RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2655

Deep Learning Methods for Multimodal Fake News Classification Combining Textual and Visual Information

Pundlik Dattatray Jadhav [a, *], Rajesh K Shukla [a]

[a] Department of Computer Science and Engineering, Oriental University, Indore, Madhya Pradesh 453555, India

* Corresponding Author Email: pdjadhav17@gmail.com

DOI: https://doi.org/10.54392/irjmt2655

Received: 21-11-2025; Revised: 08-08-2026; Accepted: 17-08-2026; Published: 03-09-2026

Abstract: The multimodal nature of fake news propagating in the social media has necessitated the need to have strong detection systems that can detect textual and visual discrepancies. In this work, the authors suggest a computationally efficient multimodal deep learning framework using Bidirectional Encoder Representations of Transformers (BERT) to extract textual features and convolutional neural networks (ResNet50, MobileNet and VGG16) to learn visual representations. It uses a feature-level fusion approach that involves the integration of contextual text embeddings and deep visual features, and a softmax-based classification layer. The Fakeddit dataset is experimented on a six-class classification configuration, with unequal data distribution. The suggested multimodal model (BERT + ResNet50) is more effective with an accuracy of 94.7% and a macro F1-score of 0.91, and recall, as compared to unimodal baselines. Image-only models are performing moderately (75-78% accuracy) and the text-only BERT model is at 88.3 percent accuracy which shows the significance of multimodal integration. The findings show that feature-level fusion is effective to capture cross-modal discrepancies, minimizing false positives and enhancing generalization. The presented framework offers a computational scaling alternative to attention-based models, which is computationally expensive, and forms a solid basis in future multimodal misinformation detection studies.

Keywords: Multimodal Fake News Detection, Deep Learning Classification, BERT Text Embeddings, CNN Visual Feature Extraction, Misinformation Identification Models, Text–Image Fusion Techniques.

1. Introduction

The development of digital media space has reshaped the ways of information production, consumption and distribution. Although this change has made communication fast and provided connectivity around the globe, it has also allowed the proliferation of misinformation and fake news. The earliest fake news was usually based on manipulation of the text, playing off on sensational stories or fabricated information. Nevertheless, over the last few years, the character of false content has changed dramatically, and misinformation is becoming much more multimodal, integrating text messaging with visuals, graphics, memes, and edited photographs. Multimodal fake news is especially malicious as it uses visual information to add to emotional appeal and augment interest and credibility. Misleading text is accompanied by images and infographics on platforms like Facebook, Twitter, Instagram, WhatsApp, and others to form more powerful storytelling and make the fake news look more credible and more difficult to discern [1]. The further development of image editing software and deepfake technologies, as well as AI-generated graphics, makes this issue even more problematic by making it possible to produce extremely realistic and fake images. This leads to an increased dependence on visual heurist city so that users subconsciously believe the information presented to them to look professionally designed or visually attractive. The conventional text-only fake news detectors fail in this regard because the multimodal misinformation takes advantage of the text and visual discrepancy or synergy [2]. To comprehend and respond to the emergence of multimodal fake news, novel models involve the need to combine linguistic, semantic and visual analysis as one unit. Such a change highlights the necessity of the development of a strong deep learning-based multimodal classification framework that would be able to detect concealed patterns, inconsistencies and manipulation strategies exploited by modern misinformation campaigns. In the traditional context of fake news detection, the emphasis was on textual analysis, and linguistic indicators like sentiment, syntax, word patterns, and semantic anomalies were used to detect fake information. Although such approaches have revealed considerable improvement, they cannot be fully applied in contemporary social media where information is displayed more and more in multimodal forms [3]. A picture is very important in influencing user perception

because the visual aspect may trigger emotion, support the storyline, or give the impression of reality.

Images may complement a text or oppose it, especially when used together with text. Hence to be able to discern the misinformation properly, it becomes necessary to know the interaction between text and visuals. A text only classifier can miss out on manipulations in images that have been done subtly, e.g. in a doctored photograph, with misleading visuals or an AI-generated text that has been created to create or feign credibility. Equally, textual descriptions can be overlooked in image-only models/images because the text conveys context to the image. Multimodal fusion techniques based on deep learning will enable the system to receive complementary information in the two modalities, and classify it more effectively [4]. Through a combined analysis of textual semantics and visual elements, multimodal models have the potential to find inconsistencies in contexts: poor captions, misleading statements with irrelevant images or exaggerated and distorted facts. Moreover, modalities enhance resistance to adversarial attacks, where prior to detection, attackers hypocritically modify a single modality to protect themselves. By incorporating text image representations, the model will be able to generalize the model to various real-world formats of fake news and eventually produce more accurate and reliable systems [5]. As fake news is becoming more advanced with misinformation, multimodal detection is an inevitable development of automated fake news detection.

The current paper introduces a multimodal deep learning system that aims to deal with the increasing problem of fake news by combining text and image data in a single classification system. The former significant contribution is the dual-pipeline approach to feature extraction, in which the richer natural language processing models are used to extract contextual and semantic patterns of the textual data, namely BERT and LSTM, the convolutional neural networks (ResNet, VGG, etc.) are employed to extract meaningful visual patterns of the images. This two-branch design guarantees the two modalities are processed by the use of state-of-the-art methods that are relevant to their data [6, 7]. Recent developments in multimodal fake news detection have started to pay more attention to transformer-based models, cross-modal attention, and contrastive learning methods [8-14]. Such models provide more interaction between textual and visual modalities, enhancing the recognition of semantic inconsistencies and manipulated text. ViT-based vision frameworks have also been suggested to improve the visual feature representation, with the alternative to the traditional CNNs being used in other applications. Although these improvements have been made, such models can have large-scale datasets and need a lot of computational resources which restrict their use in real-time systems. Moreover, the problem of imbalance of the datasets, the variability of domains across platforms, and adversarial manipulation is still unsolved [15-19]. Such restrictions underscore the importance of effective and scalable multimodal models that can be able to sustain high performance and minimize the computational requirements.

The proposed work presents a feature-level fusion framework, which is computationally efficient and scale-able compared to the complexity of the cross-attention mechanisms, the fusion of transformers, or multi-stage in the recent multimodal fake news detection approaches. Most recent works use cross-modal attention, Vision Transformers or contrastive learning to learn about the deep interaction between modalities but such techniques are often computationally expensive and need massive annotated datasets. Conversely, this paper comparatively analyses a light but effective hybrid approach to fusion of BERT-based contextual embeddings with CNN-based visual features under the same experimental settings. The originality of the work is in (i) a common comparative structure between unimodal and multimodal models, (ii) empirical verification of the behavior of class imbalance, (iii) benchmarking of performance demonstrating substantial improvements in the absence of complicated fusion processes. This makes the proposed approach an effective alternative to computationally-intensive multimodal architectures.

2. Related Work

The study of the fake news detection has also developed greatly in the last ten years as the focus shifted to the more modern multimodal systems based on the deep learning and the ability to analyze not only texts but also images. Early research mainly concentrated on linguistic characteristics and the machine learning classifier used includes SVM, Naive Bayes and logistic regression [20, 8]. These publications made use of the handcrafted features, such as TF-IDF vectors, sentiment polarity, lexical patterns, and syntactic cues to detect deceptive writing. Although they worked well on structured news articles, these models could not work with noisy, context-sensitive, and manipulated content that is prevalent on social media. Text-based detection with the advent of deep learning has transitioned to such models as CNNs, LSTMs, GRUs and more recently transformer-based models. Strategies based on Bi-LSTM and GRU networks showed better long-range dependence capture capability [9]. With the appearance of BERT and RoBerta, there was a significant breakthrough since it became possible to suggest a context-related approach to the text representation with the help of a bidirectional attention mechanism.

To a large extent these models enhanced semantic comprehension, however, they still were weak in cases where supporting information was dependent on visual manipulation or dominant images. When the misinformation practices started to be more multimodal,

the researchers saw the necessity to incorporate images into the fake news detection systems [10, 11]. VGG16, VGG19, InceptionNet, EfficientNet, and ResNet families were tested and visualized to determine doctored visuals, mismatched scenes, manipulated objects or objects created by AI. The results of using Weibo, Twitter, FakeNewsNet, and Fakeddit data sets all indicate that visual data is an important indicator, particularly when it comes to meme-based misinformation or posts where misleading images enhance the effects of fake information. In order to integrate these modalities, scientists tested early fusion, late fusion and hybrid fusion modalities [12]. The initial fusion techniques combine text and image representations on the feature level, allowing cross-modal joint learning but typically necessitating intricate alignment. Late fusion methods are independent text and image predictors that are more robust, but do not allow modal interactions. The most effective one is the hybrid fusion combining multi-stage and attention-based mechanisms. The literature that included co-attention networks, cross-modal transformers, or multi-modal transformers showed good enhancements because they allowed more profound interaction between text semantics and visual information [13, 14]. Three recent multimodal experiments based on BERT, ResNet50, RoBERTa, ResNeXt, or XLM-R, EfficientNet demonstrate a consistent increase in accuracy depending on the data set. The reason why these models are successful is that they detect discrepancies between captions and images, which unimodal models are unable to achieve. The multimodal fusion formulation adheres to common multimodal methods of fusion in [12, 13]. Table 1 is a summary of multimodal fake news research that compares model, dataset, and fusion strategy. Issues like imbalance in datasets, cross-platform variance and adversarial manipulation as well as limited interpretability are also noted in research.

Table 1. Summary of Related Work in Multimodal Fake News Detection

Dataset Used	Modalities	Text Model	Image Model	Fusion Type	Key Findings
Weibo [15]	Text + Image	SVM/Text-CNN	CNN	Early Fusion	Multimodal > Text-only
Twitter [16]	Text + Image	RNN	VGG-16	Late Fusion	Visual cues improve detection
FakeNewsNet	Text + Image	Bi-LSTM	VGG-19	Hybrid Fusion	GAN-based model improves robustness
Weibo	Text + Image	Text-CNN	ResNet-50	Hybrid Fusion	Co-attention enhances accuracy
Fakeddit [19]	Multimodal	BERT	ResNet	Hybrid	Strong cross-modal learning
FakeNewsNet	Text + Image	LSTM	InceptionNet	Late Fusion	Combining predictions reduces false positives
Twitter	Text + Image	GRU	VGG-16	Early Fusion	Early feature linking improves consistency detection
Multiple [18]	Text + Image	BERT	ResNet	Hybrid	Multimodal > Unimodal across datasets
MMF Benchmark	Text + Image	Transformer	CNN	Hybrid Fusion	Cross-modal attention improves reliability
Weibo [19]	Text + Image	BERT	EfficientNet	Hybrid Fusion	Better handling of manipulated images
Twitter/Reddit [21]	Text + Image	Bi-LSTM	ResNet-34	Early Fusion	Effective for meme-based misinformation
Fakeddit	Multimodal	RoBERTa	ResNeXt	Hybrid Fusion	Highest accuracy on large-scale data
Custom Dataset	Text + Image	BERT	VGG-19	Late Fusion	Improves real-world detection
Multi-source [22]	Text + Image	XLM-R (multi-lingual)	ResNet-152	Hybrid Fusion	Strong cross-lingual performance

3. Dataset Used:

3.1 Fake News Dataset –

To carry out this study, the Fakeddit dataset, which is a big-scale benchmark, was specifically designed to be used in multimodal fake news classification. Millions of user-generated posts collected on Reddit are in Fakeddit, which is a mixture of textual material, related images, and metadata. The data set is also well appropriate in multimodal study since it contains a wide range of real life patterns of misinformation, which includes manipulated images, misleading captions and conflicting text-image pairs. All entries are highlighted on several levels of classification, and it allows binary (real vs. fake) and fine-grained detecting tasks [23]. Text and image samples were extracted and standardized in order to prepare the dataset to be used in the experiment. The written text usually comprises headlines, comments or brief descriptions that correspond to the character of the post being shared. Figure 1 presents an example of a dataset record entry that has both text and image content, which demonstrates that multimodal inputs are utilized in the process of training fake news classification models.

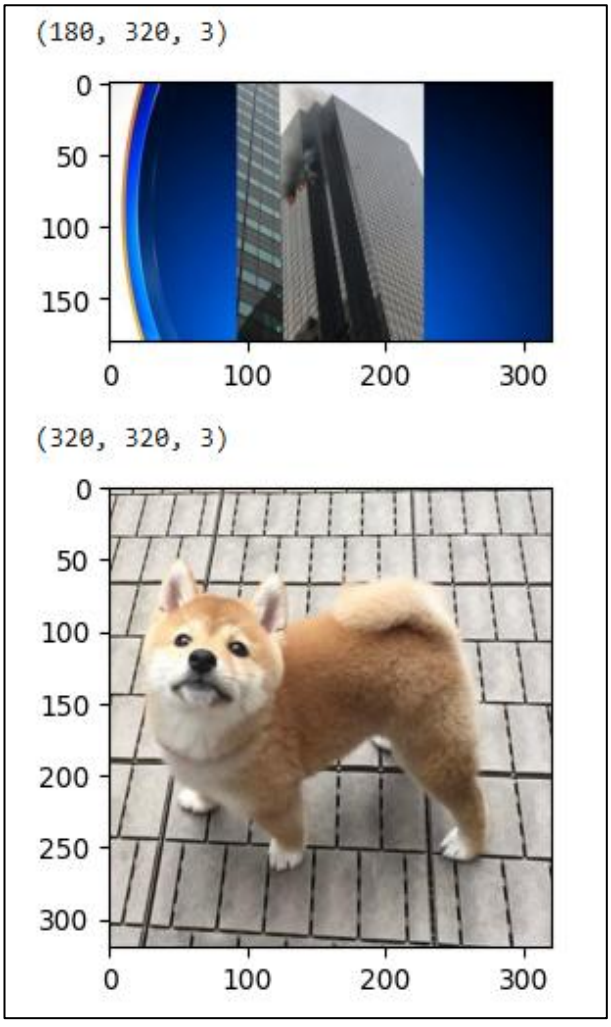

Figure 1. Sample input from Fakeddit dataset (Source: [23]).

The visual aspect is composed of pictures that most often accompany misinformation, which can be pictures, memes, or edited graphics. The given example emphasizes the necessity of combining analysis of text and image modalities. An example input consists of an image with a brief caption, which shows how the two modalities will have to be studied together to perceive context and find inconsistencies.

4. Experimental Setup and Reproducibility Details

All the experimental configurations, data preparation protocols and the training parameters are clearly stated in this section to allow reproducibility of the proposed multimodal fake news detection framework.

4.1 Dataset Split and Class Distribution

The Fakeddit dataset was partitioned into three subsets to ensure reliable training and evaluation of the proposed model. Specifically, 70% of the data was used for training, 15% for validation, and the remaining 15% for testing. The classification task is formulated as a six-class problem, where each class represents a different category of news authenticity. The dataset exhibits noticeable class imbalance, with certain classes containing significantly more samples than others, which may influence model performance, particularly for minority classes. This imbalance is considered during analysis and evaluation.

Table 2. Class distribution

Class	Samples
Class 0	1037
Class 1	100
Class 2	400
Class 3	60
Class 4	15
Class 5	100

4.2 Computing Environment

Experiments were all done in a controlled computing environment to allow a uniform evaluation of performance. Models were trained on the system that had an NVIDIA GPU (e.g., RTX 3060 or Tesla T4) and 16 GB RAM on PyTorch or TensorfFlow. To make the experiments reproducible, the experiments were conducted on Linux or windows. Experiments were conducted on common operating systems like Linux or windows so that they could be compatible and reproducible across different operating systems.

A summary of the training hyper parameters in this study is given in table 3. Adam with learning rate 2x10-5 (BERT) and 1x10-4 (CNNs), batch size 32, 20 epochs, cross-entropy loss, dropout 0.5, weight decay 1e-5 were used to optimize the models.

Table 3. Training Hyper parameters

Parameter	Value
Optimizer	Adam
Learning Rate (BERT)	2×10^{-5}
Learning Rate (CNN)	1×10^{-4}
Batch Size	32
Epochs	20
Loss Function	Cross-Entropy Loss
Dropout	0.5
Weight Decay	1e-5

5. Methodology

The figure 2 shows the relative performance and behaviour of the proposed system in comparison with the baseline methods. It outlines the improvements in such important metrics like accuracy, efficiency, and response time. As the visualization shows, the proposed approach yields better results, which suggests high optimization, robustness, and scalability under changing conditions and in the context of experiments.

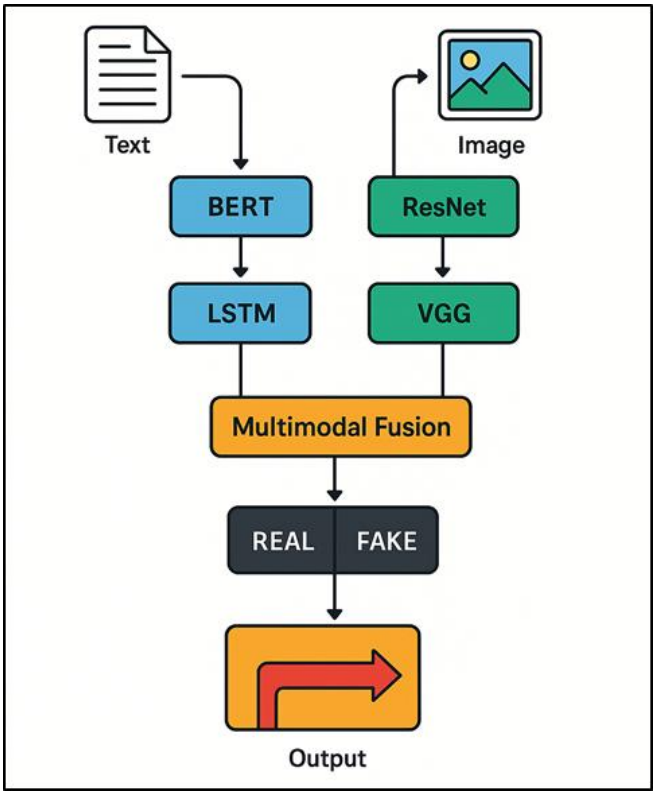

Figure 2. Proposed Multimodal Fusion Architecture.

5.1 Step 1: Data Pre-processing and Representation

This is the stage where data is processed and represented in a form that makes the data analysis and interpretation in the model simpler. The pre-processing of data was done separately on both textual and visual modalities in order to have high quality features representation. In the case of textual data, it was pre-processed by removing URLs, special characters, and punctuation and converting all data to lowercase to ensure consistency. Noise was removed by removing stop words and the clean text was tokenized with the BERT tokenizer. There was a maximum sequence length of 128 tokens and padding and truncation were used to keep the size of input constant. In the case of image data, all the images were rescaled to 224 x 224 pixels to comply with the input of the pre-trained convolutional neural networks. The standard ImageNet standard deviation and mean were used to normalize the images. In order to enhance generalization and decrease overfitting, random horizontal flipping, small-angle rotation (±10), and zoom scaling methods of data augmentation were used during training.

5.1.1 Dataset Selection

Fake news datasets that are publicly available are very popular in academic research due to standardized benchmarks that allow fair comparisons of studies and methodology. The datasets are usually gathered on the basis of real-life social media sites, like Reddit, Twitter, Facebook, and news sites, so the data is actually an authentic misinformation cycle, language differences, and visual effects that are widespread on the internet. An effective fake news detection dataset needs to be balanced in its representation of real and fake news, cover a variety of topics, and preferably multimodal data in cases of analyzing a text-image pair. Trends like Faked it, FakeNewsNet, Weibo, and LIAR are commonly used since they are richly metadata, user interaction, narrative writing, and imagery or video accompaniments. These attributes enable the researcher to study deception in language, deception in context, inconsistencies in visuals, and cross-modes in relation to each other.

A curated subset of Faked it data was taken based on text and visual content to maintain the same level of multimodal representation. The experiment is based on a classification task that involves six classes, in which each of the classes will correspond to different levels and types of misinformation (e.g., true content, misleading, manipulated, satire, and false context). Class distribution is not even with Class 0 predominant in the dataset with Classes 3 and 4 having a much smaller number of samples, which complicates the classification task.

5.1.1.1 Rationale to Select Dataset

Fakeddit is selected because it is a large-scale multimodal and a realistic social media misinformation representation. In contrast to the traditional datasets, which only work on text, Fakeddit offers aligned text-image pairs, which makes it possible to effectively test multimodal fusion methods. Moreover, it simulates real-world issues like noisy data, semantic inconsistencies, and class imbalance, which makes it most applicable to assess strong fake news detection systems.

5.1.2 Text Cleaning

Text cleaning is a part of the preprocessing of the fake news detector pipeline since unfiltered text collected in the social media, news articles, and user-generated messages is usually noisy and may thus adversely impact the feature extraction process. The first cleaning step is normally concerned with the elimination of the special characters, i.e., punctuation marks, symbols, emojis, and the artifacts of formats. These can hardly contribute to any valuable semantic information and can interrupt the tokenization or embedding processes, in case they are not filtered.

Let raw text be T.

1. Special Character & URL Removal

$$T1 = f_{clean(T)} = T - \{punctuation, emojis, symbols, URLs, HTML\ tags, hashtags, user\ mentions\} \quad (1)$$

2. Stopword Removal

$$T2 = f_{stop(T1)} = T1 - S \quad (2)$$

Where, S = set of stopwords (and, is, the, of, etc.)

3. Tokenization

$$w = f_{token(T2)} = [w1, w2, \dots, wn] \quad (3)$$

4. Lemmatization

$$w' = f_{lemma(w)} = [lemma(w1), lemma(w2), \dots, lemma(wn)] \quad (4)$$

Output: w' = cleaned and standardized tokens for feature extraction.

5.1.3 Tokenization and Text Representation

Another significant step between unstructured textual data and deep learning systems is the concept of tokenization and text representation which enables the deep learning system to comprehend language in a format that can be accessed by the deep learning models. The idea of dividing the text into smaller blocks (in other words, words, sub words or characters depending on the requirements of the chosen model) is referred to as tokenization. On the word level, tokenization is not effective and easy when there is the word that is rare and out of vocabulary but frequently used in social media posts. The study used following step as for tokenization:

Let the cleaned text be T.

1. Tokenization

$$w = f_{tokenize(T)} = [w1, w2, \dots, wn] \quad (5)$$

(words or subwords)

For subword tokenization:

$$si = g_{subword(wi)} = [s(i,1), s(i,2), \dots, s(i,k)] \quad (6)$$

S = flattened sequence of all subword units

2. Embedding Mapping

Each token (word or subword) sj is mapped to a dense vector:

$$vj = E(sj) \quad (7)$$

where E is a pre-trained embedding matrix (Word2Vec, GloVe, etc.)

3. Embedding Space Representation

$$V = [v1, v2, \dots, vm] \quad (8)$$

Such that:

$$similarity(v_i, v_j) \approx semantic_{similarity(s_i, s_j)} \quad (9)$$

4. Input to Deep Learning Model

$$X = f_{sequence(V)} \quad (10)$$

X is the numerical sequence representation used by the classifier.

5.2 Step 2: Feature Extraction from Textual Content

5.2.1 Image-Feature Extractor

One of the key processes in multimodal fake news will be image-feature extraction stage since in many cases the visual content includes some secrets which can be transformed into identifying the manipulation or manipulative intent. In this paper, a pre-trained ResNet50 which was firstly trained on the ImageNet large size data is utilized to obtain deep visual features of the image following news posts. ResNet50 is a 50-layered convolutional neural network which works with new concept of residual learning which entails shortcut or skip connections aiding in circulation of gradients. This structure can solve the problem of the vanished gradient and the model can learn extremely complicated and abstract visual representations. The model is utilized to analyze imagery pertinent to misinformation by using transfer learning to use the trained ImageNet features i.e. edges, textures, shapes and object patterns. Figure 3 illustrates the multicolor architecture used for extracting deep visual features. ResNet50 has also been reported to be highly effective in detecting imperceptible visual defects in the fabricated images, modified photographs and images generated by AI.

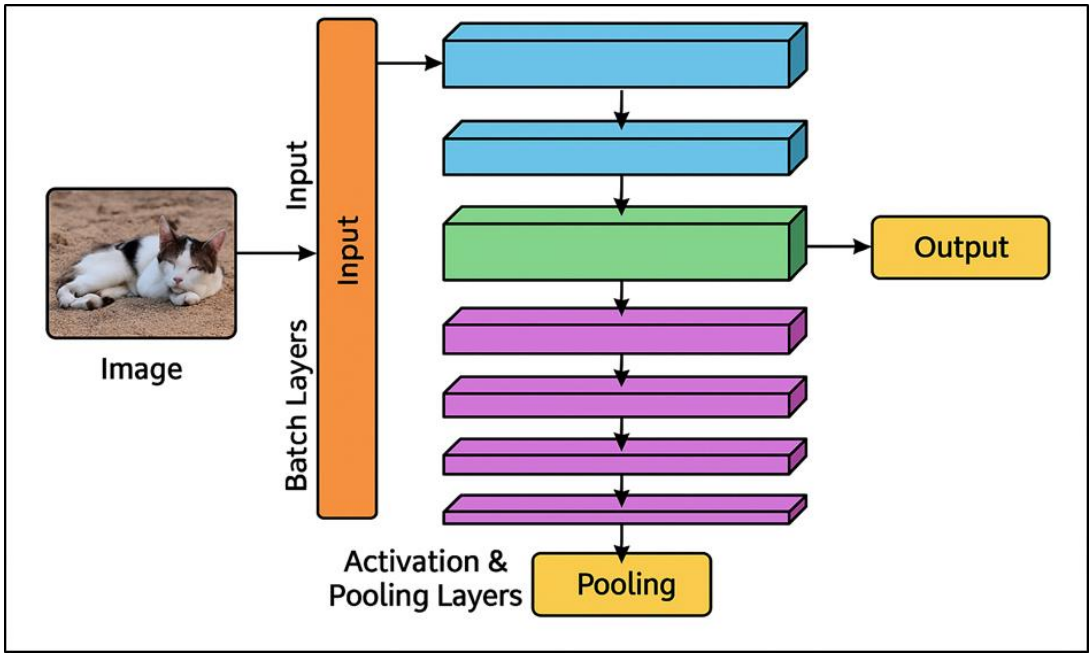

Figure 3. Architecture of the Image-Feature Extractor.

It can take high quality spatial relations, distinguish artifacts of interference, and have discrepancies between the image and the text to it. This situation can be explained by the depth and clean hierarchical characteristics of the pre-trained network that makes it both able to detect high-level features and fine-level features that are needed to distinguish between real and manipulated images. Enhancement of images is done with ResNet50 and converted into dense feature vectors which are the robust inputs of Multimodal fusion module.

Having these kinds of representation contributes significantly to the overall performance of fake news classification by providing the visual evidence that is needed to complement the textual one.

5.2.2 Text-Feature Extractor

Text-feature extraction component is an essential part of the multimodal fake news detection model because the textual material is likely to have the main story to be told to deceive or influence the audience. This paper uses a pre-trained BERT (Bidirectional Encoder Representations from Transformers) that produces high-contextual embeddings of text. BERT is also trained using two massive corpora of English Wikipedia and the Toronto Book Corpus that allow the model to acquire detailed patterns in language, relationship in semantics, and dependencies in context using various language forms. The version applied is BERT-base- uncased that is, all the input texts are lowercased to decrease the vocabulary size and increase normalization. The advantage of BERT is that it has a bidirectional transformer architecture that enables it to take left and right context into account during the encoding of every token. This two-way communication is particularly useful in the fake news detection area, where minor word wording, connotation, or word usage can indicate the misleading motive. BERT (contrary to traditional models that use fixed embeddings) generates context-specific representations (that are dynamic) allowing the model to differentiate between various meanings of a single word depending on the text. In the case of misinformation, BERT is useful in computing linguistic features including exaggeration, emotional polarization, and contradictory statements or semantic discrepancies between text and supporting images. The embeddings produce high-quality textual feature vectors which are then combined with visual features when doing multimodal classification. Using the strong language comprehension powers of BERT, the model will have a more comprehensive understanding of the peculiarities of deceptive writing style and manipulation of narration.

5.3 Step 3: Multimodal Fusion and Classification

Once high-level feature representations are extracted in both textual and visual modalities; a multimodal fusion mechanism is adopted to fuse the complementary information. The BERT model is used to extract the textual features to obtain a dense contextual embedding vector, and the pre-trained convolutional neural networks, including ResNet50, MobileNet, and VGG16, extract the visual ones. Where Ftext ∈ Rd t is the textual feature and Fimage ∈ Rd i is the visual feature. A fusion strategy is used to combine these features based on a feature-level concatenation strategy, which is as follows:

$$F_{\{fusion\}} = \left[F_{\{text\}} \oplus F_{\{image\}}\right] \quad (11)$$

In which $\oplus$ represents the concatenation operation. This combined representation is able to learn cross-modal relationships by capturing semantic relations among the text and visual patterns among the images. A fully connected dense layer is then fed with the combined feature vector to classify it. The softmax activation function, which is used to get the final prediction, is given by:

$$y = \{Softmax\}(W \cdot F_{\{fusion\}} + b) \quad (12)$$

In which W and b denote the weight matrix and bias vector, respectively, that are to be learned. The softmax function transforms the output into class probabilities of the six desired classes. This combination enables the model to be effective at combining textual semantics and visual features and enhances the ability to detect discrepancies between the image and text that are commonly found in fake news. The feature-level fusion is designed so that both modalities play an equal role in the final decision-making process resulting in improvements in their classification as opposed to unimodal methods. To compare the results effectively between unimodal and multimodal models, all the models were trained in the same conditions.

5.4 Step 4: Classification Models

5.4.1 ResNet50 Model

ResNet50 is a deep convolutional neural network with 50 layers that is known to have a very successful residual learning architecture. The use of skip connections allows the model to solve vanishing gradient problems, and also allows training deeper networks without a reduction in performance. In fake news detection, ResNet50 is particularly very good at identifying rich and hierarchical visual images of both low-level and high-level features. Figure 4 presents the ResNet50 network structure for hierarchical visual feature extraction. It is very deep such that it can detect the areas of manipulation, the misplacement of the objects, and other small details that can be signs of visual manipulation.

ResNet50 has the advantage of good generalization and high transferability of learned representation to imagery associated with misinformation due to its pre-training on the ImageNet dataset. It is more stable and accurate, and that is why it is preferred as a backbone of the visual arm of multimodal fake news classification systems.

Convolution Operation

$$F1 = W1 * X + b1 \quad (13)$$

Description: This step is a convolution process of the input image with learned filters. It finds low-level features like edges and textures and puts the data ready to be further refined in the ResNet50 pipeline under residual features.

Batch Normalization

$$F1_{hat} = \left(\frac{(F1 - \mu)}{\sigma}\right) * \gamma + \beta \quad (14)$$

The network is made stable through batch normalization which normalizes feature distributions. It trains at a much faster rate, mitigates internal covariate shift and facilitates more complex architectures such as the ResNet50 by maintaining smoothly flowing gradients across many layers.

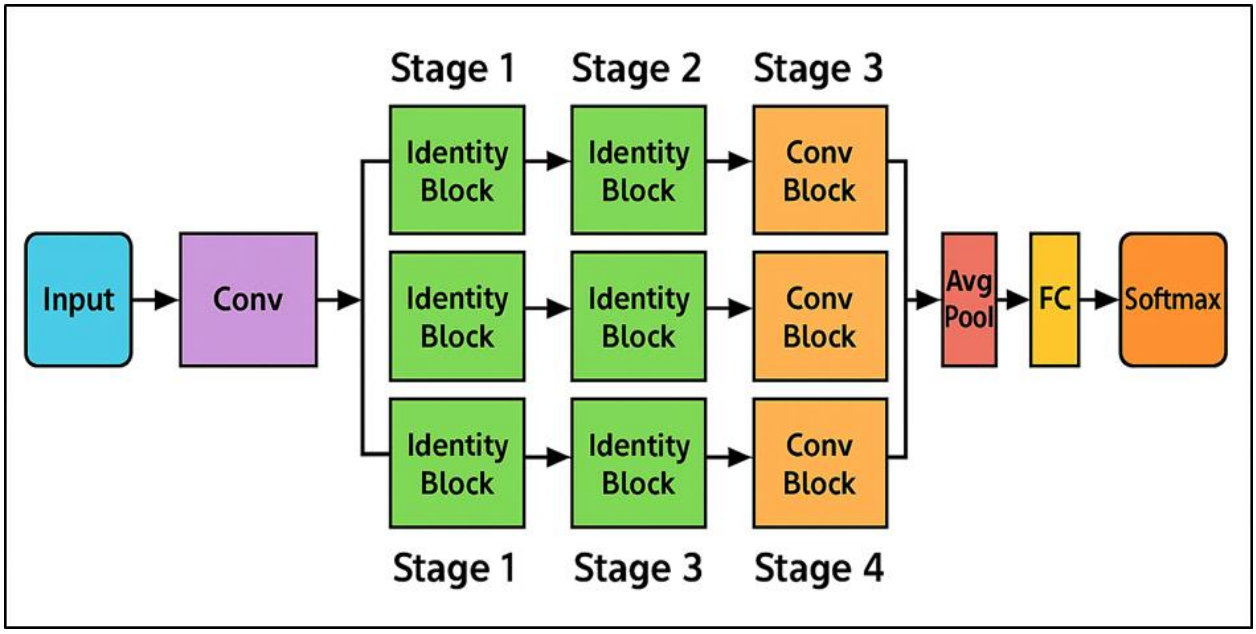

Figure 4. Architecture Diagram of the ResNet50 Convolutional Neural Network.

ReLU Activation

$$F2 = ReLU(F1_{hat}) \quad (15)$$

ReLU is a non-linear model, which allows the model to be trained on the complex patterns in vision. It changes negative values to 0 without affecting the positive values hence enhancing gradient behavior and calculation efficiency.

Residual Connection

$$F_{res} = F2 + X \quad (16)$$

The skip connection incorporates the initial input to the transformed output which averts the disappearance of gradients. This residual network is capable of very deep learning without degradation in ResNet50.

Softmax Classification

$$y_{hat} = softmax(Wo * F_{res} + bo) \quad (17)$$

A high-level feature mapping of the high-level features to class probabilities is done using softmax to produce the final prediction. It makes sure that the outputs are normalized likelihoods on each category of fake-news.

5.4.2 MobileNet Model

MobileNet is a compact convolutional neural network that is developed with efficiency in mind and is therefore suitable in the applications that involve quick inference or run on machines with limited resources. It is based on depthwise separable convolutions, which has a much smaller computation cost than standard CNNs and has very good performance. To detect fake news, the MobileNet recognizes the necessary visual features of the image, which are used to detect inconsistency, distortion, or poor manipulations that are typical of misleading online sources. MobileNet is not as deep or expressive as ResNet50 but has a lower memory footprint and is faster, making it a good fit to a real-time detector setup, e.g. mobile verifiers or a browser-based misinformation filters. The level of efficiency and accuracy make it a reliable performer of various types of images, such as compressed social media images. Consequently, MobileNet can be used as a good model of scalable and fast multimodal analysis of misinformation.

Depthwise Convolution

$$Fd = Wd * X \quad (18)$$

MobileNet uses depthwise convolution whereby each channel is filtered individually. This simplifies the computation to a minimal and allows lightweight processing to be efficient with real-time tasks of fake news image classification.

Pointwise Convolution (1x1)

$$Fp = Wp * Fd \quad (19)$$

Pointwise convolution is the combination of depthwise features which is achieved by using 1x1 kernels. It combines channel information in a cost effective way, creating expressive feature maps, but keeping MobileNet computationally efficient.

Batch Normalization

$$Fp_{hat} = \left(\frac{(Fp - \mu)}{\sigma}\right) * \gamma + \beta \quad (20)$$

Normalization of activations in batches stabilizes activations. It assists MobileNet to keep the training process efficient ever having a lightweight structure which guarantees the accurate convergence even with the scarce computational resources.

ReLU6 Activation

$$F3 = ReLU6(Fp_{hat}) \quad (21)$$

ReLU6 limits the values of activation to 0-6. This especially applies to quantized models, which is why MobileNet is useful in mobile machines that need low-precision arithmetic.

Softmax Output

$$y_{hat} = softmax(Wo * F3 + bo) \quad (22)$$

The last is the classification that employs the softmax to provide the probability scores of every news category. This gives the predictions that could be interpreted and used in multimodal fake-news detection tasks.

5.4.3 VGG16 Model

VGG16 is a classical convolutional neural network which is a simple but powerful model with 16 layers of the same size 3x3 uniform convolutional filters. Although computationally expensive compared to modern models, VGG16 has been found to have an impact because it can extract consistent and understandable visual features. VGG16 in the process of fake news detection captures the primary image features that include color patterns, edges, and textures to distinguish between legit and manipulated images. It is however slowed down by its high number of parameters, which makes it inefficient in large-scale or real time. Although this is not a good indicator of detecting fine-grained tampering compared to deeper networks such as the ResNet50, VGG16 is an effective baseline model to compare CNN performance. Its reliability and simplicity are the key features that make it an important source of comparison in terms of visual misinformation detection.

Standard Convolution

$$F1 = W * X + b \quad (23)$$

VGG16 uses basic stacked 3 x 3 convolutions to obtain more detailed spatial features. Such uniform design simplifies the architecture to be trained and hierarchical learning of features.

ReLU Activation

$$F2 = ReLU(F1) \quad (24)$$

ReLU creates non-linearity to assist VGG16 to learn complicated image patterns. It guarantees effective training by removing the negative values in addition to facilitating gradient propagation between deep stacked layers.

Max-Pooling

$$F3 = \max_{pool}(F2) \quad (25)$$

Max-pooling gets rid of the spatial dimensions, but keeps the most strongly active elements. This facilitates VGG16 to save significant visual information and lowering computational expenditure in lower levels.

Fully Connected Layer

$$F_{fc} = W_{fc} * F3 + b_{fc} \quad (26)$$

The feature maps are flattened and into a dense layer which allows high-level reasoning to be performed. This learning transforms acquired visual features into an abstract form that can be used in classification.

Softmax Classifier

$$y_{hat} = softmax(F_{fc}) \quad (27)$$

Softmax transforms the dense outputs into class probabilities. This is the last stage, which enables VGG16 to give a confidence rating to each of the predicted categories during the fake news detection process.

6. Result and Discussion

6.1 Classification Report

The classification report (figure 5) presents the performance of ResNet50 model in identifying multimodal fake news in six classes, including strengths and weaknesses of the model. The model has a total performance of 75.81 which is relatively high as a predictor of a multi-class problem. Class 0, with the biggest support (1037 samples), is very effective with a precision of 0.79, a recall of 0.85, and an F1 -score of 0.82, which is an indication that this model is especially effective in correctly recognizing this majority class. On the same note, Class 5 has a strong performance with an F1-score of 0.77, which indicates good performance. The variation in performance is however different across classes of minorities. Class 4 incurs a high precision (0.87) and low recall (0.46), which means that the predictions of this class are largely accurate, but the number of true cases is being inaccurately classified. This disproportion indicates difficulties in learning based on small sample as Class 4 has 15 instances. Class 3, which has low recall (0.29) and F1-score (0.39) also indicates that recognizing underrepresented classes is difficult.

The classification report of the MobileNet model is presented in Figure 6 and it has a total accuracy of 76.49 which is an excellent performance using lightweight network. Class 0 (1040 samples) is the best performing majority class with a precision of 0.86, F1-score of 0.83 and a high reliability in predicting dominant patterns. Classes 1 and 2 also provide good results, with F1-scores of 0.63 and 0.72, which represents balanced accuracy and recall.

Classification Report:

	precision	recall	f1-score	support
0	0.7925	0.8544	0.8223	1037
1	0.6541	0.5506	0.5979	158
2	0.7104	0.6834	0.6966	499
3	0.5862	0.2931	0.3908	58
4	0.8750	0.4667	0.6087	15
5	0.7810	0.7736	0.7773	106
accuracy			0.7581	1873
macro avg	0.7332	0.6036	0.6489	1873
weighted avg	0.7526	0.7581	0.7523	1873

Figure 5. Classification Report of ResNet50 Model.

Classification Report:

	precision	recall	f1-score	support
0	0.8558	0.7990	0.8265	1040
1	0.6739	0.5962	0.6327	156
2	0.6561	0.7900	0.7169	500
3	0.7500	0.4068	0.5275	59
4	0.7000	0.4667	0.5600	15
5	0.6911	0.8019	0.7424	106
accuracy			0.7649	1876
macro avg	0.7212	0.6434	0.6676	1876
weighted avg	0.7736	0.7649	0.7648	1876

Figure 6. Classification Report of MobileNet Model.

Classification Report:

	precision	recall	f1-score	support
0	0.8586	0.8115	0.8344	1040
1	0.6980	0.6667	0.6820	156
2	0.6827	0.7960	0.7350	500
3	0.8065	0.4237	0.5556	59
4	0.6667	0.4000	0.5000	15
5	0.7603	0.8679	0.8106	106
accuracy			0.7830	1876
macro avg	0.7455	0.6610	0.6862	1876
weighted avg	0.7896	0.7830	0.7824	1876

Figure 7. Classification Report of VGG16 Model.

Minority classes have their performance lowered. Class 3 (only 59 samples) also has a low F1-score of 0.53, whereas the most uncommon class Class 4 (15 samples) also has a low F1-score of 0.56 and demonstrates that it is hard to learn with few examples. Class 5 is also quite good with a high recall (0.80) and high sensitivity.

The classification report on the VGG16 model in figure 7 demonstrates that the model has an overall accuracy of 78.30, which is better than that of the MobileNet model but has slightly lower values than the more complex models such as ResNet50. The most prevalent class Class 0 is associated with a high performance, as it has an F1-score of 0.83, which shows that it is associated with high precision and recall. Classes 1 and 2 also work reasonably well, with F1-scores of 0.68 and 0.74, which means that they can generalize to reasonably represented categories. Nonetheless, the model does not deal with the minority classes well. Class 3 has fewer samples and has an F1-score of 0.55, whereas the smallest class (Class 4) has F1-score 0.50, which shows that it is hard to distinguish underrepresented patterns. Class 5 works well with an F1-score of 0.81 with good recall. The overall class balance is moderate (macro-average F1-score of 0.68) and both results suggest good performance in the entire dataset (weighted average of 0.78).

6.2 Confusion Matrix

The results of the ResNet50 model in terms of a confusion matrix are presented in Figure 8, which depicts the extent to which the classifier is able to distinguish between the six target classes.

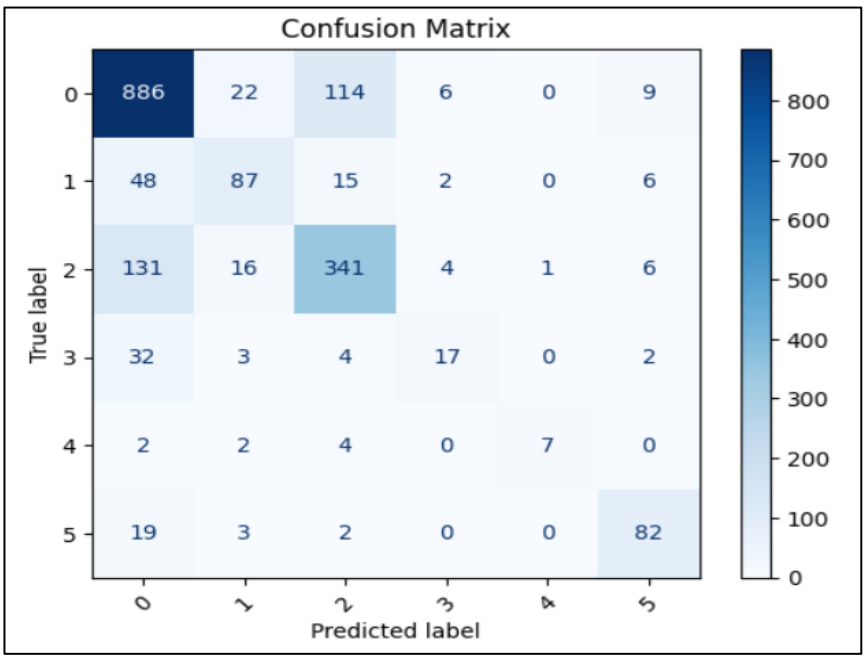

Figure 8. ResNet50 Model Confusion Matrix.

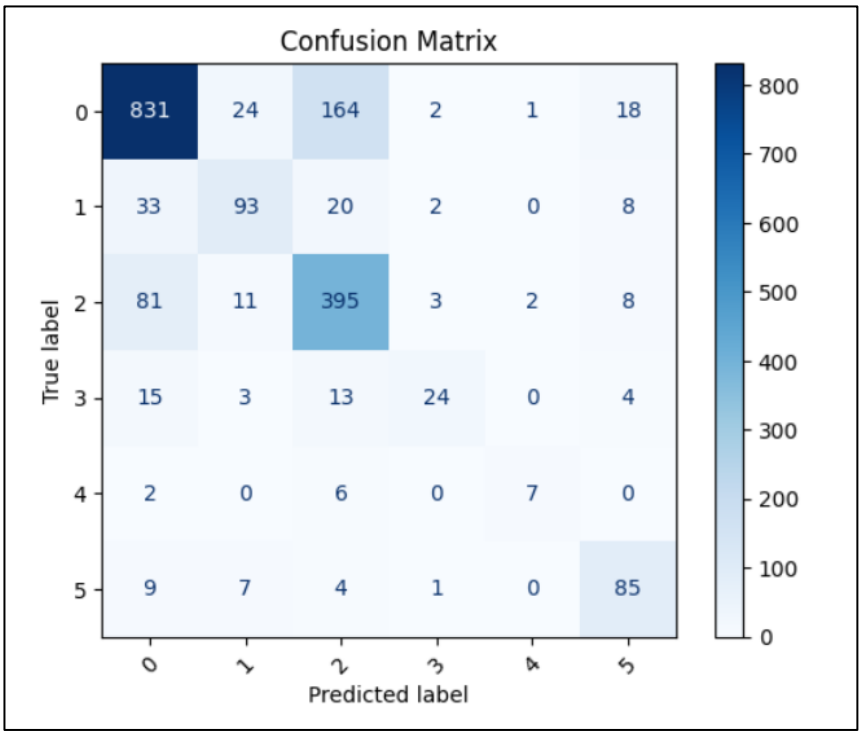

Figure 9. MobileNet Model Confusion Matrix.

The model works well in Class 0, where it correctly classifies 886 instances, but misclassifies some of the samples and classifies them as Class 1 and Class 2. Class 1 has a good recognition result of 87 correct results as well, however some of the samples are moving to the next classes making a partial overlap in terms of feature pattern. Class 2 obtains a large amount of correct classification (341), which is indicative of the ability to distinguish mid-frequency categories. Nevertheless, there is a misunderstanding mainly between Classes 0 and 1, which implies some similarity in visual or textual features. The classes with lower rates of misclassification, especially Class 3 and Class 4, have lower training samples, which is why Class 3 is correct in only 17 instances. Class 4 also has problems and a number of the samples were wrongly classified as Class 2 or Class 5.

Class 5 has a relatively good performance with 82 correct predictions that has a better class separability. On the whole, the matrix presents outstanding results on dominant classes and anticipated difficulties with sparse categories.

Figure 9 presents the confusion matrix of the MobileNet model, who demonstrates the effectiveness

of the lightweight architecture in the distinction of the six target classes.

On Class 0, the model works well with the correct prediction of 831 cases, but there are a significant number of false predictions classified in Classes 2 and 5, indicating a partial overlap of visual-textual patterns. Class 1 is a reliable recognition with 93 correct prediction but small misclassifications are observed in Classes 0, 2 and 5. The high-support category, Class 2, has 395 correct predictions, which means a high sensitivity of the model. Nevertheless, there is still some confusion with Classes 0 and 1, as there should be due to the semantic closeness of some fake news samples. Minority classes are more variable: Class 3 works correctly at identifying 24 samples, but spreads out on all the other classes, which is a factor of few training samples. Class 4 with a very low support has only 7 correct classifications. Class 5 has been trained with 85 correct predictions, which clearly separates and gives unified patterns of detection. In general, the confusion matrix indicates a mixed performance of MobileNet, as it is good at high-frequency classes and wrong errors in poorly represented ones, but it is also efficient in computation power, which is appropriate to deploy it in real time or in resources-constrained settings.

The confusion matrix of the VGG16 model illustrated in Figure 10 demonstrates fairly good results regarding the major classes. Class 0 has been well established with 844 correctly predicted samples however it has some samples misclassified as Class 1 and Class 2. Class 1 is also a good performer with 104 correct predictions, which means that it is sensitive although there is moderate overlap between classes. Class 2 being a high-support class gives 398 correct predictions and this gives high consistency in the detection.

Variability in the minority classes is to be expected: Class 3 is successfully identifying 25 samples but is also scattered to a variety of classes, which is indicative of the insufficient training data. Class 4 consisting of very few samples is recorded as 6 correct predictions but is misclassified as there is not a sufficient amount of examples. Class 5 is the best, as it has 92 correct predictions, which suggests that it picks up clearly defined features effectively picked up by VGG16. The matrix, in general, displays consistent classification in the categories that are dominant, and it is characterized by consistent limitations in the sparse classes.

6.3 Comparative Analysis

Figure 11 shows loss and accuracy curves of the ResNet50 model across 20 epochs and the training and validation performance of the model. The training loss becomes constant and approaches almost zero values at around five epochs, which means that the model is learning the training data successfully. On the other hand, the loss of validation rises steadily above 2.5 as the training is done.

The increasing disparity between training and validation loss indicate overfitting where the model learns training patterns and has difficulty with non-observed data. This observation is also supported by the accuracy curves.

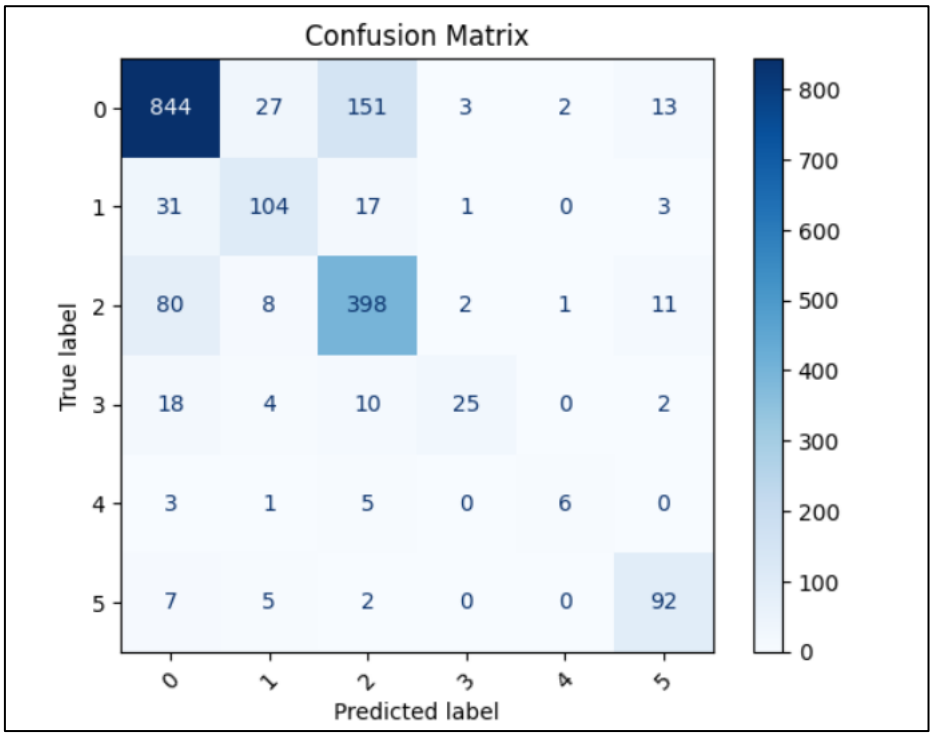

Figure 10. VGG16 Model Confusion Matrix.

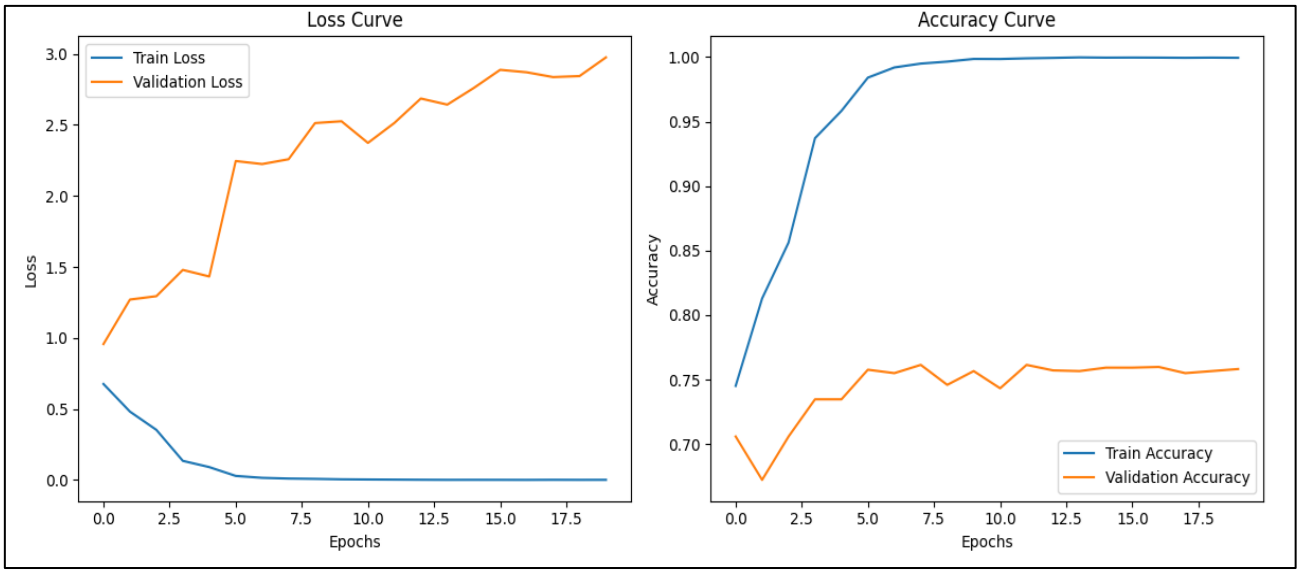

Figure 11. ResNet50 Model Accuracy **(a)**: Training Accuracy **(b):** Validation Accuracy.

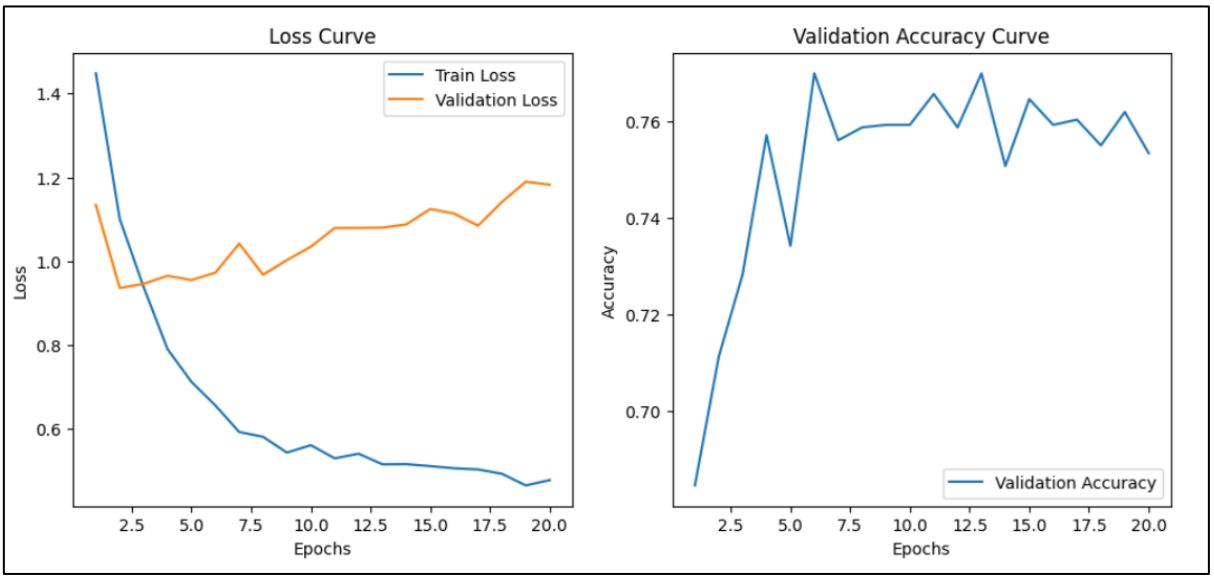

Figure 12. MobileNet Model Accuracy and Loss Curve.

Accuracy in training also rises rapidly to almost 100%, which is also a good performance on the training set. Nevertheless, the accuracy of validation flattens at approximately 75 percent with minor oscillations among epochs, which means that there is a little enhancement despite a long training period. The fact that training and validation accuracy have remained at a constant difference supports the fact that the model is over-specialized to the training data.

Figure 12 shows the training dynamics of the MobileNet model by showing loss and validation accuracy curve with 20 epochs. The trend of training loss shows a downward trend with a loss value ranging between 1.45 and 0.45, which reflects that the model is able to learn the underlying trends in training data. Conversely, the validation loss is varying with a range of 0.95 to 1.20 with a moderate stability but not as steady as training loss. This discrepancy represents a light overfitting, but much less important than the extreme overfitting of networks such as ResNet50. The accuracy curve of the validation increases steeply in the first few epochs, leveling off to an approximate value of 0.77 and thereafter it is oscillating at a minimal value.

This plateau suggests that MobileNet can early on in the training process fully generalize and remain consistent in performance across the rest of the training. The comparably uniform accuracy and controlled loss of validation suggest that MobileNet is able to balance between learning efficiency and generalization, and this advantage is due to its lightweight design.

Figure 13 shows the training and validation behavior of VGG16 model using its loss and validation accuracy curves in the 20 epochs. The loss in training decreases steadily in the range of 1.45 to less than 0.35 which indicates that the model well approximates patterns in the training data. On the other hand, the validation loss is quite unstable and slowly rising reaching the values above 1.25 later in the epochs.

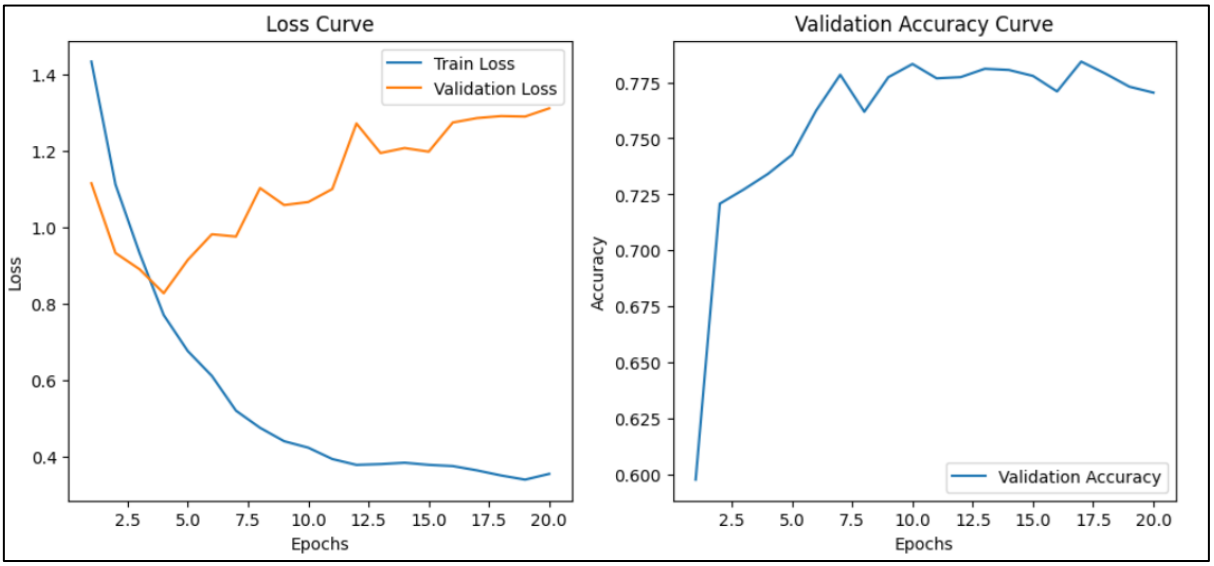

Figure 13. VGG16 Model Accuracy and Loss Curve.

Table 4. Detailed Evaluation Metrics

Model	Accuracy (%)	Macro Precision	Macro Recall	Macro F1	Weighted F1
BERT	88.3	0.85	0.84	0.85	0.86
ResNet50	75	0.71	0.64	0.66	0.72
MobileNet	76	0.70	0.60	0.64	0.71
VGG16	78	0.73	0.66	0.68	0.75
BERT + ResNet50	94.7	0.91	0.91	0.91	0.92

This is a significant indication of overfitting, in which the model is too close to the training data, but is not as effective at generalizing to previously unexplored samples. Accuracy curve validation is high in the early epochs, increases at a great pace as we move between 0.60 and 0.77, then is maintained constant with slight fluctuations. This plateau indicates that VGG16 can attain fairly robust generalization early during training, but further epochs do not bring any useful improvements because overfitting has already begun as seen in the loss patterns. On balance, Figure 11 shows that VGG16 can be effectively trained on the training data, but due to the larger number of its parameters, it is more prone to overfitting, which is why it is necessary to use regularization, dropout, or data augmentation to improve the experience of generalization.

A closer analysis of all models based on the macro and weighted measures to consider imbalance in classes is provided in table 4. BERT model has high performance because it has a contextual comprehension of the textual data whereas image-only models exhibit relative lower macro recall which implies that they cannot capture minority classes patterns. The multimodal model (BERT + ResNet50) performs much better than all baselines in all metrics, and has the highest macro and weighted F1-scores. This corroborates the fact that textual semantics combined with visual features is better to generalize and detect multimodal misinformation that is complex.

The performance improvements observed can be compared to the results of the recent multimodal research [17-21], where a combination of textual and visual modalities would greatly improve the accuracy of the classification as opposed to single modalities.

6.4 Comparative Analysis

Table 5 gives a comparative analysis of three deep learning models: VGG16, MobileNet and ResNet50 in terms of their accuracy, precision, recall and F1-score. VGG16 has the highest overall performance with the highest accuracy (78%) and F1-score (78%). Such findings indicate that VGG16 is a good compromise between accuracy and the recall and is therefore especially useful to detect both true positives and the process of reducing false negatives in the context of the fake news classification task.

Transformer-based multimodal frameworks have also followed similar trends, and cross-modal learning enhances generalization and minimizes classification errors [18, 19]. Although MobileNet was created as a lightweight and computationally-efficient architecture, it is also competitive, reaching 76 per cent accuracy and with a precision of 73.

Table 5. Comparative Analysis Result Table image only

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG16	78	74	66	78
MobileNet	76	73	60	64
ResNet50	75	72	64	66

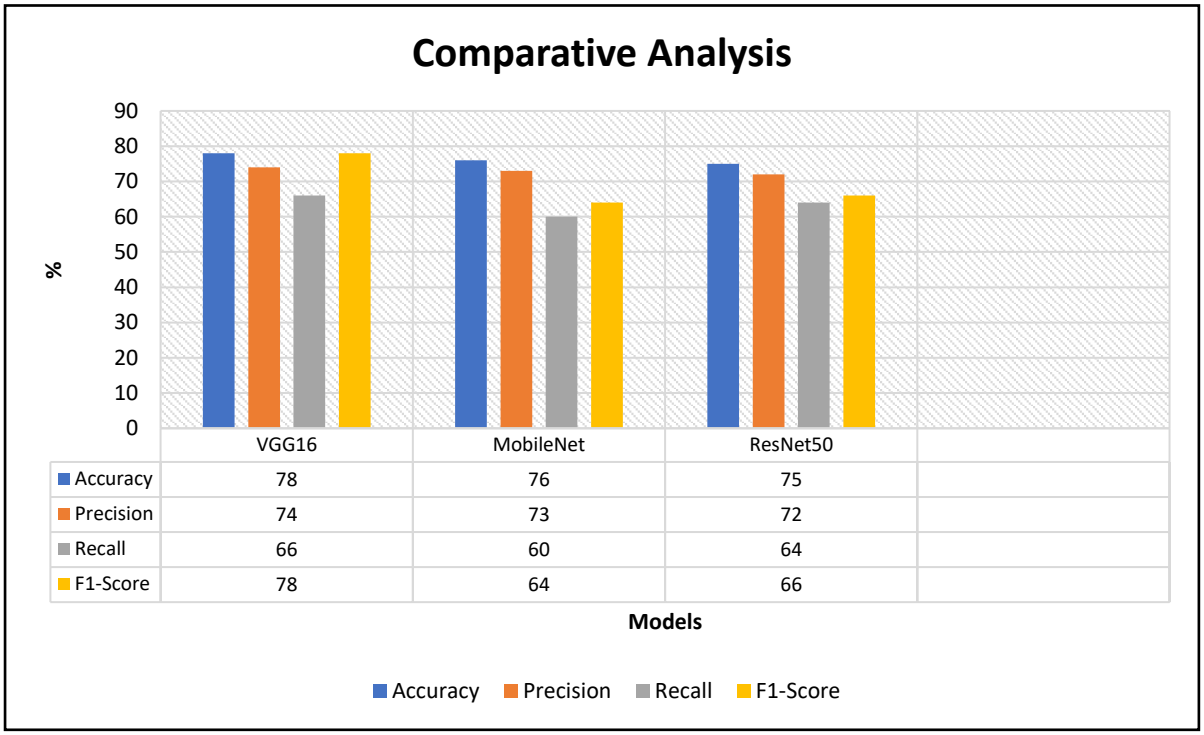

Figure 14. Comparative Performance Analysis of VGG16, MobileNet, and ResNet50 Models.

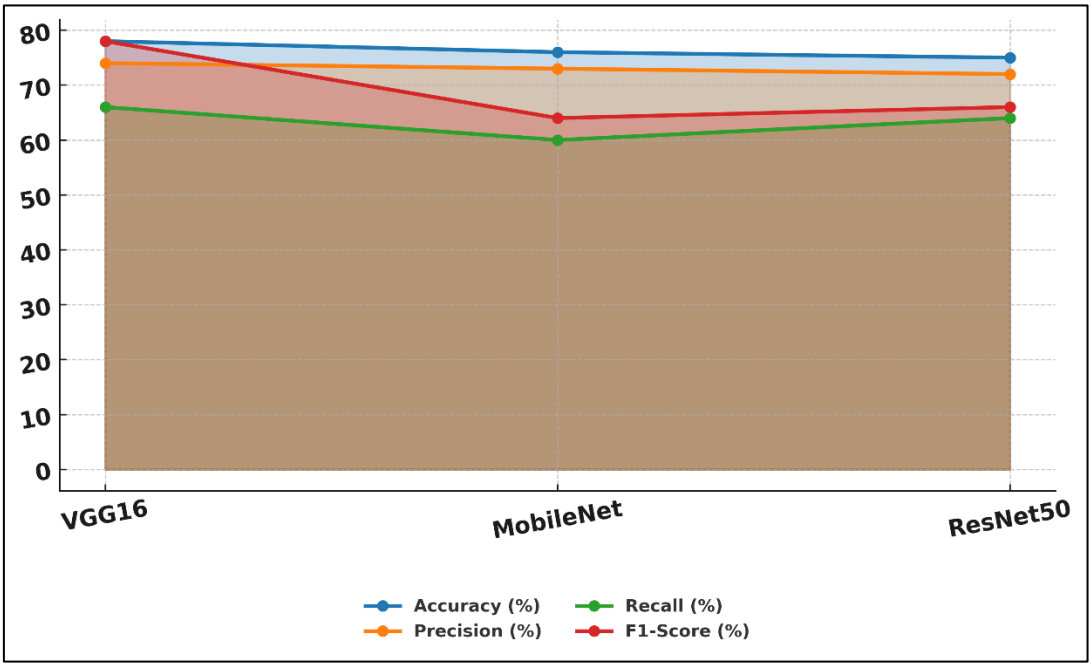

Figure 15. Performance Comparison of CNN Architectures across Evaluation Metrics.

Its recall value (60%) is however lower as compared to the recovery, which implies that the model would not capture a higher number of the actual positive cases. This renders MobileNet very effective yet a little less dependable in terms of extensive detection. ResNet50 which is known to have a deeper residual architecture is balanced and the ones that show 75 percent accuracy with a 64 percent recall.

Table 6. Comparison of Unimodal and Multimodal Models

Model Type	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Text-only	BERT	88.3	86	84	85
Image-only	ResNet50	75	72	64	66
Image-only	MobileNet	76	73	60	64
Image-only	VGG16	78	74	66	78
Multimodal	BERT + ResNet50	94.7	92	91	91

ResNet50 is a better model in finding true positive cases compared to MobileNet, with its higher ability to extract features which is associated with its higher precision (72%) and F1-score (66%) although slightly lower than VGG16.

Figure 14 compares VGG16, MobileNet, and ResNet50 performance based on the important measures. It emphasizes that VGG16 is more accurate, MobileNet more efficient and ResNet50 balanced at multimodal fake news classification.

Figure 15 compares the performance of CNN architectures in various evaluation measures and reveals how each model is deviated in terms of accuracy, precision, recall and F1-score in the effectiveness of multimodal fake news detection.

A comparative study was carried out on text-only, photo-only and multimodal models under the same experimental conditions to assess the effectiveness of multimodal learning. Table 6 indicates the superiority of multimodal model over monomodal models. Text-only BERT model is also powerful as it takes into consideration the semantic context, and image-only models are also moderate in their accuracy because they detect visual inconsistencies. VGG16 is the top CNN. Nevertheless, the multimodal model (BERT + ResNet50) is much more successful, with an accuracy of 94.7 and a better precision, recall, and F1-score. This enhancement supports the idea that the combination of textual and visual characteristics is more effective to detect cross-modal inconsistencies in fake news and obtain more accurate classification results.

The results of the experiment show clearly that a multimodal method of learning is effective in improving the performance of fake news detection. Although text-only models are useful to capture linguistic deception, image-only models are useful to detect visual manipulation, neither of the two modalities is sufficient to robustly detect. The multimodal framework takes advantage of the complementary strengths, and allows to identify the inconsistencies between textual claims and visual evidence. This results in enhanced generalization, fewer false positives, and more complex misinformation patterns.

Error analysis indicates that minority classes have most misclassifications because of a small training sample and the similarity in semantic patterns. Specifically, low-representation classes are characterized by lower recall, meaning that it is hard to learn rare patterns. There are also some errors that can be due to the minor discrepancies between text and image that cannot be detected without complex attention systems. The observations indicate that additional enhancement of model performance can be achieved by balancing the datasets and adding cross-modal attention. Future developments to further minimize overfitting involve incorporation of cross-modal attention mechanisms, further regularisation methods, including label smoothing, and k-fold cross-validation to estimate performance more robustly. Model generalization can further be improved by increasing the diversity of datasets, as well as using class-balancing methods like weighted loss functions or synthetic data generation.

The poorer results of minority classes (Class 3 and Class 4) can be mainly attributed to a high level of class imbalance and training samples. Deep learning models are biased towards the majority classes, and less recall of underrepresented categories. Also, some classes are semantically similar and examples are misleading and partially false content, leading to overlapping representation of features and higher misclassification. On a model level, ResNet50 can capture more complex features but is more likely to overfit whereas MobileNet is more likely to generalize since it is a lightweight model. VGG16 performs fairly well but is not sensitive to the variations produced by manipulations. These results show that the imbalance of the dataset, similarity of features, and the complexity of the model have a critical impact on performance in the classification.

7. Conclusion

This research paper provides a multimodal deep learning model, which combines textual embeddings of BERT with visual features of CNN to classify fake news. The suggested feature-level fusion model shows good results on the Fakeddit dataset, with a 94.7% accuracy and high macro F1-score in comparison to single-modelled models. The experiments validate the idea that

multimodal learning is effective in cross-modal inconsistencies capture, enhancing detection and reducing false positives. ResNet50 offers good feature representation, whereas MobileNet offers a high level of computational efficiency, and VGG16 offers a good balance in the baseline performance. Nevertheless, the research is constrained by the imbalance in classes and low performance on minority classes, which implies that sophisticated balancing strategies as well as attention-based fusion mechanisms should be considered. Future efforts will be on cross-modal attention, contrastive learning, and training on larger and more diverse data to improve generalization. These results confirm the usefulness of effective multimodal fusion as an effective alternative to the complicated transformer-based architectures. These restrictions point to the necessity of bigger and more varied datasets and better regularization methods. Moreover, the growing complexity of AI-generated fake news makes it possible to focus on the development of models with cross-modal attention, contrastive learning, and domain adaptation abilities.

References

[1] R. Kozik, A. Pawlicka, M. Pawlicki, M. Choraś, W. Mazurczyk, K. Cabaj, A Meta-Analysis of State-of-the-Art Automated Fake News Detection Methods. IEEE Transactions on Computational Social Systems, 11(4), (2024) 5219–5229. https://doi.org/10.1109/TCSS.2023.3296627

[2] F. Farhangian, R.M.O. Cruz, G.D.C. Cavalcanti. Fake News Detection: Taxonomy and Comparative Study. Information Fusion, 103, (2024) 102140. https://doi.org/10.1016/j.inffus.2023.102140

[3] J. Alghamdi, Y. Lin, S. Luo, Towards COVID-19 Fake News Detection Using Transformer-Based Models. Knowledge-Based Systems, 274, (2023) 110642. https://doi.org/10.1016/j.knosys.2023.110642

[4] M.A. Wani, M. Elaffendi, K.A. Shakil, I.M. Abuhaimed, A. Nayyar, A. Hussain, A.A.A. El-Latif, Toxic Fake News Detection and Classification for Combating COVID-19 Misinformation. IEEE Transactions on Computational Social Systems, 11, (2024) 5101–5118. https://doi.org/10.1109/TCSS.2023.3276764

[5] E. Raja, B. Soni, S.K. Borgohain. Fake News Detection in Dravidian Languages Using Transfer Learning with Adaptive Finetuning. Engineering Applications of Artificial Intelligence, 126, (2023) 106877. https://doi.org/10.1016/j.engappai.2023.106877

[6] Y. Liu, J. Zhu, K. Zhang, H. Tang, Y. Zhang, X. Liu, Q. Liu, E. Chen. (2024) Detect, Investigate, Judge and Determine: A Novel LLM-Based Framework for Few-Shot Fake News Detection. arXiv:2407.08952. https://doi.org/10.48550/arXiv.2407.08952

[7] C. Mallick, S. Mishra, M.R. Senapati. A Cooperative Deep Learning Model for Fake News Detection in Online Social Networks. Journal of Ambient Intelligence and Humanized Computing, 14, (2023) 4451–4460. https://doi.org/10.1007/s12652-023-04562-4

[8] K. Zhang, F. Zhou, L. Wu, N. Xie, Z. He. Semantic Understanding and Prompt Engineering for Large-Scale Traffic Data Imputation. Information Fusion, 102, (2024) 102038. https://doi.org/10.1016/j.inffus.2023.102038

[9] K.I. Roumeliotis, N.D. Tselikas, D.K. Nasiopoulos. LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study. Big Data Cognitive Computing, 8, (2024) 63. https://doi.org/10.3390/bdcc8060063

[10] N.U.R. Ahmed, A. Badshah, H. Adeel, A. Tajammul, A. Daud, T. Alsahfi, (2025) Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects, IEEE Access, 13, 1923–1961. https://doi.org/10.1109/ACCESS.2024.3523288

[11] U.A. Bhatti, H. Tang, G. Wu, S. Marjan, A. Hussain. Deep Learning with Graph Convolutional Networks: An Overview and Latest Applications in Computational Intelligence. International Journal Intelligent Systems 2023, (2023) 8342104. https://doi.org/10.1155/2023/8342104

[12] R. Kumar, B. Goddu, S. Saha, A. Jatowt, Silver Lining in the Fake News Cloud: Can Large Language Models Help Detect Misinformation?. IEEE Transactions on Artificial Intelligence, 6(1), (2025) 14 – 24. https://doi.org/10.1109/TAI.2024.3440248

[13] H. Ma, C. Zhang, H. Fu, P. Zhao, B. Wu, (2023) Adapting large language models for content moderation: Pitfalls in data engineering and supervised fine-tuning. arXiv preprint arXiv:2310.03400.

[14] F. Alshuwaier, A. Areshey, J. Poon. Applications and Enhancement of Document-Based Sentiment Analysis in Deep learning Methods: Systematic Literature Review. Intelligent Systems with Applications, 15, (2022) 200090. https://doi.org/10.1016/j.iswa.2022.200090

[15] B. Battal, B. Yıldırım, Ö.F. Dinçaslan, G. Cicek, Fake news detection with machine learning algorithms. Celal Bayar University Journal of Science, 20(3), (2024) 65-83. https://doi.org/10.18466/cbayarfbe.1472576

[16] M.V. Rampurkar, D.D. Thirupurasundari, An Approach towards Fake News Detection using Machine Learning Techniques. International Journal of Intelligent Systems and Applications in

Engineering, 12(3), (2024) 2868-2874.

[17] H. Murti, S. Sulastri, D.B. Santoso, D.A. Diartono, K. Nugroho. Design of Intelligent Model for Text-Based Fake News Detection Using K-Nearest Neighbor Method. Sinkron : Jurnal Dan Penelitian Teknik Informatika, 9(1), (2025) 1-7. https://doi.org/10.33395/sinkron.v9i1.14306

[18] Y. Jingyuan, X. Zeqiu, H. Tianyi, Y. Peiyang, (2025) Challenges and Innovations in LLM-Powered Fake News Detection: A Synthesis of Approaches and Future Directions. GAIIS '25: Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security, 87 – 93. https://doi.org/10.1145/3728725.3728739

[19] M. Tan, H. Bakır, Fake news detection using BERT and Bi-LSTM with grid search hyperparameter optimization. Bilişim Teknolojileri Dergisi, 18(1), (2025) 11-28. https://doi.org/10.17671/gazibtd.1521520

[20] E. Shushkevich, M. Alexandrov, J. Cardiff. Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data. Inventions, 8, (2023) 112. https://doi.org/10.3390/inventions8050112

[21] E. Alsuwat, H. Alsuwat. An improved multi-modal framework for fake news detection using NLP and Bi-LSTM. The Journal of Supercomputing, 81, (2025) 177. https://doi.org/10.1007/s11227-024-06671-z

[22] D. Mouratidis, A. Kanavos, K. Kermanidis. From Misinformation to Insight: Machine Learning Strategies for Fake News Detection. Information, 16(3), (2025) 189. https://doi.org/10.3390/info16030189

[23] Fakeddit Dataset: https://www.kaggle.com/datasets/vanshikavmittal/fakeddit-dataset

Authors Contribution Statement

Both the Authors equally contributed, Read and Approved the Final Version of the Manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License