Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Integrating NLP and Computer Vision for Multimodal Fake News Detection with Explainable AI

Riya Mary Raju [1], Ajay Kumar Singh[2]

[1] Masters' Student, Computer Science and Engineering, Jain Deemed-to-be University
[2] Professor, Computer Science and Engineering, Jain Deemed-to-be University

Abstract

The proliferation of fake news through online channels is a real menace that has posed considerable problems in determining the authenticity of online information. This is a multimodal misinformation detection project that is an integrated project using both text and visual data in order to find out misleading information in a more effective way. The system uses transformer-based models including BERT, RoBERTa and XLNet to perform text classification besides CNN and FastText models to receive semantic and contextual features of the news articles. To verify images, convolutional neural networks and mobile net are used in order to identify manipulated or tampered pictures. This compilation of models makes it possible to create a single framework that will increase the accuracy of detection in a variety of data sets. Besides, Explainable AI methods, in particular SHAP, are included to explain decisions of a model and visualize the importance of features to guarantee transparency and trust in predictions. The entire system is deployed as a web application in Flask, whereby it has user registration modules, log in modules, content submission modules, and modules to display the classification results. The findings of the experiment prove that the suggested scheme has a better reliability level than single-modality systems. On the whole, the current project allows creating a viable and interpretable misinformation detection system based on natural language processing, computer vision, and explainable AI.

Keywords: Misinformation Detection, Fake News, Transformer Models, CNN, Explainable AI, SHAP

1. Introduction

In the digital age we are living in currently, communication has spread more information than ever before via social media, news websites, and communication platforms. On the one hand, this accessibility has increased connectivity and information exchange among the world and contributed to the increased rapid spread of misinformation, or information that is not true, misleading, or deliberately manipulated. The speed of spreading fake news may result in distortion of opinion and change in social and political choices, and weaken the reputation of the real sources of information. Identifying this misinformation has thus become a very important issue in ensuring the integrity of information and that the content of the digital world is still reliable and worthy.

The more conventional methods of misinformation detection have been primarily concerned with the analysis of the text or images alone. Text methods depend on language hints, number of words or classic machine learning techniques such as Naive Bayes and Support Vector Machines (SVM), which often miss the fine-grained meaning and contextual information about the data. On the same note, methods based on the first images only, use simple feature detection or simple convolutional networks that are not stable enough to recognize minor visual manipulations. Nevertheless, misinformation in real life situation usually involves a mixture of deceptive text with distorted pictures, and single mode detection methods are ineffective and unreliable in such cases. This has caused an urgent demand of multimodal systems that can help in coming up with a holistic estimation of information through incorporating many sources of information.

The proposed case Multimodal Misinformation Detection proposes a new framework, where the news text and accompanying images are analyzed in one framework. Text analysis consisting of transformers-based models like BERT, RoBERTa, XLNet are used for the modeling of contextual dependencies and semantic meaning, whereas the classification of CNN and FastText including the latter as feature extraction in layers gets improved by the aforementioned technologies. Convolutional neural networks (CNNs) and MobileNet are applied to detect manipulated, tampered, or fabricated visual material to perform image verification. The combination of text and image analysis can guarantee an increased or more valid evaluation of misinformation.

In addition to improve the levels of trust, Explainable Artificial Intelligence (XAI) methods like SHAP are integrated. The tools are interpretable as they visually depict which features have the greatest effect on model predictions, so that the user can understand the way and the reason behind individual classification. The whole system is adopted as a web application with Flask and offers different modules in registration, login, and content submission and classification output. The user-friendly interface facilitates easy interaction and at the same time, it provides efficient model deployment.

This project is also highly accurate in misinformation detection and it prioritizes transparency and interpretability- important attributes in developing trust in AI-powered systems. The suggested framework can be used to counteract misinformation through a scalable, explainable, and efficient method by integrating natural language processing, computer vision, and explainable AI. Finally, it leads to the creation of a more stable digital information ecosystem and preconditions the development of further studies in the field of multimodal content verification and automated fact-checking.

2. Related work

Misinformation has also been detected in a more sophisticated way with the development of machine learning and deep learning algorithms. Most initial studies in this field used the classical text-based classification methods like Naive Bayes, Logistic Regression and Support Vector Machines (SVM) primary techniques. These methods were moderately accurate based on the use of lexical and syntactic characteristics but were incapable of resolving deeper semantic connections of textual information [1],

[2]. Later research proposed deep learning algorithms that could learn contextual representations and enhanced the accuracy and strength of the fake news detection systems [3].

More progress was made on the application of convolutional neural networks (CNN) and recurrent neural networks (RNN) to extract textual features. Text and image feature models like TI-CNN were suggested to combine the features to achieve further detection performance and learn hierarchical patterns with multimodal data [4]. Likewise, CNN-RNN-based hybrid deep learning models have also shown great potentiality to capture spatial and sequential dependencies in news sources thus enhancing the postulation of models and accuracy of classification [5].

The use of transformer-based architectures was a significant breakthrough in the detection of textual misinformation. The models like BERT, RoBERTa and XLNet used the bidirectional attention procedures and context embeddings to comprehend semantics and dependencies within language and outperformed the other traditional and shallow neural networks [6][7][8]. Moreover, hybrid networks with convolutional neural networks (CNN) and recurrent neural networks (RNN) have been investigated to enhance the design of features and generalization of different data sets [9].

There has been a similar development in image-based misinformation detection with visual forgery detection being an important aspect of detecting altered or modified media. The CNN-based models, such as VGGNet, ResNet, and MobileNet, have been found to be very precise in identifying the inconsistencies within the pixel distribution, texture, and space features [10], [11]. In addition, the performance of lightweight deep learning models like the so-called MobileNet has been demonstrated to achieve dependable performance at a lower level of computation, which can be deployed in real-time applications [12].

The recent studies have moved towards what is known as multimodal frameworks which are the textual and visual modalities developed together to analyze misinformation holistically. The experiments of BERT-based multimodal fusion have demonstrated the high accuracy of fake news detection with the combination of semantic and visual cues [13]. On the same note, thorough surveys have also highlighted that multimodal fusion techniques including early, late and hybrid fusion has the potential to provide cross-modal interactions, which result in better model reliability and robustness [14].

Besides it, the concept of **Explainable Artificial Intelligence (XAI) has become critical in guaranteeing transparency and interpretability in the detection models of misinformation. Methods based on the usage of SHAP and LIME have been discovered to be useful in visualizing and explaining model choices to boost user confidence and comprehension of predictions made by AI-powered methods [15]. In general, natural language processing, computer vision, and explainable AI are an influential and understandable base that can help to fight misinformation in digital ecosystems.

3. Dataset

Figure 1 Dataset Preview

4. Proposed Methodology

The proposed methodology is expected to create a framework, which is automated and explainable in terms of detecting misinformation based on text and visual content. The method combines two primary pipelines text-based analysis and image-based analysis, which are individually processed and then combined to create a single classification decision. The system adheres to a sequence of phases which include data acquisition, preprocessing, the extraction of features, classification, fusion, and the explainability.

4.1. Text-Based Analysis:

The text analysis pipeline is aimed at detecting the false information in the written text like news articles, headlines, or social media posts.

4.1.1. Data Preprocessing:

The text data is then cleansed and normalized to eliminate noise. This involves removal of URLs, punctuations, digits, special characters and unnecessary white spaces. All stop words are eliminated and the rest of the words are tokenized. The standardization of word forms is done through stemming or lemmatization. This makes the textual data always to be consistent and have sense to go further.

4.1.2. Feature Extraction:

Text processing is followed by conversion to numerical values that can be processed by computation. This is accomplished by embedding words in context to acquire the semantic meaning of words as well as their syntactic structure. These characteristics enable the model to learn the association between words, which enhance the identification of minor patterns of misinformation.

4.1.3. Classification:

The text characteristics that are processed are loaded into a deep learning-based classifier which separates authentic and deceptive content. The model acquires the rhythms like tone, context and linguistic messages that tend to be the signs of deceitful writing. The result is a probability score or binary label of the reliability or falsity of the input text.

4.2. Image-Based Analysis

The analysis of the image part is aimed at testing the credibility of the visual contents presented with the text.

4.2.1. Image Preprocessing:

All the pictures are resized, normalized and converted into a similar format to be analyzed. Histogram equalization, filtering, and pixel scaling are some of the techniques that are used to improve the quality of the image and eliminate noise.

4.2.2. Feature Extraction:

The processed image is also fed into a convolutional architecture that obtains visual features such as texture, color distribution, and spatial structure. This makes the system identify minute discrepancies or intrusion like splicing, merging or manipulation of image areas.

4.2.3. Image Classification:

Features extracted on an image are fed through a classification layer that forecasts an authentic or a manipulated image. The classifier is trained on discriminative patterns that assist in distinguishing between honest and fake images as regards structural and contextual integrity of the image.

4.2.4. Multimodal Fusion and Explainability:

Upon receiving the results of classification of both the text and image models, the results are combined to get the final multimodal decision. Fusion process integrates textual and visual probabilities either through weighted averaging or decision-level fusion that produces a complete analysis of the authenticity of the content. This multimodal choice minimises false predictions by exploring the semantic and visual cues at the same time. To improve the transparency, the system has an Explainable AI (XAI) layer. It plots the influential characteristics and points out the areas which made the most contribution to the final prediction. This allows users to see why a given content was deemed as misleading or real and they should be encouraged to interpret and trust the information.

5. Results and Discussion

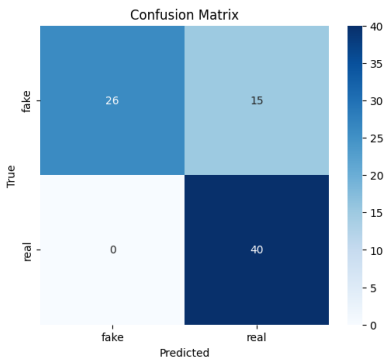

Figure 2 Confusion Matrix For Cnn

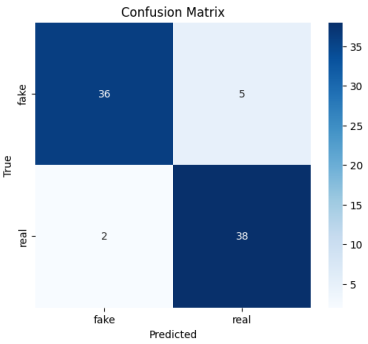

Figure 3 Confusion Matrix For Mobilenet

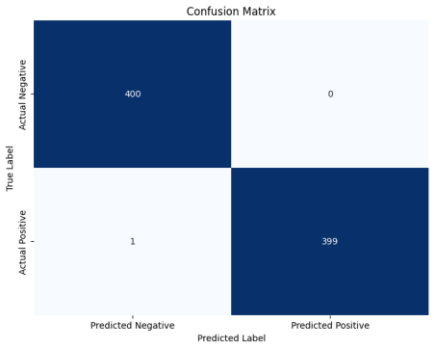

Figure 4 Confusion Matrix For Xlnet

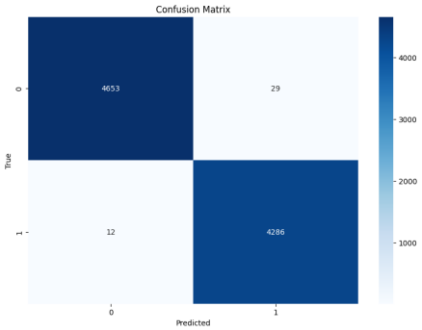

Figure 5 Confusion Matrix For Fast-Text+Cnn

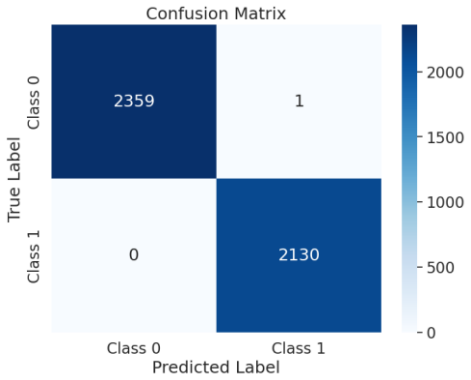

Figure 6 Confusion Matrix For Roberta

The effectiveness of the suggested multimodal fake news detector was measured by single model evaluation on the basis of the confusion matrix. All the models showed to differ in their levels of accuracy and precision in distinguishing between genuine and fake content.

To analyze the text, the XLNet showed excellent accuracy, having 400 and 399 positive and negative examples respectively, and the error rate was one. This shows that XLNet has a high contextual knowledge and a low error margin. Equally, RoBERTa scored almost perfect classification of 2,359 correct negative and 2,130 correct positive with a single false positive. The models based on transformers achieved better results in comparison to the traditional models because of their capability to state the bi-directional dependencies and subtle language characteristics.There was also the FastText + CNN hybrid model that performed and identified 4,653 negative and 4,286 positive correctly and misclassified a total of 41 items.

CNN model was used in image-based analysis and it was found to correctly identify 26 fake images and 40 real images whereas it wrongly identified 15 fake images as real. MobileNet, in its turn, demonstrated a higher generalization score, with the highest correct predictions of 36 fake images and 38 real ones, and reduced errors (five false negatives and two false positives).

In general, the findings indicate that textual models based on transformers, as well as deep CNN models of image verification, are highly reliable. Their combination in a multimodal fusion system increases the robustness of the system and makes it more reliable and consistent in detecting misinformation in a variety of data sources.

6. Conclusion

The given proposal, named Integrating NLP and Computer Vision to Multimodal Fake News Detection with Explainable AI, manages to prove the efficiency of multimodality use (text and images) in enhancing the accuracy of misinformation detection. Conventional single-modality systems do not have much ability to discover the whole context of deceitful material, whereas the multimodal system designed in this study addresses this issue by considering text and pictures together. Transformer-based models BERT, RoBERTa and XLNet turned out to be very effective in contextual semantic understanding, whereas CNN and MobileNet architectures recognized visual manipulations effectively.

Transformer models were tested and found to have close to perfect accuracy in classification with few instances of misclassification, which is indicative of their strength in detecting linguistic indicators of deception. Similarly, the visual models proved to be quite effective when it comes to detecting tampered images. All these models performed better when they were compounded using multimodal fusion to reduce the number of both false positive and false negative results.The strongest aspect of this project is the integration of Explainable Artificial Intelligence (XAI), in this case, SHAP visualizations, which allowed gaining insight into the way decisions were made by the model. This guarantees user trust and interpretability which is important to be applied in real world in misinformation monitoring systems. The fact that the entire framework itself was implemented as a Flask-based web application also confirms that it can be and is practically used in practice, having a convenient interface that users can use to analyze and check the digital content.

7. Future Scope

The suggested multimodal misinformation detection model has a number of future research and development opportunities. Though the present system successfully combines the use of text and visual modalities, the future research can develop the model to include other types of data like audio and video. The growth would allow full multimodal investigation, which is necessary because misinformation spreads faster and bigger using multimedia such as short videos and podcasts.

The other notable point of development is the application of larger, multilingual and cross-domain datasets. Increasing the diversity of the dataset would enhance the generalisation ability of the model to diverse languages, cultures and topics, which would improve its flexibility to global information condition. Secondly, it would be possible to introduce the measures of real-time misinformation detection that might allow dynamically tracking news feeds and social media streams and intervene timely in the case of misleading information.
The scope of improvement can also be established in the **Explainable AI (XAI) component. Future studies can be based on creating more interactive and understandable explanation interfaces, as this would enable non-technical and technical users to more intuitively understand model predictions. This transparency is essential to improve user trust and confidence in the decision system that is based on AI.

Lastly, this framework can be integrated with the fact-checking organizations, governmental agencies, and social media platforms, which can be used to deploy the framework at large scale and put it into practice. It is based on these developments that the system can become a scalable, intelligent, and transparent

misinformation detection solution that plays a great role in ensuring the integrity, accountability, and authenticity of the digital information ecosystem.

References

1. U. Sharma, S. Saran, and S. M. Patil, “Fake News Detection using Machine Learning Algorithms”, Accessed: Nov. 10, 2025. [Online]. Available: www.ijert.org
2. A. Yadav, D. V. Rao, A. Yadav, and D. V Rao, “Fake News Detection Using Naive Bayes Classifier: A Comparative Study,” Journal of Management and Service Science (JMSS), vol. 3, no. 1, pp. 1–14, Apr. 2023, doi: 10.54060/JMSS.2023.22.
3. O. Bashaddadh, N. Omar, M. Mohd, and M. N. A. Khalid, “Machine Learning and Deep Learning Approaches for Fake News Detection: A Systematic Review of Techniques, Challenges, and Advancements,” IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3572051.
4. Y. Yang et al., “TI-CNN: Convolutional Neural Networks for Fake News Detection,” Jun. 2018, Accessed: Nov. 10, 2025. [Online]. Available: https://arxiv.org/pdf/1806.00749
5. J. A. Nasir, O. S. Khan, and I. Varlamis, “Fake news detection: A hybrid CNN-RNN based deep learning approach,” International Journal of Information Management Data Insights, vol. 1, no. 1, p. 100007, Apr. 2021, doi: 10.1016/J.JJIMEI.2020.100007.
6. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Nov. 10, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805
7. Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” Jul. 2019, Accessed: Nov. 10, 2025. [Online]. Available: https://arxiv.org/pdf/1907.11692
8. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” Adv Neural Inf Process Syst, vol. 32, Jun. 2019, Accessed: Nov. 10, 2025. [Online]. Available: https://arxiv.org/pdf/1906.08237
9. S. Sharma, M. Saraswat, and A. K. Dubey, “Fake News Detection Using Deep Learning,” Communications in Computer and Information Science, vol. 1459 CCIS, pp. 249–259, 2021, doi: 10.1007/978-3-030-91305-2_19.
10. L. Verdoliva, “Media Forensics and DeepFakes: An Overview,” IEEE Journal on Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.
11. M. Patel, K. Rane, N. Jain, P. Mhatre, and S. Jaswal, “Image Forgery Detection using CNN,” 2023 3rd International Conference on Intelligent Technologies, CONIT 2023, 2023, doi: 10.1109/CONIT59222.2023.10205377.
12. S. Shikalgar, R. K. Yadav, and P. N. Mahalle, “International Journal on Recent and Innovation Trends in Computing and Communication Lightweight MobileNet Model for Image Tempering Detection”, doi: 10.17762/ijritcc.v11i5.6524.
13. M. Al-alshaqi, D. B. Rawat, and C. Liu, “A BERT-Based Multimodal Framework for Enhanced Fake News Detection Using Text and Image Data Fusion,” Computers 2025, Vol. 14, Page 237, vol. 14, no. 6, p. 237, Jun. 2025, doi: 10.3390/COMPUTERS14060237.

14. X. Li et al., “A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective,” IEEE Trans Emerg Top Comput Intell, vol. 9, no. 4, pp. 2658–2675, 2025, doi: 10.1109/TETCI.2025.3543389.
15. A. B. Athira, S. D. M. Kumar, and A. M. Chacko, “A systematic survey on explainable AI applied to fake news detection,” Eng Appl Artif Intell, vol. 122, p. 106087, Jun. 2023, doi: 10.1016/J.ENGAPPAI.2023.106087.