International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 1 (2026)
https://doi.org/10.31436/ijpcc.v12i1.644

Berita Debunked: Real-time Fake News Detection and Alert System

Ahmad Faisal Daniell bin Mohd Yusoff, Aiman Kamil bin Zainuddin, Raini binti Hassan*

Department of Computer Science, International Islamic University Malaysia (IIUM), Kuala Lumpur, Malaysia

*Corresponding author: hrai@iium.edu.my
(Received: 5th December 2025; Accepted: 22nd December, 2025; Published on-line: 30th January, 2026)

***Abstract*—** BeritaDebunked is an AI-driven near real-time fake news detection and alert system designed to combat misinformation in Malaysia, particularly on platforms such as WhatsApp. The system combines natural language processing and multimodal deep learning by using BERT for textual analysis and BLIP-2 for image–text evaluation. Deployed as a browser extension, it flags suspicious messages and allows continuous model updates through a scalable backend. Evaluation on the Fakeddit benchmark dataset demonstrates that the proposed hybrid architecture achieves an accuracy of (83.3%), with a precision of (82.6%) and an F1-score of (84.9)%. While unimodal text baselines achieved marginally lower raw accuracy (82.9%), the hybrid model demonstrates superior robustness in detecting multimodal context mismatches. The system demonstrates real-time capability with an average inference latency of 56.42 ms. By enabling timely detection and user-friendly alerts, BeritaDebunked aims to support digital literacy efforts, reduce the spread of misinformation, and contribute to Sustainable Development Goal 16 by strengthening information integrity.

Keywords— Hybrid, Fake news, SDG 16, BERT, BLIP-2, Multimodal deep learning, NLP

I. Introduction

Due to the widespread use of social media and messaging platforms, Malaysia is facing a challenge in controlling the rapid spread of fake news, particularly on WhatsApp, Facebook and Twitter. The process of manually verifying the authenticity of forwarded messages is difficult and unfeasible due to high volume and speed. Although initiatives by Malaysia Communication and Multimedia Commission (MCMC) Sebenarnya.my exist, the current approach remains slow, reactive and unable to prevent early public impact [1] [2].

The research also focuses on evaluating and comparing different machine learning models for fake news detection. The algorithms and results are presented and compared in a detailed yet concise manner using multiple evaluation metrics to identify the most reliable approach.

To address this problem, this project proposes the development of a real-time AI-driven fake news detection and alert system designed specifically for WhatsApp, leveraging natural language processing (NLP), machine learning, and deep learning techniques including models such as BERT for text classification and BLIP-2 for multimodal content analysis. The system prioritizes user privacy by analysing only message content and excludes personal metadata or private chat logs. It is built using Python, Flask/FastAPI, and a front-end browser extension, the platform ensures accessibility, scalability, and practical deployment for public use.

Despite its potential, developing such a system introduces several challenges, including privacy concerns due to WhatsApp's end-to-end encryption, compliance with Malaysia's Personal Data Protection Act (PDPA), and the need for fast, real-time performance supported by scalable infrastructure. Ethical issues such as algorithmic bias, transparency, and responsible alerting must also be addressed to prevent user distrust or over-reliance. The scarcity of labelled WhatsApp datasets further complicates model training, while risks of false positives, maintenance requirements, and evolving platform features pose additional hurdles.

Overall, the development of this AI-powered real-time detection system is critical for protecting information integrity, enhancing public digital literacy, and supporting organisations such as MCMC, MyCERT, fact-checkers, and media outlets in combating misinformation. This project also sets a technological precedent for misinformation detection on encrypted platforms, aligning with the goals of SDG 16 by promoting peace, justice, and strong institutions in the digital age.

This paper makes three main contributions: (1) a hybrid BERT+BLIP-2 multimodal model trained on Fakeddit for fake news detection; (2) an end-to-end architecture that integrates the model into a browser extension for real-time alerts; and (3) a comparative evaluation against unimodal baselines using standard metrics on a multimodal benchmark dataset.

https://doi.org/10.31436/ijpcc.v12i1.644

II. Literature Review

The proposed paper draws directly from the reviewed literature to establish an effective framework for real-time fake news detection and alerting across text and images. Insights from both unimodal and multimodal models have been adapted to address core limitations and gaps that are identified in prior studies, ensuring the system is methodologically robust, scalable and aligned with current research trends. *Evolution of Detection Models: From Unimodal to Multimodal*

Early research into fake news detection predominantly focused on unimodal approaches, utilizing machine learning and deep learning to analyze textual features. A few studies employing models like BERT, LSTM, and XGBoost have achieved exceptional accuracy rates. For instance, Cavus et al. [3] and Sharma et al. [4] demonstrated that semantic analysis can effectively identify false narratives, reaching accuracy levels up to 99.9%. However, these models face significant limitations in real-world applications. Unimodal systems are blind to visual context, rendering them ineffective against multimedia misinformation. Furthermore, approaches such as those by Rashad et al. [5] and Limbachia [6] rely on query-based inputs or domain-specific training (e.g., COVID-19 data), limiting their generalizability and scalability. To address these deficiencies, recent scholars have shifted toward multimodal architecture that processes both text and images.

To address these deficiencies, recent scholarship has shifted toward multimodal architectures that process both text and images. Advanced hybrid frameworks have emerged to tackle this complexity. Yan et al. [7] utilized BERT and BLIP-2 as feature extractors, integrating them through a 1D-CCNet attention mechanism and Heterogeneous Cross-Feature Fusion Method (HCFFM). This approach demonstrated the superior capability of BLIP-2 in capturing visual semantics compared to traditional CNNs. Similarly, Ojo et al. [8] employed a BiLSTM + VGG19 architecture, achieving 97.2% accuracy. While these systems demonstrate strong performance, current research is often hindered by high model complexity, small datasets, and class imbalances [9], [10]. Additionally, most existing multimodal models are restricted to image-text pairs and struggle with cross-domain generalization.

A. *Comparison of Existing Fake News Detection Tools*

Beyond academic models, several consumer-facing systems attempt to mitigate misinformation, though they rely largely on source-level credibility rather than real-time content analysis. NewsGuard [11] and Media Bias/Fact Check (MBFC) [12] operate primarily as browser extensions that rate the reliability of news domains. NewsGuard employs human analysts to grade sites based on journalistic criteria, while MBFC categorizes sources by political bias and factual reporting. While valuable for digital literacy, their source-level approach is a critical limitation. They cannot flag individual false articles hosted on generally credible sites, nor can they assess viral content on encrypted platforms. Furthermore, their reliance on human curation introduces subjectivity and scalability issues, with ratings often criticized for being US-centric or potentially biased.

In contrast, ClaimBuster [13] utilizes NLP to detect check-worthy factual claims in real-time. While it automates the detection process, surpassing the speed of human evaluators, it remains limited to textual content. It relies heavily on matching claims against existing fact-checking databases, meaning it often fails to detect novel misinformation or nuanced context-dependent falsehoods that involve imagery.

B. *Synthesis and Research Gap*

Literature reveals a distinct gap in current countermeasures. Unimodal models [3]- [6] lack of visual context while existing multimodal architectures, such as attention-heavy mechanisms proposed by Yan et al. [7], often prioritize architectural novelty over the latency requirements of real-time detection systems and struggles with generalization and deployment scalability. Furthermore, commercial tools either prioritize source reputation over content analysis [11]- [12] or ignore multimedia entirely [13].

Consequently, there is a critical need for a hybrid, real-time detection system capable of analyzing both text and images within encrypted environments. Building upon the robust feature extraction capabilities established by Yan et al. [7], the proposed system integrates BERT for deep semantic text analysis and BLIP-2 for visual reasoning. However, unlike prior complex fusion methods, this paper employs a direct concatenation and dense classification approach to balance high accuracy with privacy-preserving, real-time performance required for platforms like WhatsApp.

III. Methodology

This paper proposes a hybrid multimodal deep learning framework designed to detect misinformation by analysing both textual and visual components of a message. The system integrates state-of-the-art pre-trained models, BERT for text and BLIP-2 for images into a unified classification pipeline.

A. *Dataset*

The core dataset that will be used is Fakeddit, a publicly available dataset multimodal fake news dataset that includes both text and images, along with multi-class labels

https://doi.org/10.31436/ijpcc.v12i1.644

representing different levels of truthfulness [14]. The system utilizes this dataset to learn the patterns and features that distinguish real news from fake news across multiple modalities. Although the model is evaluated on the Fakeddit benchmark, which captures multimodal news content, WhatsApp messages in Malaysia may differ in style, language, and media usage. Therefore, the current results should be interpreted as an initial validation of the architecture, with future work focusing on collecting or adapting datasets that more closely reflect local WhatsApp communication patterns.

B. *Experimental Configuration and Reproducibility*

To ensure the reproducibility of our results, all experiments were conducted using a fixed random seed (seed=42) for both data splitting and model initialization. The dataset was partitioned into training (80%) and testing (20%) sets using stratified random sampling to preserve the class distribution of the original Fakeddit dataset. The hybrid model was implemented using PyTorch and the Hugging Face Transformers library. We utilized the AdamW optimizer and a batch size of 16 to fit within the memory constraints of a standard NVIDIA T4 GPU. No additional class balancing techniques (such as SMOTE or weighted loss) were applied.

C. *Proposed System Architecture*

System Analysis and Design Diagram are essential tools for meddling, understanding, and communicating the structure and behaviour of this system [15]. They will visually map the information to support specific goals and enhance cognitive processing during task performance. Effective diagrams such as system architectural diagrams are well known. The architectural diagrams provide a high-level overview of system components and their interactions, supporting communication, design and maintenance [16].

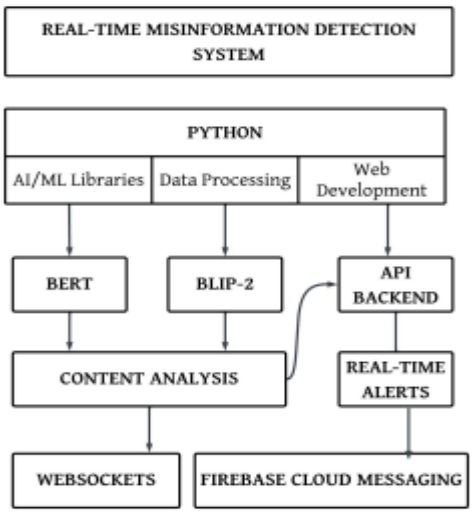

Fig. 1 System Architecture Design for BeritaDebunked

The architectural framework of the proposed system is illustrated in Figure 1. The core processing layer is built upon a Python-based backend that orchestrates the multimodal analysis. Incoming data from the browser extension is routed to the Content Analysis module, where the BERT and BLIP-2 models operate in parallel to extract linguistic and visual features, respectively. These features are fused to generate a credibility score, which is then transmitted via a high-performance FastAPI backend. To ensure real-time responsiveness, the system leverages WebSockets for low-latency communication, delivering immediate verification alerts to the user interface while asynchronously caching results in the Firebase cloud database for scalability.

IV. Results And Discussion

The developed system integrates BERT for text analysis and BLIP-2 for multimodal understanding within a Python-based framework. BERT enables nuanced detection of sentiment and bias in textual claims, while BLIP-2 analyses image-text alignment to identify inconsistencies characteristic of manipulated media. This dual-model approach addresses a critical gap in existing tools, which often rely on single-modality analysis.

While the full Fakeddit dataset contains over one million samples, this paper utilized a focused subset of 30,000 samples to balance training time with statistical confidence. Research by standard deep learning benchmarks indicates that validation sets exceeding 6,000 samples are sufficient to achieve model convergence and reliable performance estimates. Consequently, the results reported in this paper offer a high degree of confidence regarding the system's real-world applicability.

The dataset is split into training (80%) and testing (20%) subsets. This 80:20 ratio was selected as a standard convention in machine learning to maintain a balance between sufficient data for the model to learn complex multimodal feature representations and a large enough unseen validation set to rigorously test generalizability and prevent overfitting.

The qualitative analysis of the hybrid model reveals distinct behavioural advantages over unimodal approaches. While the BERT component successfully flags sensationalist text typical of 'clickbait' news, it struggles with posts where the text is neutral, but the accompanying image provides a misleading context. The integration of BLIP-2 addresses this semantic gap by generating image captions that are cross-referenced with the textual claims. This multimodal fusion allows the system to detect mismatch where the visual evidence contradicts the textual narrative a key indicator of sophisticated misinformation that text-only models often miss. This behaviour suggests that future improvements should focus on fine-tuning the cross-modal attention mechanisms rather than simply increasing dataset size.

https://doi.org/10.31436/ijpcc.v12i1.644

Performance benchmarking confirms the system's suitability for real-time deployment. On a standard T4 GPU environment, the model achieved an average inference latency of 56.42 ms per message with a throughput of 17.72 messages/second, well within the latency tolerance for instant messaging applications.

Table 1 reveals that the unimodal BERT model achieved an individual performance (82.9% accuracy), indicating strong textual cues in the dataset. However, the Hybrid model demonstrated competitive performance (83.3% accuracy), significantly outperforming the BLIP-2 baseline. While slightly higher than the unimodal text baseline, this trade-off is justified by the hybrid model's ability to detect multimodal context mismatches. This result highlights that while text remains the primary indicator of credibility in this dataset, the Hybrid architecture successfully integrates visual context with minimal loss in accuracy, providing a more holistic detection mechanism than text-only approaches.

TABLE I
RESULT COMPARISON OF MACHINE LEARNING MODELS

Modality	Machine Learning Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Unimodal	BERT	82.9	86.5	86.2	83.6
Multimodal	BLIP-2	60.7	60.6	70.9	65.4
	Hybrid (BERT+BLIP-2)	83.30	82.62	81.8	84.9

Figure 2 illustrates the Hybrid BERT+BLIP-2 model's training progression over 2 epochs. The validation accuracy stabilizes at approximately 83.3%, while the validation loss decreases to 0.39, indicating that the model successfully learned generalizable patterns without overfitting. The convergence of training and validation loss confirms the stability of the fine-tuning process on the 30,000-sample subset.

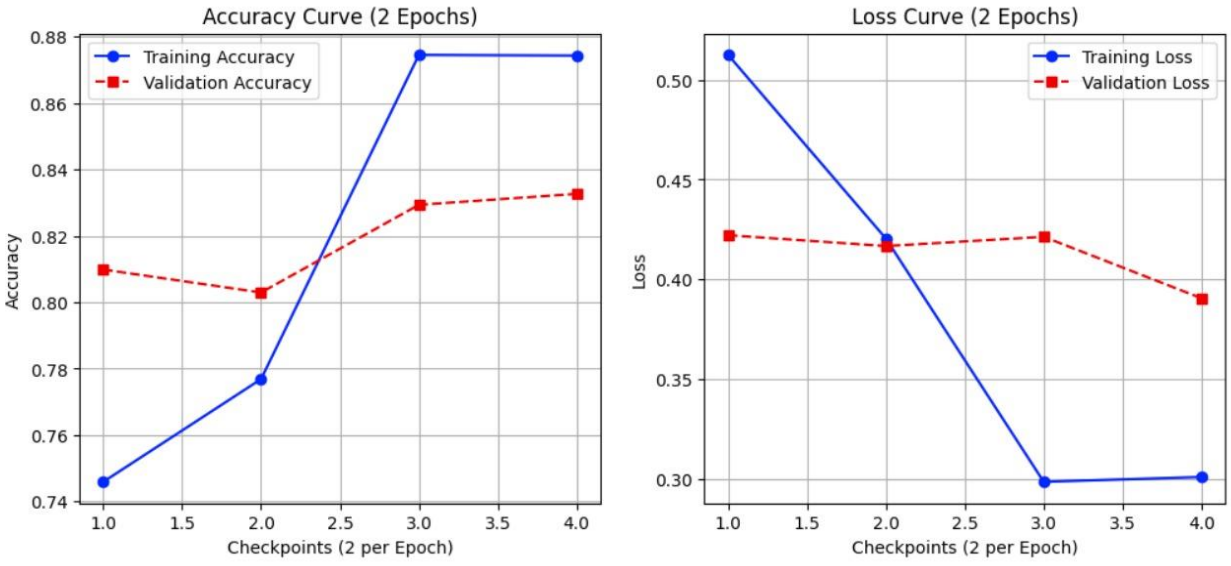

Fig. 2 Training and Validation Performance Curves for BERT+BLIP-2 model

It is important to note that these preliminary results are derived from a subset of the Fakeddit dataset due to computational resource constraints. While the current sample size is sufficient to validate the hybrid architecture's logic, scaling the training process to the full multi-terabyte dataset in future iterations is expected to further improve the model's precision and recall stability.

As detailed in

TABLE , the proposed hybrid model's accuracy (83.3%) presents a realistic performance baseline when balanced against the constraints of real-time deployment. While prior

https://doi.org/10.31436/ijpcc.v12i1.644

unimodal studies such as Cavus et al. [3] and Sharma et al. [4] reported accuracies exceeding 99%, these models were often trained on small, topic-specific datasets (e.g., COVID-19), which limits their ability to generalize to the broad-spectrum misinformation found on social media.

Similarly, in the multimodal domain, architectures like Yan et al. [7] achieved higher accuracy (92.5%) but relied on computationally intensive attention mechanisms (1D-CCNet) that are unsuitable for browser-based extensions. In contrast, the slightly lower accuracy observed in this paper reflects the trade-off required to achieve near real-time latency and privacy preservation. Unlike Ojo et al. [8] whose high-accuracy model operates offline, the proposed system successfully integrates verification into the user's workflow, prioritizing immediate impact and accessibility over raw metric maximization on a noisy benchmark like Fakeddit.

TABLE II.
COMPARATIVE ANALYSIS

Author, Year	Modality	Methodology	Dataset	Performance	Key Limitation
Cavus et. al. [3]	Unimodal (Text)	CRIPS-DM, BERT, MS Azure	COVID-19 News	Acc. up to 99.9%	Domain-specific training
Babar et. al. {18}		Hybrid N-Gram + LSTM	Social media	Acc. 96.5%	High computation cost
Sharma et. al. {4}		XGBoost, LSTM	News Dataset	Acc. up to 99.9%	Small dataset
Rashad et. al. [5]		TF-IDF Random Forest, Logistic Regression, LSTM	News Dataset	Acc. up to 99.8%	Query-based input
Limbachia [6]		Random Forest	News Dataset	Acc. 100%	Poor generalization
Yan et. al. (2024)	Multimodal	BERT + BLIP-2 (Model encoders text & images extractor) 1D-CCNet Attention Mechanism Heterogeneous Cross-Feature Fusion Method (HCFFM)	Multimodal News	Acc. 92.5–96.7%	Weak cross-domain support
Segura-Bedmar et. Al. [9]		CNN + BiLSTM	Multimodal News	Acc. 87%	Small dataset
Ojo et. al. [8]		BiLSTM + VGG19	Social media	Acc. 97.2%	No audio/video support
Saha [17]		DeBERTa + ConvNeXT	Multimodal News	Acc. 91.2%	Image-text only
Dellys et. al. [19]		ViLBERT + SVM	Multimodal News	Acc. 77%	Class imbalance
Proposed System		BERT + BLIP-2	Fakeddit	Acc. 83.3%	Computational constraint

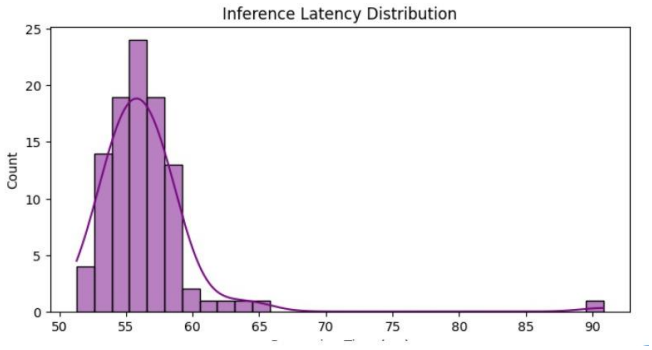

Fig. 1 Inference Latency Distribution and Throughput Analysis.

To validate the system's real-time capabilities, we conducted a latency stress test on a standard NVIDIA T4 GPU environment. As shown in Figure 3, the system achieved an average inference latency of 56.42 ms per message, with a 95th percentile lag of 60.11 ms. The system demonstrated a throughput of 17.72 messages per second.

These results contradict initial concerns regarding the computational overhead of the BLIP-2 component. With an average response time well below the 100ms threshold often cited for perceived instantaneity, Berita Debunked successfully meets the 'near real-time' requirement for interactive user verification workflows. Consequently, the system is characterized as offering prototype-level responsiveness suitable for user verification workflows, rather than high-frequency automated filtering.

For deployment, the backend architecture leverages asynchronous processing capabilities inherent in the FastAPI framework to handle concurrent requests efficiently.

Furthermore, the integration of Firebase Cloud Messaging ensures that alert delivery is decoupled from the heavy model inference process, preventing bottlenecks during high-traffic periods. Future work will focus on optimizing the BLIP-2 backbone to further reduce computational overhead.

V. Conclusions

To summarize, this paper successfully developed and validated a near real-time prototype, multimodal fake news detection system designed to combat misinformation on encrypted platforms. By integrating BERT for textual analysis and BLIP-2 for visual-semantic reasoning, the proposed hybrid model achieved a classification accuracy of 83.3% on a robust test set of 6,000 samples from the Fakeddit dataset. These results demonstrate that combining linguistic and visual features provides a more holistic verification mechanism than unimodal approaches, capable of identifying multimedia content in near real-time while maintaining user privacy through a browser-based extension architecture. However, this paper acknowledges certain limitations. The reliance on the Fakeddit benchmark as a proxy for WhatsApp messages introduces a domain shift, as the linguistic style of Reddit posts differs from private messaging patterns in Malaysia. Additionally, computational constraints necessitated the use of a 40,000-sample subset of the dataset, which, while statistically significant, does not capture the full variance of the complete dataset.

Lastly, the future work will focus on bridging this domain gap by fine-tuning the model on localized, anonymized Malaysian datasets to better capture regional dialects and specific forwarding behaviours. Furthermore, the project will prioritize the optimization and public deployment of the WhatsApp Web extension. This strategic focus acknowledges the current technical limitations of mobile operating systems in supporting real-time message interception, positioning the browser-based solution as the most viable path for immediate, scalable impact in combating misinformation.

Acknowledgment

The authors hereby acknowledge the review support offered by the IJPCC reviewers who took their time to study the manuscript and find it acceptable for publishing.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author(s) Contribution Statement

A.F.D. Mohd Yusoff contributed to the conceptualization, methodology, software development, and writing of the original draft. A.K. Zainuddin was responsible for data curation, validation, visualization, and reviewing and editing the manuscript. R. Hassan provided supervision and oversaw the project.

Data Availability Statement

The data that support the findings of this paper are openly available in the Fakeddit repository at [https://github.com/entitize/Fakeddit]. This dataset is a public benchmark for multimodal fake news detection. The source code and scraped samples used for the system demonstration are available from the corresponding author upon reasonable request.

Ethics Statement

This study did not require ethical approval.

References

[1]. A. Mat Isa, A. Z. H. Samsudin and M. R. Hendrawan, "Dissemination of Fake News and Information Disorder in Malaysia: A descriptive analysis," Environment-Behaviour Proceedings Journal, vol. 7, no. SI10, p. 53–58, 2022. doi: 10.21834/ebpj.v7isi10.4101.

[2]. "Kementerian Komunikasi," 19 May 2025. [Online]. Available: https://www.komunikasi.gov.my/awam/berita/23980-fake-news-spreaders-deserve-heavier-penalty. [Accessed Nov. 2025].

[3]. N. Cavus, M. Goksu and B. Oktekin, "Real-time fake news detection in online social networks: FANDC Cloud-based system," Scientific Reports, vol. 14, no. 1, 2024. doi: 10.1038/S41598-024-76102-9.

[4]. S. Sharma, M. Saraswat and A. K. Dubey, "Fake News Detection Using Deep Learning. Communications in Computer and Information Science," Communications in Computer and Information Science, vol. 1459, p. 249–259, 2021. doi: 10.1007/978-3-030-91305-2_19.

[5]. M. Rashad, N. Khalid, A. Hamza, S. Javed and K. B. Majeed, "A Semantic Fake News Detection System Using Machine Learning Classifier," Kashf Journal of Multidisciplinary Research, vol. 1, no. 12, p. 264–279, 2024. doi: 10.71146/KJMR171.

[6]. D. Limbachia, "Real-time Fake News Detection System Using AI," International Journal for Research in Applied Science & Engineering Technology, vol. 13, no. III, p. 560–565, 2025. doi: 10.22214/ijraset.2025.67294.

[7]. Y. Yan, H. Fu and F. Wu, "Multimodal Social Media Fake News Detection Based on 1D-CCNet Attention Mechanism," Electronics, vol. 13, no. 18, 2024.

[8]. A. O. Ojo, F. Najar, N. Zamzami, Z. T. Himdi and N. Bouguila, "SmoothDetector: A Smoothed Dirichlet Multimodal Approach for Combating Fake News on Social Media," IEEE Access, vol. 13, pp. 39289-39305, 2025. doi: 10.3390/electronics13183700.

[9]. I. Segura-Bedmar and S.-B. Alonso-Bartolome, "Multimodal Fake News Detection," Information, vol. 13, no. 6, p. 284, 2022. doi: 10.3390/INFO13060284.

[10]. H. Dellys, Mokeddem, Halimal and L. Sliman, "On the Integration of Social Context for Enhanced Fake News Detection Using Multimodal Fusion Attention Mechanism," AI, vol. 6, no. 4, 2025. doi: 10.3390/ai6040078.

[11]. "NewsGuard: Global Leader in Information Reliability.," [Online]. Available: https://www.newsguardtech.com/.

[12]. "Media Bias / Fact Check (MBFC)," [Online]. Available: https://chromewebstore.google.com/detail/media-bias-fact-

check/ganicjnkcddicfioohdaegodjodcbkkh?utm_source=item-share-cb.

[13]. "ClaimBuster: Automated Live Fact-checking," [Online]. Available: https://idir.uta.edu/claimbuster/.

[14]. K. Nakamura, S. Levy and W. Yang Wang, "r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, 2020.

[15]. S. Kumari, "Visual Modeling: Unlocking Ideas and Enhancing Understanding: The Power of Visual Modeling," International Journal of Engineering & Technology, vol. 12, no. 2, pp. 20-25, 2023. doi: 10.14419/ijet.v12i2.32334.

[16]. M. Malinova and J. Mendling, "Cognitive Diagram Understanding and Task Performance in Systems Analysis and Design," MIS Quarterly, vol. 45, no. 4, pp. 2101-2158, 2021. doi: 10.25300/misq/2021/15262.

[17]. K. Saha, "DeBERTNeXT: A Multimodal Fake News Detection Framework," Lecture Notes in Computer Science, vol. 14074, p. 348–356, 2023. doi: 10.1007/978-3-031-36021-3_36.

[18]. M. Babar, A. Ahmad, M. U. Tariq and S. Kaleem, "Real-Time Fake News Detection Using Big Data Analytics and Deep Neural Network.," IEEE Transactions on Computational Social Systems, vol. 11, no. 4, p. 5189–5198, 2024. doi: 10.1109/TCSS.2023.3309704.

[19]. S. K. Hamed, M. J. Ab Aziz and M. R. Yaakub, "Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments.," Sensors, vol. 23, no. 4, 2023. doi: 10.3390/s23041748.