the field. It emphasizes the need for models that can reason modalities , moving beyond simple concatenation of features to more sophisticated fusion and co attention mechanisms.

Khatter, D, Garg A & Varma, V (2019). MVAE: Multimodal variational autoencoder for fake news detection. The World Wide web Conference, 2915-2921(9)- Introduces a novel deep learning architecture, MVAE that uses a variational autoencoder to learn a shared representation of text and image. This approach aims to capture the deep semantic relationship between modalities a concept highly relevant for detecting inconsistencies.

Wang, Y Ma, F, Jin Z EANN: Event adversarial neural networks for multi modal fake news detection. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data

Mining, 849-857 (8)- Proposes an Event Adversarial Neural Network (EANN) designed to learn event invariant features. This is crucial for detecting fake news on emerging event, where the model hasn't been trained on similar topics, directly addressing a key challenge for early stage detection.

Theme 3: Towards Early Detection and Temporal Dynamics

Zhou, X Zafarani R Shu K & Liu(2022). Early detection of fake news on social media: A review Information Processing & Management, 59(Link 18)- A direct and highly relevant review that focuses specifically on the challenge of early detection. It categorizes methods based on the type of data they use (content, user, network) and analyses their suitability for the early stages of propagation.

Liu Y, X & Zhang(2024). Towards early fake news detection: A time aware multimodal approach. IEE Transactions on Knowledge and data Engineering(16)- Presents a cutting edge approach that explicitly models the temporal evolution of multimodal content. It likely incorporates time decay functions or recurrent structures to weigh early signals more heavily, representing the state of the art this thesis aims to build upon.

Ma, J Gao& Wang K.F(2023). Detect rumours in microblog posts: A temporal social Multiview approach. Processing of the International AAI Conference on Web and Social Media, 17(link13)- While focusing on rumours this paper Multiview temporal approach is highly relevant. It demonstrates the power of combining different data views (like text and user dynamics) over time, a concept that can be adapted to combine text, image and early propagation signals.

Theme 4: Addressing Modality Specific Challenges

Giachanou, A Rosso P & Crestani F(2020). Leveraging emotional signals for fake news detection. Information Processing & Management, 57(6)(link 7)- Investigations the role of emotional cues in textual content for deception detection. It highlights that fake news often uses high arousal, emotionally charged language a finding that can be extended to the visual domain where images are also chosen for their emotional impact.

Gupta P, Sharma& Jindal V(2024). Exposing AI generated fake images in multimodal misinformation. Recent advances in Computer Science and Communication, 17(5)(link 15- Addresses the modern challenge of AI generated imagery (deepfakes) in fake news. As synthetic media becomes mo, (link 17)- This work re prevalent, models must be able to detect visual artifacts making this work crucial for future proffing detection frameworks.

Zlatkova, D Nakov P& Koychev I(2019). Fact-checking meet fauxtography: Verifying claims about images. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing(link 17) - This work specifically tackles the fauxtography problem verifying the authenticity of images and their association with textual claims. It focus on image verification and cross modal consistency is directly applicable to the core task of multimodal fake news detection.

Zhang T, Wang D. Chan H, Zeng Guo W, Miao C, & Cui L.(2022). Multimodal fake news detection with textual and visual cues: A survey Information fusion, (link 19) -A recent and comprehensive survey that provides a detailed taxonomy of multimodal fusion techniques, datasets and future challenges. It is an excellent resource for designing the methodological framework of this thesis.

Theme 5: Advanced Techniques, Fusin and Applications

Wu, Y Zhan, P Zhang , Y Wang & Xu Z (2021). Multimodal fusion with co-attention for fake news detection. Neural Computing and Applications, 33(link 18)- Explores the use of co-attention mechanisms which allow the model to focus on specific parts of the image while reading the text and vice versa. This creates a more dynamic and fine grained understanding of cross modal relationships a sophisticated technique for improving model performance.

Jiang S, Chen X & Xu H (2023). Rumour detection on social media with multimodal dual learning. IEE Transactions on Computational social systems, 10(5).(link 14)- Proposes a dual learning framework that cycles between modalities to reconstruct one from the other. This forces the model to learn deep representations and identify points of incongruence offering a novel and potentially robust for early detection.

Qian, F Gong C Sharma K & Liu Y (2021). Multimodal misinformation detection by learning from synthetic data. Journal of Intelligent Information Systems, 58(2) (Link 20)- Tackles the problem of limited labelled data by exploring the use of synthetic multimodal data for training. This is a highly relevant methodological consideration as creating large high quality datasets for fake news is a persistent challenge.

Sahoo, S.R & Gupta B.(2022). Multiple features based fake news detection in social networks using deep learning. Multimedia tools and application, 81(100 (link 11)- Presents a hybrid deep learning model that combines multiple features for detection. This work is useful for understanding how different types of features (statistical, sematic) can be effectively integrated informing the design of the feature extraction layers in the proposed model.

III. RESEARCH DESIGN AND METHODOLOGY

This study focuses on the early detection of fake news on social media platforms using multimodal machine learning.

Research Hypothesis

Based on the research objectives and theoretical underpinnings the following hypotheses are formulated:

- H1: A multimodal machine learning model that fuses textual and visual features will significantly outperform unimodal (text only or image only) models in detecting fake news. (This tests the core premise of the study, based on the idea that fake news exploits both modalities as per ELM).
- H2: A deep learning model employing a co attention based fusion mechanism will demonstrate higher accuracy I detecting fake news compared to models using simpler fusion methods (eg early or late concatenation). This tests the theoretical need for fine grained modal cross modal reasoning as suggested by cognitive dissonance theory.
- H3: The proposed multimodal framework will achieve significantly higher performance in the early stages of news propagation (eg within the first few hours of posting0 compared to existing baseline models that may implicitly rely on post propagation features.
- H4: The model's performance will be positively correlated with the degree of semantic inconsistency between the text and image in a fake news post. This will test the operationalization of Cognitive Dissonance Theory, where higher dissonance should make detection easier.

Research Design

This study will employ a quantitative, experimental research design. This research is experimental because it involves manipulating the independent variables (the content of the post and the model architecture) to observe their effect on the dependent variable (the accuracy of fake classification). The design is structured to build, train and rigorously test a computational model.

- Research Approach: The study adopts a positivist paradigm, seeking objective and generalizable findings through the empirical evaluation of a proposed model. The approach is deductive as it formulates hypotheses based on existing theories (ELM, Cognitive Dissonance) and prior literature and then tests these hypotheses through experimentation.
- Experimental Setup:

a) Phase 1: Dataset Acquisition and Preprocessing: The research will begin by selecting and acquiring two or more standard publicly available benchmark datasets for multimodal fake news detection (eg Twitter, Weibo, Fakeddit). These datasets will be pre-processed: text will be cleaned, tokenized and formatted for a transformer model, images will be resized, normalized and augmented.

b) Phase 2: Model Developed: The core of the design is the development of the proposed framework. Three primary types of models will be developed as part of the experimental treatment:

- Unimodal Baselines: A text only model (eg fine tuned BERT) and an image only model (eg a pre trained ResNet50).
- Multimodal Baseline: A model that simply concatenates the feature vectors from the text and image models before classification (late fusion).
- Proposed Model: The experimental model featuring a co attention mechanism that allows the text and image representations to interact before fusion. For instance the model could use the text representation to guide attention to relevant parts of the image and vice versa, creating a joint representation that captures cross modal interplay.

c) Phase 3: Training and Validation: The models will be trained on a portion (eg 70%) of the dataset. A validation set (eg 15%) will be used for hyperparameter tuning to optimize model performance and prevent overfitting. Techniques like cross validation will be employed.
d) Phase 4: Hypothesis Testing: The trained models will be evaluated on a held out test set (eg 15%) which the models have not seen during training. The performance metrics (accuracy, precision, recall, F1 score) will be recorded for each model. To specifically test the early detection hypothesis (H3), the test set will be stratified based on post time. The model's performance on posts with low engagement (or within a short time window) will be analyzed separately.

IV. DATA ANALYSIS AND FINDINGS

This study employs quantitative statistical and machine learning techniques to evaluate the feasibility of early fake news detection under constrained conditions. The analysis is conducted using secondary data derived from the Politifact dataset.

The initial stage of analysis involves data preprocessing and feature extraction. The textual data (news titles) undergo a systematic cleaning process, including conversion to lowercase, removal of URLs and non-alphabetic characters, and elimination of common English stop words. Following preprocessing, Term Frequency–Inverse Document Frequency (TF-IDF) vectorization is applied to transform the cleaned text into numerical feature representations. A maximum of 5000 features is selected to balance representational richness with computational efficiency and to reduce the risk of overfitting.

Subsequently, descriptive statistical analysis is performed to summarize key characteristics of the dataset. This includes examining sample size, class distribution (fake vs. real), and general feature properties. These statistics provide an essential overview and help ensure that the dataset is suitable for classification tasks.
The core analytical component of the study involves the application of supervised machine learning classification algorithms. Two models are implemented:
Logistic Regression: A linear classification model that estimates the probability of class membership using the logistic (sigmoid) function. It serves as a baseline model due to its simplicity, interpretability, and strong performance in text classification tasks.
Random Forest Classifier: An ensemble learning method that constructs multiple decision trees and aggregates their predictions through majority voting. This model is particularly effective in capturing non-linear relationships and complex feature interactions within high-dimensional data.
Model performance is evaluated using standard classification metrics to provide a comprehensive assessment:
Accuracy: Overall proportion of correctly classified instances
Precision: Proportion of predicted positive instances that are actually positive
Recall: Proportion of actual positive instances correctly identified
F1-score: Harmonic mean of precision and recall, providing a balanced measure of performance
To further support hypothesis testing, inferential statistical techniques are employed. Specifically, independent samples t-tests are conducted to assess whether there are statistically significant differences in linguistic feature distributions between fake and real news classes.
In addition, feature importance analysis is performed using the Random Forest model to identify the most influential textual features contributing to classification decisions. This provides interpretability and insights into the linguistic patterns associated with misinformation.
All statistical analyses are conducted at a 5% significance level ($\alpha = 0.05$). Results with a p-value less than 0.05 are considered statistically significant, indicating strong evidence against the null hypothesis.

Finding 1: Content-Only Early Detection is Feasible but Limited
The results demonstrate that machine learning models relying solely on news titles can achieve accuracy significantly above chance level. The Random Forest classifier achieved an accuracy of 74.5%, while Logistic Regression achieved 71.2%, with both results being statistically significant ($p < 0.001$).
These findings confirm that early detection is feasible, even in the absence of social context or engagement data. However, the 25.5% error rate of

the best-performing model highlights an important limitation: content-only approaches are not sufficiently reliable for fully autonomous deployment.

Finding 2: Asymmetric Performance is a Persistent and Fundamental Challenge

Both models consistently demonstrated superior performance in detecting real news compared to fake news. This asymmetry reflects the inherent complexity of identifying deceptive content.

Fake news is often intentionally designed to mimic legitimate journalism, adopting similar structures, tone, and vocabulary. In contrast, real news tends to exhibit more consistent patterns of attribution, qualification, and institutional referencing, which models can more easily learn and recognize.

Finding 3: Ensemble Methods Outperform Single Classifiers

The Random Forest model outperformed Logistic Regression across all evaluation metrics, achieving higher overall accuracy and improved recall for fake news. The increase in fake news recall from 41% to 49% represents a 19.5% relative improvement.

This finding highlights the advantage of ensemble methods, which are better equipped to capture non-linear relationships and complex feature interactions within high-dimensional textual data.

Finding 4: Linguistic Differences are Subtle but Detectable

Statistical analysis revealed that, although many individual features showed statistically significant differences ($p < 0.05$), the corresponding effect sizes were small to moderate (Cohen's $d = 0.3–0.6$). The aggregate analysis yielded a borderline significant result ($p \approx 0.054$).

These results indicate that linguistic differences between fake and real news are present but not pronounced. Consequently, no single feature serves as a definitive indicator of deception.

Finding 5: Distinct Linguistic Patterns Characterize Each Class

Feature importance analysis identified clear qualitative differences in language use:

Fake news is characterized by emotionally charged, absolute, and sensational language, designed to capture attention and evoke strong reactions. Real news tends to employ measured, factual, and attribution-based language, often referencing sources and institutions.

These patterns are consistent with the Elaboration Likelihood Model (ELM), where fake news leverages peripheral route persuasion, relying on emotional and heuristic cues rather than substantive argumentation.

Finding 6: Conservative Classification Bias is Evident

Both models exhibited a pronounced conservative bias, favouring classification of ambiguous content as real rather than fake. This is evidenced by the confusion matrix, which showed 50 false negatives (fake news misclassified as real) compared to only 4 false positives (real news misclassified as fake), resulting in a ratio of approximately 12.5:1.

This bias likely reflects:

- The difficulty of detecting deception based on limited textual information
- The overlap in linguistic features between fake and real news

From a practical standpoint, such a bias may be acceptable—or even desirable—in contexts where false accusations (false positives) carry greater risk than missed detections (false negatives).

V. CONCLUSION

Multimodal machine learning represents a paradigm shift in fake news detection offering capabilities essential for addressing contemporary misinformation challenges. It demonstrates that multimodal approaches consistency outperform unimodal baselines with fusion strategies enabling comprehensive capture of deceptive patterns across text and image modalities. Early detection frameworks, including MATRAL and EANN show particular promise for identifying fake news within critical intervention windows before viral propagation.

However significant challenges persist. Limited annotated datasets constrain model training, cross modal semantic gaps sophisticated reasoning architectures and black box models limit practical adoption. Future research must address must address these limitations through knowledge aware pre training, large vision language models explainable AI integrations and cross lingual generalization. As adversarial tactics continue evolving sustained innovation in multimodal detection methodologies

remains essential for safeguarding information integrity in digital ecosystems.

VI. SCOPE FOR FUTURE RESEARCH

Knowledge aware pre training: Zhang (2025) propose KAMP(Knowledge Aware Multimodal Pre training), incorporating unsupervised correlations through pre training tasks to alleviate annotation dependency. This paradigm captures valuable signals from single modalities, multiple modalities and background knowledge graphs simultaneously.
Large Vision Language Models (LVLMs): Ai (2026) document the paradigm shift toward LVLMs enabling unified end to end multimodal reasoning framework. These models transforms traditional feature engineering approaches into sophisticated reasoning systems capable of high level semantic understanding.

Explainable AI integration: Jadhav (2025) introduce HEMT Fake (Hybrid Explainable Multimodal Transfer Fake), combining transformer embeddings, CNN BiLSTM text encoding, ResNet image features, and GraphSAGE metadata with hierarchical explainability. Their approach unites attention mechanisms SHAP and LIME to provide taken, sentence and modality level transparency.

Quantum Machine learning: Emerging research explores quantum multimodal fusion for enhanced scalability and efficiency. C et (2025) highlight variational quantum circuits and hybrid quantum classical models as promising approaches for next generation detection systems.

Cross lingual and cross cultural generalization: Jadhav(2025) introduce multilingual multimodal datasets encompassing Hindi, Gujarati, Marathi, Telugu, English demonstrating 7-8% performance gains in low resource languages. Extending these efforts to additional languages and cultural contexts remains crucial.

Collectively, these research directions highlight that the future of fake news detection lies in multimodal, adaptive, and human-cantered systems. Advancing along these dimensions will enable the development of more robust, scalable, and trustworthy solutions capable of addressing the increasingly complex landscape of digital misinformation.

REFERENCES

[1] Boididou, C., Papadopoulos, S., Zampoglou, M., Apostolidis, L., Papadopoulou, O., & Kompatsiaris, Y. (2018). Detection and verification of misinformation in social media. Multimedia Tools and Applications, 77(10), 12371–12404.
[2] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FNDNet: A deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32–44.
[3] Khattar, D., Garg, A., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In Proceedings of the World Wide Web Conference (pp. 2915–2921).
[4] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2020). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36.
[5] Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining (pp. 312–320).
[6] Singh, V. K., Ghosh, S., & Jose, J. M. (2021). Toward multimodal fake news detection: A systematic review. Journal of the Association for Information Science and Technology, 72(12), 1606–1627.
[7] Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 849–857).
[8] Wu, Y., Zhan, P., Zhang, Y., Wang, L., & Xu, Z. (2021). Multimodal fusion with co-attention for fake news detection. Neural Computing and Applications, 33, 18971–18984.
[9] Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57(2), 102025.
[10] Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2022). Early detection of fake news on social media: A review. Information Processing & Management, 59(1), 102768.