MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3973-3975

Dynamic and Explainable Multimodal Fake News Detection Using Transformer-Based Text–Image Fusion Framework

Divya Rupesh Pimple
M.Tech CSE (AI&ML) Scholar,
Department of Computer Science and Information Technology
Chhatrapati Shivaji Maharaj University, Navi Mumbai
Dr Vikas Kumar
Professor & Head, Department of Computer Science and Information Technology
Chhatrapati Shivaji Maharaj University, Navi Mumbai

Abstract
The widespread use of social media and digital news platforms has made it easier for information to reach large audiences within seconds. However, the same platforms have also contributed to the rapid spread of misleading or false news content. Detecting such misinformation has become an important challenge for researchers and technology developers. This study presents a multimodal fake news detection system that evaluates both textual and visual information associated with online news. Instead of training a new model from scratch, the proposed system makes use of pre-trained transformer models for inference. Textual content is analyzed using the RoBERTa language model to understand contextual meaning, while images are examined using the Data-efficient Image Transformer (DeiT). To make the system more transparent for users, an explainability layer is included which identifies important keywords through TF-IDF analysis and evaluates the emotional tone of the text. The framework supports different input formats such as text, images, and URLs, allowing flexible evaluation of news content obtained from online sources. Functional testing of the system shows consistent prediction behavior across these input types. The results indicate that combining transformer-based text and image analysis with simple explainability techniques can provide a practical and understandable approach for detecting potentially misleading news content.

Keywords : *Fake News Detection, Multimodal AI, Explainable Artificial Intelligence, Transformer Models, RoBERTa, DeiT.*

Introduction

In recent years, the way people consume news has changed significantly. With the increasing popularity of social media platforms and online news portals, information can now be shared and accessed almost instantly. While this rapid dissemination of information has many advantages, it has also created an environment where misleading or fabricated news can spread easily. Fake news often appears similar to legitimate news articles, making it difficult for readers to verify its authenticity. In many cases, such misinformation can influence public opinion, create confusion, or even lead to social and political consequences.Traditional fake news detection techniques have largely focused on analyzing textual features present in news articles. Machine learning models are commonly used to examine writing patterns, word usage, and semantic relationships within the text. Although these approaches can detect certain linguistic signals associated with misinformation, they often overlook visual elements such as images that accompany the news content. In practice, manipulated or unrelated images are frequently used to strengthen false narratives, which means that analyzing only the textual component may not provide a complete understanding of the content.Recent advancements in deep learning, particularly transformer-based architectures, have improved the ability of artificial intelligence systems to understand both language and visual information. Pre-trained models such as RoBERTa have demonstrated strong performance in natural language processing tasks, while vision transformers have achieved promising results in image analysis. These models can capture complex relationships within data and can often be applied to new tasks without extensive retraining.Considering these developments, this work proposes a multimodal approach for fake news detection that combines textual and visual analysis. The system relies on pre-trained transformer models operating in inference mode, which removes the need for large training datasets or complex model development. In addition, the system incorporates simple explainability techniques to provide insight into the reasoning behind its predictions. The objective of this approach is to create a practical and transparent system that can assist users in identifying potentially misleading news content available online.

Objective

1. To design a framework capable of analyzing news inputs such as text, images, and URLs.
2. To analyze textual information using the RoBERTa model for understanding contextual meaning in news content.
3. To evaluate associated images using the Data-efficient Image Transformer to examine visual cues related to the news.
4. To incorporate explainability by extracting important keywords using TF-IDF and performing tone analysis.
5. To combine textual and visual results through a multimodal fusion process to generate the final prediction.

Literature Review

The problem of fake news detection has attracted considerable attention in the fields of data science and artificial intelligence. Early research mainly relied on traditional machine learning algorithms that used handcrafted linguistic features such as word frequency, sentiment scores, and stylistic characteristics of news articles. Although these methods provided useful insights, their ability to capture deeper contextual meaning was limited.The introduction of transformer architectures brought significant improvements in natural language processing. The Transformer model proposed by Vaswani et al. introduced the self-attention mechanism, which allows models to understand relationships between words more effectively than earlier sequential models. Building upon this architecture, Devlin et al. introduced BERT, a bidirectional language model capable of capturing contextual dependencies in text. These developments have greatly enhanced the performance of text classification tasks, including misinformation detection.Researchers have also explored the role of social context in identifying fake news. Shu et al. discussed how user interactions, network structures, and content characteristics can be analyzed together to better understand the spread of misinformation on social media platforms. At the same time, the importance of interpretability in artificial intelligence systems has been emphasized in several studies. Doshi-Velez and Kim highlighted the need for machine learning models to provide understandable explanations so that users can trust the system's predictions.Deep learning models have also been applied to fake news detection. For example, hybrid architectures combining convolutional neural networks and long short-term memory networks have shown promising results in capturing both spatial and sequential textual patterns. However, many of these approaches rely primarily on textual data and do not fully consider visual elements that may accompany online news articles.With the emergence of transformer-based vision models, image analysis has also improved significantly. Vision Transformers and Data-efficient Image Transformers have demonstrated strong performance in various image classification tasks. Despite these advancements, relatively few studies have attempted to integrate both textual and visual analysis within a single framework while also providing interpretability. The approach proposed in this study addresses this gap by combining pre-trained transformer models for text and images within an explainable multimodal system.

Proposed Methodology and System Implementation

1. Textual Analysis

The system employs RoBERTa in inference mode to process user-provided text. The model generates contextual embeddings and outputs a probability score indicating whether the content is likely to be real or fake. No fine-tuning or retraining is performed.

2. Image Analysis

Image inputs are evaluated using DeiT, a transformer-based vision model. The model processes image patches through self-attention mechanisms to generate classification confidence scores. Image predictions complement textual analysis in multimodal scenarios.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3973-3975

3. URL -Based Content Extraction

When a URL is provided, the system automatically extracts the article's textual content and primary image using web scraping techniques. This enables dynamic evaluation of live online news sources without manual preprocessing.

4. Explainability Layer

Explainability is achieved using TF-IDF keyword extraction to identify influential textual terms. Additionally, tone analysis evaluates emotional cues such as exaggerated or sensational language. These components operate exclusively during inference to enhance user understanding.

5. Multimodal Fusion

Textual and visual predictions are combined using a weighted fusion strategy to generate the final classification outcome. This approach ensures balanced contribution from both modalities.

The system is implemented using Python and open-source libraries. The architecture consists of a user interface, inference engine, and explainability module. The interface allows users to input text, upload images, or provide URLs. Transformer models are invoked in evaluation mode, ensuring efficient real-time inference. The system evaluates linguistic patterns, emotional tone, and visual context associated with news content rather than performing direct factual verification against external databases. No dataset construction, labeling, or model training is involved. The system exclusively relies on publicly available pre-trained models. Although a large-scale benchmark evaluation was not conducted, the system consistently produced stable predictions across different input modalities, validating its practical applicability.

The architecture of the proposed system consists of multiple interconnected modules that collectively perform multimodal fake news detection. The workflow begins with user input, which may include textual content, an article URL, or an image link. If a URL is provided, the system extracts the relevant text and image from the webpage using the URL extraction module. The extracted textual information is passed to the transformer-based text analysis module, while the image is forwarded to the vision-based image analysis module. The explainability layer processes the textual data to extract keywords and determine the emotional tone of the content. Finally, the outputs from these modules are combined in the multimodal fusion component to generate the final prediction.

Figure 1. Proposed Multimodal Fake News Detection System Architecture.

Working of the System

The proposed system performs fake news detection by analyzing textual content, associated images, and extracted URLs using a multimodal analysis framework. The system processes user-provided input through a sequence of modules that evaluate linguistic patterns, visual context, and source-related information before generating a final prediction.

Initially, the user provides news content to the system in the form of text, a URL, or an image. The system extracts relevant information from the input and preprocesses the data for further analysis. If a URL is provided, the system retrieves the article content and associated media using a web extraction module. This extracted information is then forwarded to the respective analysis modules.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 36 Issue 1, 2026, Pages: 3973-3975

The textual content of the news is analyzed using a pretrained transformer-based language model. The model evaluates the semantic structure, contextual meaning, and emotional tone present in the text. Based on this analysis, the system generates an initial prediction regarding the authenticity of the news content along with a confidence score. Simultaneously, the associated image is analyzed using a vision–language model that measures the semantic similarity between the image and the textual content of the news. This step helps identify cases where unrelated or misleading images are used to manipulate the perception of the news article. The similarity score generated by the model indicates whether the visual content supports the textual information.

Finally, the outputs from the text analysis, image–text similarity evaluation, and explainability module are combined to generate the final prediction. The system displays the authenticity label (Real or Fake), the associated confidence score, and an explanation of the factors influencing the decision. A sample system output illustrating the final prediction and explanation is presented in Figure 2.

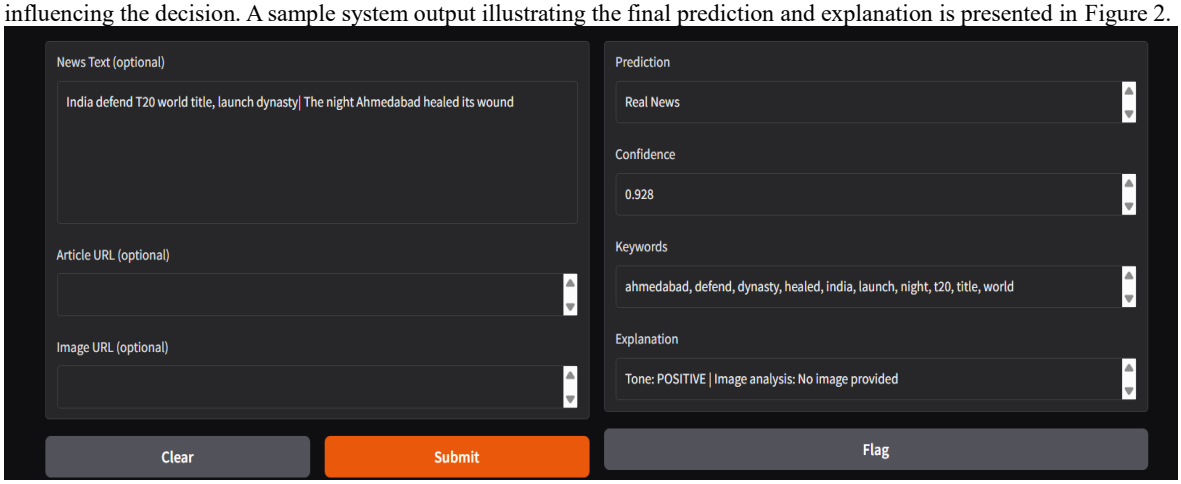

Figure 2. Final output of the system showing authenticity classification and confidence score.

Through this workflow, the proposed system enables efficient detection of potentially misleading information while maintaining transparency in the decision-making process.

Conclusion

This study presented a multimodal system designed to identify potentially misleading news by analyzing both textual and visual content. The proposed framework utilised pre-trained transformer models to evaluate news information without requiring additional training or dataset preparation. Textual data is analyzed using the RoBERTa model to understand contextual meaning, while visual elements are examined using the Data-efficient Image Transformer. By supporting different types of inputs such as text, images, and URLs, the system allows flexible evaluation of news content obtained from various online sources.

Another important aspect of the proposed system is the inclusion of an explainability component. By highlighting significant keywords and examining the emotional tone of the text, the system provides users with a clearer understanding of the factors that influence its predictions. This transparency helps improve user trust in the automated detection process. The implementation and testing of the system demonstrate that the framework can generate stable predictions across different input scenarios. Overall, the approach shows that combining transformer-based models with simple explainability techniques can provide a practical solution for detecting misleading information in digital media environments.

Future Scope

Future improvements to the system could focus on expanding the range of media types that the model analyses . For instance, incorporating video and audio analysis may allow the system to evaluate multimedia news content more effectively. Another potential direction is the integration of external fact-checking resources or knowledge bases to strengthen the verification process. Futhermore, exploring more advanced multimodal fusion strategies could further enhance the reliability of the predictions. Developing browser extensions or mobile applications based on the proposed system may also enable real-time detection of misleading news for everyday users.

References

[1] A. Vaswani et al., “Attention Is All You Need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[3] 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.
[4] F. Doshi-Velez and B. Kim, “Towards a Rigorous Science of Interpretable Machine Learning,” arXiv preprint arXiv:1702.08608, 2017.
[5] P. Kaur and M. Sharma, “Hybrid Deep Learning Model for Fake News Detection,” IEEE Access, vol. 9, pp. 134613–134624, 2021.
[6] A. Dosovitskiy et al., “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. International Conference on Learning Representations (ICLR), 2021.
[7] Y. Jin et al., “Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs,” in Proc. ACM Multimedia, 2017.
[8] S. Vosoughi, D. Roy, and S. Aral, “The Spread of True and False News Online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.
[9] X. Zhou and R. Zafarani, “Fake News: A Survey of Research, Detection Methods, and Opportunities,” ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020.