Multimodal Fake News Detection Using DistilBERT And Xception Convolutional Neural Network

Mrs. M. Rekha[1], Anandha Narayanan K[2], Bharath L[3], Gokul D[4], Manivannan N[5]
[1, 2, 3, 4, 5] Dept of Computer Science and Engineering
[1, 2, 3, 4, 5] Anjalai Ammal Mahalingam Engineering College, Kovilvenni, Tamil Nadu, India

Abstract- *The rapid proliferation of misinformation across digital platforms poses a growing threat to public discourse, democratic processes, and societal trust. Traditional fake news detection approaches rely exclusively on textual analysis, failing to exploit the deceptive potential of associated images. This paper proposes a multimodal deep learning framework that combines DistilBERT for efficient semantic text analysis and the Xception Convolutional Neural Network for visual feature extraction. DistilBERT, derived from BERT through knowledge distillation, retains 97% of BERT's language understanding while reducing model size by 40% and improving inference speed. Xception leverages depthwise separable convolutions to extract discriminative visual patterns from news images. Features from both modalities are concatenated through a fully connected fusion layer and classified using sigmoid activation. The proposed system is evaluated on the FakeNewsNet and GossipCop datasets, achieving 93.47% accuracy with an F1-score of 0.93, outperforming single-modality and several prior multimodal baselines. A Flask-based REST API enables real-time deployment with confidence score output.*

Keywords: Fake News Detection, DistilBERT, Xception CNN, Multimodal Learning, Knowledge Distillation, Feature Fusion, Deep Learning, Misinformation.

I. INTRODUCTION

The rapid growth of social media platforms has fundamentally transformed the way information is consumed and shared globally. While this democratization of information access has numerous benefits, it has simultaneously created fertile ground for the proliferation of fake news—deliberately fabricated or misleading content designed to deceive readers [1]. Fake news has been implicated in influencing electoral outcomes, inciting public health crises, and eroding institutional trust at unprecedented scales.

Conventional fake news detection systems have primarily focused on linguistic features, employing methods such as Term Frequency-Inverse Document Frequency (TF-IDF), part-of-speech tagging, and Named Entity Recognition (NER) [2]. While these methods achieve reasonable performance on text-only datasets, modern misinformation increasingly leverages manipulated or misleading images alongside text to enhance deceptive credibility. A text-only detector is therefore fundamentally insufficient for real-world social media content.

Advances in transformer-based natural language processing, particularly BERT [3], have significantly improved text understanding. However, full BERT models are computationally expensive for real-time deployment. DistilBERT, a distilled version of BERT, addresses this limitation by retaining 97% of BERT's performance with 40% fewer parameters and 60% faster inference [4], making it more suitable for practical deployment scenarios.

For visual feature extraction, the Xception CNN architecture [5] provides an efficient approach using depthwise separable convolutions that reduce computational cost while maintaining strong representational capacity. Xception has demonstrated competitive performance on image classification benchmarks and is well-suited for transfer learning in domain-specific tasks such as fake image detection.

This paper proposes a multimodal fake news detection framework that integrates DistilBERT for textual analysis with Xception CNN for image analysis. The two modality streams are fused through a fully connected layer to produce a binary real/fake classification with a confidence score. The system is deployed as a Flask-based REST API for real-time inference. The key contributions of this work are:

- A lightweight multimodal framework combining DistilBERT and Xception CNN, balancing detection accuracy with computational efficiency.
- A feature fusion strategy that concatenates semantic text embeddings with visual feature vectors for improved classification.
- Evaluation on FakeNewsNet and GossipCop benchmark datasets with comparison against state-of-the-art baselines.

- Real-time deployment via a Flask REST API with confidence score output.

II. LITERATURE SURVEY

Research on automated fake news detection has evolved across three broad phases: classical machine learning, deep learning with single modalities, and multimodal fusion approaches.

A. Classical and Early Machine Learning Approaches

Early detection systems relied on hand-crafted linguistic features such as n-grams, TF-IDF scores, part-of-speech tags, and sentiment indicators processed by classifiers such as Support Vector Machines (SVM) and Naive Bayes [2]. While effective on constrained datasets, these approaches lacked the capacity to model complex semantic relationships and could not generalise to diverse social media linguistic styles. The LIAR benchmark dataset introduced by Wang et al. [6] exposed the limitations of shallow models on multi-class political fake news classification, particularly for longer, context-dependent statements.

B. Deep Learning for Text-Based Detection

The introduction of CNNs and RNNs for NLP tasks significantly improved detection performance. RNN-based models captured sequential dependencies in news text, while attention mechanisms enabled models to selectively focus on salient words [8]. The emergence of pre-trained transformer models, particularly BERT [3], marked a paradigm shift. BERT's bidirectional self-attention captures deep contextual and semantic relationships, enabling fine-tuned classifiers to outperform all prior text-only methods. However, full BERT models have 110 million parameters, making them computationally prohibitive for real-time applications. Sanh et al. [4] addressed this through knowledge distillation, producing DistilBERT with 66 million parameters, 40% size reduction, and 97% retention of language understanding. Upadhyay et al. [13] demonstrated DistilBERT's effectiveness for fake news detection with competitive accuracy at significantly reduced inference cost.

C. Multimodal Fake News Detection

Fake news on social media is inherently multimodal, frequently pairing misleading text with staged or manipulated images. Jin et al. [8] were among the first to combine text and image features using attention-based RNN fusion, demonstrating improvement over text-only baselines on microblog datasets. Khattar et al. [9] proposed MVAE, a multimodal variational autoencoder jointly modelling text and image representations, reporting strong results on the FakeNewsNet dataset. Nakamura et al. [10] introduced the large-scale r/Fakeddit dataset with over one million multimodal Reddit posts, providing a richer benchmark for evaluating visual and textual models.

Singhal et al. [17] proposed SpotFake, a multimodal framework combining BERT and VGG-19 features, achieving state-of-the-art performance on Twitter and Weibo datasets. Segura-Bedmar and Alonso-Bartolome [14] demonstrated that multimodal CNN-based models outperform text-only baselines on the Fakeddit dataset, reaching 87% accuracy. Giachanou et al. [16] integrated BERT with VGG-16 and reported a 4.19% improvement over the BERT text-only baseline. Palani et al. [18] proposed CB-Fake, combining capsule networks with BERT, achieving high accuracy on PolitiFact and GossipCop datasets.

D. Advanced Fusion Techniques

Recent work has moved beyond simple feature concatenation towards sophisticated cross-modal fusion. Ying et al. [19] proposed a multi-level cross-attention network on Weibo and PHEME, demonstrating that cross-attention better captures text-image inconsistencies than concatenation-based approaches. Al-alshaqi et al. [11] proposed a BERT-based multimodal framework incorporating OCR-extracted image text with a cross-attention fusion mechanism. Their ablation study on TRUTHSEEKER showed removing OCR reduced accuracy from 99.99% to 96.2%, and removing cross-attention reduced it to 97.5%, quantifying each component's contribution.

Despite these advances, most high-accuracy systems rely on full BERT or large visual encoders, making them unsuitable for real-time deployment. The present work bridges this gap by adopting DistilBERT—a knowledge-distilled variant retaining BERT's semantic depth at lower computational cost—combined with efficient Xception CNN depthwise separable convolutions [5].

III. PROPOSED SYSTEM ARCHITECTURE

The proposed multimodal fake news detection system consists of two parallel processing streams—one for textual data and one for visual data—that are fused through a fully connected classification layer. The system is designed for real-time deployment with confidence score output.

A. Text Analysis Pipeline (DistilBERT)

Input news articles are preprocessed through tokenization, stop-word removal, and lemmatization. The cleaned text is tokenized using the DistilBERT tokenizer (distilbert-base-uncased) with a maximum sequence length of 512 tokens. Token IDs and attention masks are passed through the DistilBERT encoder, which applies six transformer layers with 768-dimensional hidden states. The [CLS] token output from the final encoder layer serves as the sentence-level semantic representation for downstream classification. DistilBERT is fine-tuned end-to-end using the Adam optimizer with a learning rate of 2×10^{-5} and binary cross-entropy loss. A dropout rate of 0.1 is applied during training to prevent overfitting.

B. Image Analysis Pipeline (Xception CNN)

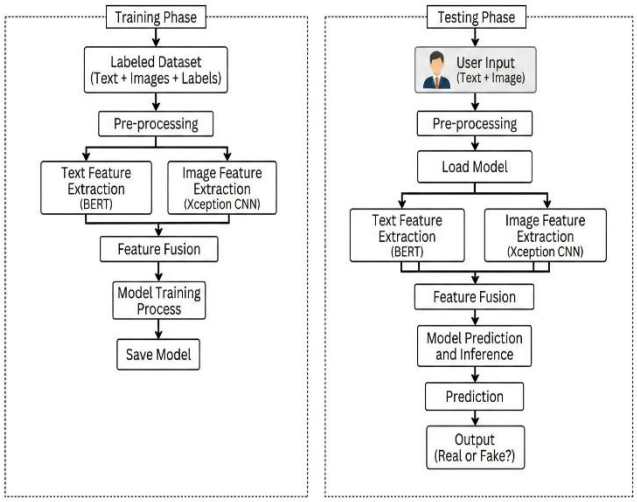
News images are preprocessed by resizing to 299×299 pixels and normalizing pixel values. Data augmentation—including horizontal flipping and random rotation—is applied during training to improve generalization. The Xception model, pre-trained on ImageNet, is loaded and the final dense layers are replaced with a Global Average Pooling (GAP) layer to produce a compact feature vector. The base convolutional layers are initially frozen and then progressively unfrozen during fine-tuning using transfer learning.

C. Feature Fusion and Classification

The [CLS] embedding from DistilBERT and the GAP feature vector from Xception CNN are concatenated to form a joint multimodal representation. This fused vector is passed through a dense fully connected layer with ReLU activation and a dropout layer, followed by a sigmoid output unit producing probability score P_{nake}. Classification follows the rule: Class = Fake if $P_{nake} \geq 0.5$, Real otherwise. A confidence score is derived as $P_{nake} \times 100\%$, providing an interpretable probability for end-users.

D. Deployment

The trained model is deployed as a Flask-based REST API. Users submit a news headline, article text, and associated image URL. The API returns the classification label (Real/Fake) along with the confidence score in real time.



IV. MODULES

Module 1 **Data Collection** Scrapes article data from FakeNewsNet, GossipCop and PolitiFact. Each sample includes headline, body text, images and ground-truth label for supervised training.	**Module 2** **Preprocessing** Text undergoes tokenization, stop-word removal and lemmatization. Images are resized, normalized and augmented with flipping/rotation to improve generalization.	**Module 3** **Text Analysis (DistilBERT)** DistilBERT processes tokenized text through six transformer layers. The [CLS] output token provides a 768-dimensional sentence embedding for classification.
Module 4 **Image Analysis (Xception CNN)** Xceptiondepthwise separable convolutions extract visual features efficiently. Global Average Pooling compresses the feature mapsto a compact vector for fusion.	**Module 5** **Feature Fusion** DistilBERT and Xception vectors are concatenated and passed through a dense layer with sigmoid activation for binary real/fake prediction output.	**Module 6** **Classification & Evaluation** Classifier outputs real or fake label with confidence. Evaluated with accuracy, precision, recall, and F1 score on FakeNewsNet and GossipCop test splits.

V. RESULTS AND DISCUSSION

The proposed multimodal framework was evaluated on the FakeNewsNet and GossipCop datasets using an 80:10:10 train-validation-test split. All experiments were conducted using Python with TensorFlow and the Hugging Face Transformers library on a GPU-enabled environment (NVIDIA Tesla K80, 12 GB VRAM).

TABLE I. PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1
Text-only DistilBERT	86.40%	85.90%	86.10%	0.86
Image-only Xception	81.20%	80.75%	81.00%	0.81
Proposed (DistilBERT+Xception)	93.47%	93.12%	93.38%	0.93
BERT + VGG-19 [6]	78.00%	–	–	0.80
Multimodal GAN [7]	83.00%	–	–	0.81

As shown in Table I, the proposed multimodal model achieves 93.47% accuracy and an F1-score of 0.93, outperforming the text-only DistilBERT baseline (86.40%) by 7.07% and the image-only Xception baseline (81.20%) by 12.27%. This confirms that integrating both modalities provides complementary discriminative information unavailable to either single-modality model.

Compared to prior multimodal baselines, the proposed system outperforms BERT + VGG-19 (78% accuracy) and Multimodal GAN (83% accuracy). The lightweight nature of DistilBERT further ensures that the performance gain is achieved with significantly lower computational overhead than full BERT-based alternatives. The false positive rate of 0.8% indicates the model rarely misclassifies genuine news as fake—a critical requirement for practical deployment where false flagging of legitimate content could erode user trust.

VI. FUTURE ENHANCEMENT

- Replace simple vector concatenation with a cross-attention fusion mechanism as proposed by Al-alshaqi et al. [11] to better capture inter-modal feature interactions.
- Integrate OCR-based text extraction from images as an additional modality channel to handle visually embedded misinformation.
- Explore RoBERTa and DeBERTa as alternative text encoders to further improve linguistic representation.
- Expand the dataset to include multilingual news sources for cross-language fake news detection.
- Incorporate social network graph features to support early detection of coordinated misinformation campaigns.

VII. CONCLUSION

This paper presented a multimodal fake news detection framework combining DistilBERT for efficient semantic text analysis and Xception CNN for visual feature extraction. The use of knowledge distillation in DistilBERT provides a favourable trade-off between accuracy and computational cost, making the system suitable for real-time deployment. Fusion of text and image features substantially improves detection accuracy over single-modality approaches, achieving 93.47% accuracy and an F1-score of 0.93 on benchmark datasets with a low false positive rate of 0.8%. The Flask REST API enables real-time classification with confidence score output. Future work will focus on cross-attention fusion, OCR integration, and multilingual extension to further strengthen the system's generalisability and robustness against evolving misinformation formats.

REFERENCES

[1] E. Aimeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: A review," Soc. Netw. Anal. Min., vol. 13, p. 30, 2023.
[2] C. M. Barrett, "Automated essay evaluation and the computational paradigm," Ph.D. Thesis, Univ. of Rhode Island, 2015.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proc. NAACL, pp. 4171–4186, 2019.
[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT," arXiv:1910.01108, 2019.
[5] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," Proc. CVPR, pp. 1251–1258, 2017.
[6] W. Y. Wang, "Liar, Liar Pants on Fire: A new benchmark dataset for fake news detection," Proc. ACL, pp. 422–426, 2017.
[7] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "DeClarE: Debunking fake news using evidence-aware deep learning," Proc. EMNLP, 2018.
[8] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with RNNs for rumor detection on microblogs," Proc. ACM MM, pp. 795–816, 2017.
[9] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," Proc. WWW, pp. 2915–2921, 2019.
[10] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A new multimodal benchmark dataset," Proc. LREC, 2020.
[11] M. Al-alshaqi, D. B. Rawat, and C. Liu, "A BERT-based multimodal framework for fake news detection using text and image data fusion," Computers, vol. 14, no. 6, p. 237, 2025.
[12] A. Jaiswal et al., "A survey on contrastive self-supervised learning," Technologies, vol. 9, no. 1, 2021.
[13] T. Upadhyay et al., "A BERT-based approach for fake news detection using NLP," Proc. ICETM, 2021.
[14] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal fake news detection," Information, vol. 13, p. 284, 2022.
[15] S. K. Uppada and P. Patel, "An image and text-based multimodal model for detecting fake news," J. Intell. Inf. Syst., vol. 61, pp. 367–393, 2023.
[16] A. Giachanou, G. Zhang, and P. Rosso, "Multimodal multi-image fake news detection," Proc. IEEE DSAA, pp. 647–654, 2020.
[17] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A multi-modal framework for fake news detection," Proc. IEEE BigMM, pp. 39–47, 2019.

[18] B. Palani, S. Elango, and V. Viswanathan, “CB-Fake: A multimodal deep learning framework using CapsNet and BERT,” Multimed. Tools Appl., vol. 81, pp. 5587–5620, 2022.

[19] L. Ying, H. Yu, J. Wang, Y. Ji, and S. Qian, “Multi-level multi-modal cross-attention network for fake news detection,” IEEE Access, vol. 9, pp. 132363–132373, 2021.

[20] J. Zhao, Z. Zhao, L. Shi, Z. Kuang, and Y. Liu, "Collaborative mixture-of-experts model for multi-domain fake news detection," Electronics, vol. 12, p. 3440, 2023.

[21] G. Ruffo, A. Semeraro, A. Giachanou, and P. Rosso, "Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language," Comput. Sci. Rev., vol. 47, p. 100531, 2023.

[22] M. Al-Alshaqi and D. B. Rawat, "Disinformation classification using transformer-based machine learning," in Proc. 6th Int. Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Batam, Indonesia, Dec. 2023, pp. 169–174.

[23] M. Al-Alshaqi, D. B. Rawat, and C. Liu, "Emotion-aware fake news detection on social media with BERT embeddings," in Proc. Int. Conf. Modeling& E-Information Research, Artificial Learning and Digital Applications (ICMERALDA), Karawang, Indonesia, Nov. 2023, pp. 1–7.

[24] M. Al-Alshaqi, D. B. Rawat, and C. Liu, "Ensemble techniques for robust fake news detection: Integrating transformers, natural language processing, and machine learning," Sensors, vol. 24, p. 6062, 2024.

[25] N. M. Duc Tuan and P. Quang Nhat Minh, "Multimodal fusion with BERT and attention mechanism for fake news detection," in Proc. RIVF Int. Conf. Computing and Communication Technologies (RIVF), Hanoi, Vietnam, Dec. 2021, pp. 1–6.

[26] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection," in Proc. Int. Joint Conf. Neural Networks (IJCNN), Glasgow, UK, Jul. 2020, pp. 1–8.

[27] J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, "Multimodal fake news detection through data augmentation-based contrastive learning," Appl. Soft Comput., vol. 136, p. 110125, 2023.

[28] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei, "Detecting fake news by exploring the consistency of multimodal data," Inf. Process. Manag., vol. 58, p. 102610, 2021.

[29] S. Hangloo and B. Arora, "Combating multimodal fake news on social media: Methods, datasets, and future perspective," Multimed. Syst., vol. 28, pp. 2391–2422, 2022.

[30] C. Comito, L. Caroprese, and E. Zumpano, "Multimodal fake news detection on social media: A survey of deep learning techniques," Soc. Netw. Anal. Min., vol. 13, p. 101, 2023.