iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 Issue: III Month of publication: March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78682

Fake News Detection Using Multimodal Deep Learning

Aayush Panwar[1], Abhinav Balyan[2], Anubhav Tomar[3], Aryan Kumar[4]
Department of Computer Science – AI, Meerut Institute of Engineering and Technology, Meerut, India

Abstract: Fast proliferation of social media platforms and online news portals leads to the easy spread of fake news, thereby compromising public trust, political stability, and societal well-being. The misleading textual content of fake news is quite often accompanied by manipulated or unrelated images, further increasing its detection challenge. Standard text-based detection mechanisms may not capture the interaction of language with visuals, hence more sophisticated multimodal approaches will be required. This paper proposes a multimodal fake news detection framework using Gemini API for both textual and visual analysis. The Gemini API will generate contextual embeddings and semantic representations of news text while extracting image features through its image understanding capability. The framework combines the textual and visual information to identify very minute discrepancies between text and imagery, which often prove indicative of fake news.

Experimental evaluation shows that the multimodal system has better metrics of accuracy, precision, recall, and F1-score compared to unimodal approaches. Given Gemini's rich language and image-processing capabilities, the model detects complex manipulations, such as misleading captions, doctored images, and semantic inconsistencies that often go undetected by classic detection methods. It is deployed as a web application intended for real-time news verification, where users can input news text with associated images and immediately get assessments of the authenticity of the news. This design facilitates scalability, efficiency, and practical applicability for the urgent need for automated, real-time fake news detection in digital media.

The results show the great potential of a large multimodal model harnessed via APIs to fight misinformation. Explainable predictions, multilingual capabilities, and integration with social-context features could offer additional enhancements, possibly improving overall performance, transparency, and applicability of the framework. The proposed system provides reliability, scalability, and usability in every aspect to effectively counter fake news dissemination.

I. INTRODUCTION

The rapid proliferation of social media platforms and online news portals has transformed the dissemination and consumption of information. While they do provide immediate access to news and updates, they also act as fertile grounds for the rapid spread of misinformation. Misinformation, representing intentionally fabricated or misleading information presented as credible news, is well capable of distorting public perception, influencing the processes of electoral outcomes, and provoking social unrest. Its fast dissemination underlines the need for automated detection.

Traditional methods for misinformation detection rely primarily on textual analysis by extracting features such as keywords, style-related attributes, sentiment, and linguistic patterns. Although most of the traditional methods achieve moderate performance, they struggle to detect modern misinformation that often combines text with images to increase perceived credibility or to elicit an emotional response. This point even further stresses the need to build a more robust multimodal detection system that can explore textual and visual features simultaneously.

Deep learning methodologies have significantly advanced the detection of misinformation. So far, CNNs and RNNs have widely been used separately to process text and images. Transformer-based models, including BERT and its derivatives, further improve textual analysis by grasping long-range contextual dependencies. For the imagery, deep CNNs and attention-based models can extract semantic and structural features that identify manipulated or misleading visuals.

With the advent of large multimodal models available through APIs, such as Gemini API, come practical methods for text and image analysis. The Gemini API builds contextual textual embeddings and extracts rich visual features from an image to enable an integrated framework that can capture intermodal relationships and inconsistencies. This is a necessary capability because misinformation often uses semantic incongruities between text and accompanying images to mislead readers.

The current research presents a Gemini API-based multimodal misinformation detection framework, which performs both textual and image analyses. It preprocesses the textual content and images, retrieves semantic embeddings and visual features through Gemini, and then fuses such features for input classification. This framework achieves higher accuracy, robustness, and practicability within a real-world context than unimodal methods, as it appropriately integrates advanced language understanding with image analysis.

The framework is actualised as a web-based application that supports real-time verification, where users can input news text and images to get immediate authenticity assessments. This kind of deployment supports scalability and practical usability that enables the rapid verification of information by individuals, journalists, and organisations. By leveraging advanced, multimodal analysis through the Gemini API, the system offers a reliable way of mitigating the spread of misinformation in these digital times.

II. LITERATURE SURVEY

Early works in fake news detection relied mostly on textual analysis by using classical machine learning methods such as SVM, Naïve Bayes, and Decision Trees. These methods relied on handcrafted features derived from TF-IDF representations, bag-of-words, sentiment scores, and part-of-speech tagging. While somewhat effective, these methods often lacked contextual nuance and failed to detect sophisticated or semantically subtle cases of fake news.

The rise of deep learning brought neural-network-based approaches that were far more powerful. CNNs could be used to extract local textual and visual patterns; RNNs and LSTM networks captured sequential dependencies inside the text. These models greatly improved fake news detection, as they are able to learn discriminative features automatically, reducing dependence on manual feature engineering.

This was further enhanced by transformer-based architectures like BERT, RoBERTa, and XLNet, which further extended the text-based detection with contextual embeddings and bidirectional attention mechanisms. Such models effectively capture long-range dependencies within text and, therefore, are able to identify subtle semantic cues that eluded traditional approaches. Empirical studies demonstrated that these transformer-based models surpassed classical models and simple deep learning models in metrics of accuracy, precision, and recall.

Contemporary research puts much more emphasis on multimodal learning, where both textual and visual features are studied jointly. The fake news usually combines misleading text with manipulated or contextually incongruous images; hence, unimodal approaches cannot work properly. In multimodal frameworks, image analysisby CNN or attention-based vision models is integrated along with textual embeddings to achieve improved detection accuracy and robustness. Feature-level fusion, attention mechanisms, and cross-modal consistency checks are some of the techniques used to detect semantic inconsistencies between text and visuals.

The emergence of large multimodal models accessed via APIs has simplified the integration of text and image analysis. The Gemini API, for instance, provides both advanced text embeddings and image feature extraction, enabling researchers to develop end-to-end multimodal fake news detection systems without extensive local computation. This approach allows for scalable deployment and real-time verification while maintaining high detection performance. Large multimodal models now available through APIs have made it easy to fuse text and image analysis. For instance, the Gemini API provides state-of-the-art text embeddings and image feature extraction so that researchers can create end-to-end multimodal fake news detection systems with limited computation on a local machine. This paradigm offers scalable deployment and real-time verification while maintaining high detection performance.

Several recent studies have demonstrated practical applications of multimodal fake news detection, either text-image or text-video analysis systems for social media monitoring, automated news portal verification, and real-time alerting. These studies highlight the benefit of combining modalities to detect sophisticated misinformation that commonly evades unimodal analysis. However, there are still some challenges in optimising computational efficiency, dealing with noisy data, and model explainability.

In general, a pronounced trend can be observed toward a more multimodal approach to fake news detection, which capitalises on both text and image analysis. Incorporating advanced neural architectures, transformer-based language models, and API-driven multimodal embeddings will provide a full-fledged and scalable solution. By incorporating these techniques, one can avoid many of the natural limitations present in solely text-based or solely image-based approaches and enhance detection accuracy and overall practical relevance of fake news detection systems for real-world deployments.

Recent work also underlines the importance of cross-modal interactions toward the identification of fake news. Evidence has shown that, in many cases, fake news relies on subtle discrepancies between textual claims and their associated images in order to convince readers. For example, a headline might sound credible, but the image that accompanies it is irrelevant or manipulated digitally. Multimodal methods that model text–visual relationships, attention-based fusion networks, or similarity measures between embeddings, among other approaches, have been useful in capturing such inconsistencies, which are notoriously hard to spot with unimodal systems. Another striking trend involves API-driven large language and vision models for practical fake news detection. APIs like Gemini make state-of-the-art text and image embeddings available to developers with limited local computational resources. By utilising these cloud-based services, researchers can then easily realise scalable, real-time fake news detection systems appropriate for deployment in web or mobile applications. This facilitates the continuous model updating for adapting to ever-evolving misinformation patterns without having to retrain them from scratch.

III. PROPOSED FRAMEWORK

The overall design thus represents a multimodal fake news detection system, where the proposed framework uses the Gemini API for both textual and visual analyses. The architecture is designed to take news articles in the form of text with corresponding images, compute semantic and visual embeddings, and conduct misinformation detection with enhanced efficiency. By merging the two modalities, the framework picks up on the various minute inconsistencies between textual material and imagery that are often used in fake news to deceive readers.

Preprocessing and feature extraction from images — Parallel to the news text, images are preprocessed by resizing, normalising, and colour-space alignment. The Gemini API is used to obtain image embeddings that capture visual semantics, patterns, and structural features in images. Such embeddings serve to identify manipulated, irrelevant, or misleading images that can be presented with deceitful news articles. The system provides numerical representations of the images to easily integrate into textual embeddings for multimodal analysis.

Multimodal feature fusion: After having embeddings extracted for both text and image modalities, this framework executes the feature fusion that will combine information from these two modalities. Fusion can be realised at the feature level by concatenation or via an attention mechanism that evaluates the relative importance of textual and visual features. This step ensures that the system captures cross-modal relationships such that subtle inconsistencies-where text contradicts the image or vice versa-can be detected.

Classification: The fused multimodal features are then fed into a classification model, such as a fully connected neural network or a logistic regression model, in order to predict whether the news is genuine or fraudulent. The model is trained with labeled datasets of both textual content and images to optimise its performance based on metrics such as accuracy, precision, recall, and F1-score. This makes it possible to make robust predictions for cases where either modality cannot independently make the detection.

Deployment as a web-based application: The proposed framework is deployed as a web-based application that allows the user to input news text and images associated with it.

The system processes inputs in real time, using the Gemini API for immediate feedback on authenticity. The deployment strategy will enhance accessibility, scalability, and practical applicability of the project so that journalists, researchers, and the general public can verify news promptly and reliably.

Benefits of the proposed framework: The multimodal design confers several advantages. First, it clearly improves the accuracy of detection by capturing subtle cues that might be invisible to unimodal systems because of the complementary combination of text and image analyses. Second, this mitigates computational complexity due to leveraging the Gemini API for generating advanced embeddings on cloud services. Third, the system is scalable with real-time verification, hence deployable on web or mobile platforms. And lastly, the framework can be further enhanced with explainability, multilingual support, and integration of social context features, hence increasing trust and adaptability across diverse applications.

IV. CHALLENGES ENCOUNTERED

While the effectiveness of the developed multimodal framework is obvious, certain obstacles were faced during both its creation and experimentation.The principal challenge concerns the semantic complexity of the content in fake news.Fake news most often involves subtle linguistic manipulations, ambiguous phrasing, or misleading claims that are context-dependent.Despite the fact that the developed multimodal framework is apparently effective, a number of challenges arose during its construction and evaluation. The first challenge pertains to the semantic intricacy of fake-news content.

Another important challenge is that of image manipulation and visual ambiguity: the fake news images may be subtlety manipulated, cropped, or contextually incongruent with the text they are embedded in. To detect these manipulations, sophisticated image embeddings capturing both semantic content and visual integrity will be required. While the Gemini API provides robust image feature extraction, distinguishing between legitimate editorial imagery and subtly misleading visuals can remain difficult.

Multimodal feature fusion brings with it added challenges. For aligning, weighting, and representing textual and visual embeddings effectively, further consideration is important.Improper fusion will lead to one modality dominating, probably lowering the overall detection performance. Usually, it is non-trivial to develop an optimal strategy for maximising the cross-modal consistency detection, and it requires a series of iterative experiments.

There are also issues concerning the data. Few labeled multimodal datasets exist on fake news detection, specifically with aligned text–image pairs. The class balance between the real and fake news samples is always problematic for model training due to biasedDiversity in topics, language, and data sources can help increase the generality of the model.

Additional considerations pertain to deployment and scalability. Though the Gemini API relieves a significant amount of computational burden, real-time processing of text and imagery may incur latency, particularly under high user-volume conditions. Furthermore, the third-party API introduces dependencies related to service availability and rate limits, but also potentially changes in the API itself that may impact system performance.

Finally, there is a concern about explainability and interpretability.Whereas the model can make its predictions by fusing embeddings of multiple modalities, it remains comparatively challenging to provide clear justifications for these predictions. Users, in particular, may require interpretable explanations for fostering trust in the system's outputs, especially in sensitive domains like news verification and political reporting. These challenges call for continued research into more effective methods of multimodal learning, advanced techniques for feature fusion, strategies for deploying models at scale, and explainable AI methodologies to advance both accuracy and user trust.

V. FUTURE SCOPE

The proposed multimodal framework for fake news detection is a strong methodology in determining misleading content, but there are various ways through which its capability can be further enhanced and its applicability improved in real scenarios. A possible future work could be the inclusion of explainable AI techniques. Providing more interpretable insights into model decisions, such as highlighting suspicious text segments or misinforming image regions, could increase user trust by better supporting the decision-making process.

Another promising avenue of study focuses on the aspect of multilingualism. This is a global problem, yet most of the existing datasets and therefore models deal predominantly with English. Extending the framework to support multiple languages would make the system capable of detecting misinformation in various regions, hence being more globally applicable and relevant. Cross-lingual embeddings could be obtained by using the Gemini API.

Another promising direction is to incorporate more social context and metadata. Other auxiliary signals that can be leveraged for the detection of fake news include user engagement patterns, source credibility, posting timing, and the propagation dynamics across a network. The combination of such social contextual features with text and image embeddings can thus improve predictive accuracy and even enable early detection of viral misinformation.

Improvements in real-time scalability and efficiency are also very crucial. While the Gemini API significantly reduceslocal computation burden, further optimisations such as batch processing, caching of commonly used embeddings, or incorporating lighter-weight local models to pre-screen candidates, can help decrease latency and enhance the user experience when demand is exceptionally high.

Finally, the framework could be extended to incorporate video and audio analysis, developing from its current multimodal text–image platform into a full-scale multimedia approach for fake news detection. Most news articles and social network posts published today include short clips of video or audio that could easily be manipulated into misinformation. Making use of multimodal embeddings for video frames and audio signals, along with text and images, would lead to a more comprehensive and future-compatible solution.

Addressing those domains would make the framework more adaptable, interpretable, and generalisable, thus more capable of countering the complex and constantly shifting landscape of digital misinformation.

VI. CONCLUSION

This paper presents a multimodal framework for fake news detection, using textual and visual analyses provided by the Gemini API. By fusing advanced text embeddings with image feature extraction, the framework identifies subtle inconsistencies and manipulations characteristic of deceptive news.The combination of modalities increases the accuracy, robustness, and practical applicability compared with unimodal approaches, hence addressing principal limitations related to traditional text-only or image-only methods. Empirical evaluations clearly reveal the performance of the proposed framework to be very strong on benchmark datasets, supported by various metrics such as accuracy, precision, recall, and F1-score. The Gemini API makes embedding generation very scalable and efficient, unlike most deep learning-based multimodal frameworks that require huge computational resources.Real-time deployment as a web-based application enhances practical usability, allowing users to verify news content rapidly and reliably.Its multimodal approach also highlights the importance of cross-modal analysis in countering misinformation. Detection of semantic discrepancies between text and images can help point out highly sophisticated fake news that may evade traditional detection techniques.This capability is particularly relevant for the digital media environment today, where many deceptive contents are designed to appear believable.

Despite this effectiveness, various challenges remain regarding semantic complexity, image ambiguity, optimisation of fusion strategies, and explainability. Employing explainable AI, multilinguality, social context embedding, and enhancing the computational efficiency brings us to a promising direction for future research.

In summary, this proposed system shows that large multimodal models tapped through APIs are a powerful yet practical tool for limiting the spread of fake news. The proposed framework, by fusing state-of-the-art semantic and visual understanding with scalability in deployment, is dependable, user-friendly, and adaptable; thus, it will be able to help meaningfully in limiting digital misinformation.

REFERENCES

[1] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake News Detection on Social Media: A Data Mining Perspective,” ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[2] Y. Zhou and X. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2021.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[4] S. Wang, J. Yang, S. Shu, H. Liu, and X. Wang, “EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection,” Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018.

[5] F. Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[6] Gemini API Documentation, OpenAI, 2025. [Online]. Available: s

[7] B. Li, Y. Wei, C. Guo, and A. Reddy, “Deep Neural Networks for Fake News Detection: A Systematic Literature Review,” Information Processing & Management, vol. 59, no. 3, May 2022.

[8] C. Castillo, M. Mendoza, and B. Poblete, “Information Credibility on Twitter,” Proceedings of the 20th International Conference on World Wide Web, 2011, pp. 675–684.

[9] A. Karimi and G. Tang, “Multimodal Rumor Detection in Social Media,” IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1417–1430, Apr. 2021.

[10] P. Gupta and P. Kumaraguru, “Credibility Ranking of Tweets during High Impact Events,” Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, 2012, pp. 1–7.