AI-Based Fake News Detection on Social Media Posts Combining Multimodal and Unimodal Methods for Binary Classification

Siddhi Lakkarsu[1], **Abdulla Kerimov**[2]

Received February 18, 2026
Accepted May 12, 2026
Electronic access August 15, 2026

Our study compared unimodal and multimodal transfer-learning models using the Fakeddit dataset. We analyzed a wide range of models to assess the impact of text and images for fake news detection. For image-only classification, we fine-tuned Convolutional Neural Network (CNN) and Vision Transformer (ViT) models. In addition, we fine-tuned a group of transformer-based natural language processing (NLP) models for text. Then, we combined them with a late-fusion multimodal framework. We compared model performance using machine learning (ML) metrics including accuracy, precision, recall, and F1 score. For unimodal computer vision results, CNNs achieve a 0.45-0.60 range of F1 score with ConvNeXt-XL performing best. In comparison, the F1 score ranges from 0.56-0.82 for ViTs, with EfficientViT scoring highest. Unimodal NLP models dominate performance with a range of 0.85-0.90 for F1 score. This is led by DeBERTa. Our study shows that text related clues are the major drivers for predictions within the Fakeddit dataset. Image based features provide complementary context. In comparison, multimodal models deliver a high F1 score range of 0.87-0.91, but not significantly different from text-only models. While text-based detection can be simple and highly effective, multimodal models will have a role with cross-modal inconsistencies, for example when a fake image is paired with a real caption. Our study plays an important role in developing fake news detection systems that are scalable and efficient in this era of rapidly spreading misinformation.

Keywords: *fake news detection, unimodal, multimodal, fusion, transfer learning, natural language processing, large language models, computer vision, transfer learning, fine tuning*

Introduction

Today's online world is full of information that is spreading faster and wider than ever before. This happens because of social media, news, and messaging. The fast spread of information has many benefits, but it also helps to spread misinformation and fake news[1]. Misinformation can reach millions of people in a few seconds. It moves faster than real news and people remember the fake information even after learning the truth[2]. Fake news changes public opinion and splits people apart. It lowers trust in journalism, science, and government[3]. Misinformation is also a large cause of real-world harms. It puts people's health in danger and makes them stop trusting the government[4]. These problems show that fake news is a serious problem all over the world.

Several methods have been used against fake news, like fact-checking and online safety lessons[5]. However, misinformation is still growing because of AI and deepfakes. Human-based methods cannot keep up with the large amount and speed of content online[6]. Because of this, automated AI and ML-based detection systems are strongly needed. They can work with large-scale data and find patterns of false content.

Relevant Works

Traditional fake news detection approaches rely heavily on human effort. This makes them time-consuming and inconsistent. They cannot keep up with the spread of misinformation online[3]. In contrast, AI/ ML techniques have grown in importance in recent years[7]. These methods include natural language processing (NLP) and computer vision (CV). They analyze large volumes of text, images, and social media interactions. These methods identify patterns associated with misleading content. Furthermore, AI and ML systems detect features which are difficult for humans to observe. These include hidden textual cues and unusual writing styles. These systems can analyze patterns in how information spreads. Scalable and automated systems are necessary for timely, real-world fake news detection[8].

Early efforts with AI/ML systems focused on text-based methods. This primarily included using textual features and ML classifiers. Methods focused on textual patterns such as bag-of-words and TD-IDF representations[9]. More recent approaches have adopted deep learning methods. These mod-

[1] *Monta Vista High School, California, USA*
[2] *Stanford University PhD Alumni, Houston, TX, USA*

els have a stronger ability to understand the hidden context in news stories. Methods focused on recurrent neural networks (RNNs), long short-term memory (LSTM), and transformer-based architectures[8]. Text-only transformer models especially have performed strongly on benchmark datasets. Large pre-trained models such as DeBERTa have reached up to 95.6% accuracy on the LIAR dataset[10]. Datasets like LIAR contain short political statements labeled for truthfulness. These datasets provided early foundations. However, they were limited by their small size and their topic of data[10]. Despite overall advances, text-only systems struggle with many cues. These include well-hidden changes, satire, or misinformation that looks similar to true reporting[1].

Parallel work has explored image-only detection. This is especially important due to the growing scale and spread of manipulated content and deepfakes. Generative adversarial networks (GANs) have advanced greatly. This has significantly increased the authenticity of fake images and videos[1]. These advances greatly challenge traditional detection methods. Later studies have applied CNNs and ViTs to identify image features. These include pixel-level errors, compression artifacts, and glitches in facial features[11]. For example, an image-based model named DeepFakeGuard achieved 97% accuracy and an F1 score of 0.97 on the DFDC dataset[12]. This dataset is considered a large-scale benchmark for deepfake detection. However, image-only methods remain limited by their reliance on visual features. These artifacts can be hidden by targeted strategies. Hence, the performances of image-only methods across many content formats and platforms are lowered.

The limitations of unimodal fake news detection have not gone unnoticed. Recent studies have increasingly adopted multimodal approaches. These methods combine textual and visual information. Early tests of such models took place on Twitter-based datasets with models consistently performing strongly. Their precision, recall, and F1 scores ranged between 0.74 and 0.88. Several multimodal tests showed that adding text and images improved classification compared to text-only models[13–15]. However, Twitter-based datasets are often limited to binary labels and weaker visual-textual agreement. This restricts highly detailed multimodal analysis.

To address these limitations, a large-scale multimodal dataset was introduced. The dataset, titled Fakeddit was designed specifically for fake news detection[16]. Unlike Twitter datasets, Fakeddit provides several sets of labels. They range from coarse (two classes) to fine-grained (six classes). Fakeddit has purposefully mismatched and misleading image-text pairs. This makes it more useful for analyzing cross-modal variances. Thus, Fakeddit has become a widely adopted benchmark for evaluating multimodal fusion methods[7]. These are techniques that AI uses to combine different formats of data into one understanding. However, in this study we restrict all experiments to binary (real vs. fake) classification. Also, we do not assess model performance on fine-grained labels.

As shown in Table 1, prior work on Fakeddit reports consistently strong performance for multimodal models. Recent approaches have achieved precision, recall, and F1 scores above 0.90. Studies by Uppada et al., Kumari et al., and Saha et al., demonstrate clear improvements over earlier methods[12,17,18]. Recent models using advanced fusion methods show large increases in accuracy and F1 score[19,20]. In comparison, earlier models such as by Low et al. reported lower scores. Overall, multimodal models trained on Fakeddit show the dataset's effectiveness in finding cross-modal patterns.

AI/ML systems are still having difficulties, like with new situations and large growth. These days compared to true information, the spread of misinformation has been much more rapid in pace and reach. Hence, the solutions to identify them should also be able to match their pace and volume[2]. However, multimodal models use too much compute power which prevents them from being greatly used[25]. And the information used to train these models can be biased. Most datasets only have small topics or have unequal classes of information. This can lead to models doing well in benchmarks but failing in real-world conditions[1].

Rapidly growing AI content has been identified by researchers to accelerate misinformation spread. Deepfake tools have made it harder to tell the difference between real and fake media. This means that we need tools that can keep up with generative AI[6,11]. Most of the current techniques evaluating either text or images have performed well on datasets using Twitter and Fakeddit. Despite that, we need to improve the accuracy and consistency of these systems, particularly for content that could be either hidden or vary across different modalities. The solutions should perform well with different data modalities and media platforms and keep up with the pace of fake news.

In order to better detect misinformation, new ML models are being developed to match state-of-the-art (SOTA) performance. SOTA ML models are the newest or best-performing models on tasks and benchmarks. The evaluation of these new models is a priority to ensure that they are not only efficient compared to older models but also that they perform well beyond the limits of their training data. Thus, they should be compared with baselines to ensure that results are consistent and valid across datasets which is essential for our ability to detect rapidly spreading misinformation.

Research Gap and Contribution

Large progress has been made in fake news detection but there are still a few important gaps in past research as first, models that can compete with SOTA (state-of-the-art) methods are still being made but there is very little testing comparing uni-

Table 1 Comparison with other models on the Fakeddit multimodal dataset.

Author	Dataset	Precision	Recall	F1	Accuracy
Uppada et al. 2023 [21]	Fakeddit	0.934	0.931	0.930	0.919
Emil et al. 2025 [22]	Fakeddit	0.935	0.903	0.919	0.888
Kumari et al. 2024 [17]	Fakeddit	0.930	0.950	0.930	0.930
Shao et al. 2022 [23]	Fakeddit	0.838	0.749	0.791	0.804
Low et al. 2025 [20]	Fakeddit	0.698	0.610	0.651	0.612
Feng et al. 2024 [24]	Fakeddit	0.798	0.637	0.708	0.745
Saha et al. 2023 [18]	Fakeddit	0.910	0.950	0.930	0.912

modal and multimodal methods on large datasets like Fakeddit. Second, Fakeddit has become a useful benchmark for text and image fake news detection but we do not clearly know how much the text alone versus the images alone affect how well the model makes predictions and third, past studies usually test models by themselves but this makes it hard to see how using both text and images helps find fake news in the real world.

This study fixes these gaps with three steps as first, (1) we compare unimodal (text-only and image-only) and multimodal models on Fakeddit and then second, (2) we use numbers to compare the exact testing scores from both strong image and text models and then third, (3) we look at how different models and fusion methods affect how well models can make predictions. By doing these tasks, this work shows how well using both text and images works and also, this study gives us a better idea for building strong, large-scale fake news detection systems.

Dataset

For this study we used the Fakeddit dataset [16]. This is a large-scale multimodal benchmark. It contains over one million samples with images, text, and metadata. Fakeddit has several orders of labeling which allows analysis at different levels of specificity. The 2-way setting separates real from fake news. Next, the 3-way setting further separates true, satire/parody, and false/misleading content. The 6-way labels are true, satire/parody, false connection, misleading content, manipulated content, and imposter content. This structure supports both binary and fine-grained misinformation analysis. However, all experiments in this study use the binary 2-way classification task. Figure 1 shows example labels and corresponding posts from the dataset. Extending this work to multi-class classification remains an important direction for future research.

Fake news datasets have generally been limited in several factors. These include size, number of classes, modality, source, and data category. Many are relatively small. This makes them less effective for ML models that require large-scale training data [16]. Several datasets also contain only a few classes or are restricted to specific domains. These topics can include politics, conspiracy theories, or celebrity news. Table 2 compares statistics of fake news datasets for all of which Fakeddit leads. It includes size, number of classes, modality, source, and data category. Another important factor is modality. Fake news appears in multiple forms which include both text and images. However, many datasets are unimodal and contain only text or images. Even multimodal datasets are limited by small size and weaker structure. This makes them less effective compared to Fakeddit.

Our dataset contained over 1 million images as shown in Table 2. 682,996 of these were multimodal samples. These were divided into training, validation, and testing sets. The division followed standard Fakeddit splits. Figure 2 and Figure 3 illustrate examples of fake and real image-text pairs. We faced computational and hardware constraints which limited all experiments. They were conducted on a reduced but balanced subset of the training data. Specifically, we used 99,909 multimodal training samples (49,908 fake and 50,001 real). This resulted in an about balanced (1:1) class ratio. This subset represents only a portion of the full training set. However, it remains large enough for stable fine-tuning of deep neural networks while within computational limits.

For evaluation, we used the original Fakeddit validation and test splits. This was to keep the natural class spreads. Both sets had a fake-to-real ratio of about 3:2. The validation set contained 57,557 samples with 34,396 fake and 23,161 real instances. The test set contained 57,495 samples with 34,162 fake and 23,333 real instances. There is an imbalance between the balanced training set and the skewed evaluation sets. This may affect model behavior, particularly F1 scores.

No other changes were made to lower bias in the evaluation splits. Models were trained without class-specific tuning. They predicted the class with highest confidence. This may lead to a bias toward the majority class. Also, it can lower performance on the minority class. This was usually the case for the "real" class. Additionally, models were trained on a reduced subset of the dataset. Thus, comparisons with work using the full Fakeddit training set should only be seen as ap-

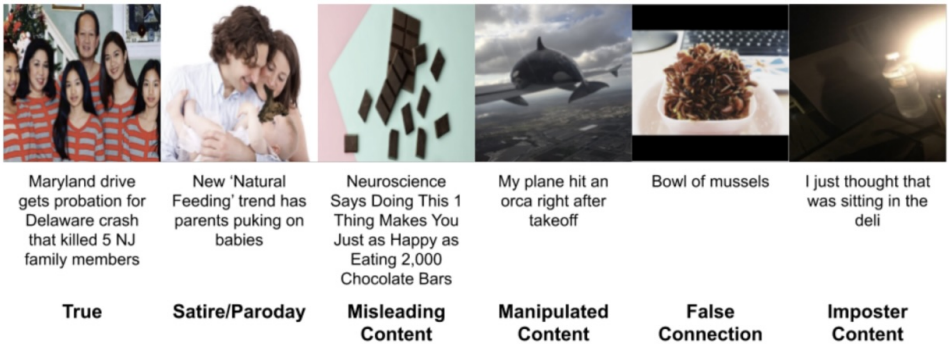

Figure. 1 Fakeddit samples with 6-way classification labels and corresponding text data [16]

Table 2 Comparison of various fake news datasets (IA = Individual Assessments) [16]

Dataset	Size (# of samples)	# of Classes	Modality	Source	Data Category
LIAR	12,836	6	text	Politifact	political
FEVER	185,445	3	text	Wikipedia	variety
BUZZFEEDNEWS	2,282	4	text	Facebook	political
BUZZFACE	2,263	4	text	Facebook	political
some-like-it-hoax	15,500	2	text	Facebook	scientific/ conspiracy
PHEME	330	2	text	Twitter	variety
CREDBANK	60,000,000	5	text	Twitter	variety
Breaking!	700	2,3	text	BS Detector	political
NELA-GT-2018	713,000	8 IA	text	194 news outlets	variety
FAKENEWSNET	602,659	2	text	Twitter	political/celebrity
FakeNewsCorpus	9,400,000	10	text	Opensources.co	variety
FA-KES	804	2	text	15 news outlets	Syrian war
Image Manipulation	48	2	image	self-taken	variety
Fauxtography	1,233	2	text, image	Snopes, Reuters	variety
image-verification-corpus	17,806	2	text, image	Twitter	variety
The PS-Battles Dataset	102,028	2	image	Reddit	Manipulated content
Fakeddit (ours)	**1,063,106**	**2,3,6**	**text, image**	**Reddit**	**variety**

Figure. 2 Example of a fake image with corresponding text information [16].

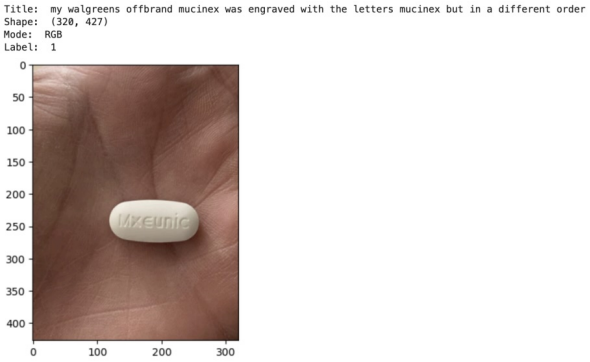

Figure. 3 Example of a real image with corresponding text information [16].

proximate. Differences in training scale may affect performance. Therefore, our results provide a general comparison rather than a direct evaluation against full-data SOTA baselines.

The Fakeddit images are shown in Figure 2 and Figure 3. They have a lot of variability in both shape and mode. As shown in Figure 4, most images have dimensions around 320×320 pixels. Some extend up to 5000 pixels in width and height. Furthermore, most images are in RGB (red, green, blue color format) mode. Smaller amounts are in L, P, or RGBA (RGB with an alpha/transparency channel) modes. This variation causes challenges for deep learning models, which take in the same input size for training. To address this, all images were resized to a 224×224 RGB format. This kept the data uniform, saved compute power, and allowed the system to pull out accurate features [26]. ImageNet is a large-scale image dataset commonly used for pretraining. Pixel values

were normalized using the standard ImageNet mean and standard deviation[27]. No other data augmentation was applied. This ensured that evaluation remained consistent across all visual models.

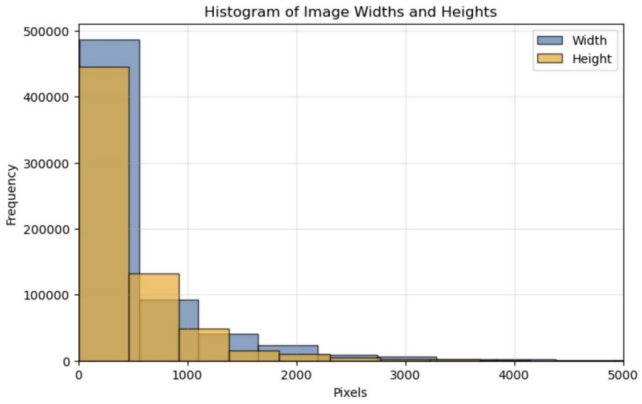

Figure. 4 Count distribution of image width and height in pixels for the Fakeddit dataset.

The Fakeddit dataset has strong textual, visual, and metadata attributes. This makes it very useful for multimodal fake news detection. Textual data consists of the original title and a preprocessed clean title[28]. The original title was created by the user for each post. The clean title is a normalized version of the original title as it eliminates punctuation, irregular spacing, and noise to have consistent model inputs[29]. The analysis of word counts of clean titles was performed using whitespace tokenization which showed that most headlines were short. They usually range from five to fifteen words with fewer long-form titles[30]. This reflects the short, attention-driven nature of social media[16,31]. To prepare the data for NLP testing, we tokenized all the clean titles with pretrained tokenizers without cutting or padding text.

Methodology

Unimodal vs. Multimodal Approaches

In our study, we evaluated both unimodal and multimodal methods for fake news detection. The initial unimodal methods used either only text or image-based models. Later, both image and text were combined for multimodal techniques. This enabled us to assess the performance, strengths and weaknesses of all these models across different media platforms. We performed stepwise analysis to compare these methods to older models in detecting fake news. The additional impact of context on model performance was also assessed. Overall, both text and image data is valuable to detect misinformation online.

Unimodal approaches to fake news detection use a single data modality (text or images). For the detection based on images computer vision architectures are used. This study used both CNNs and ViTs which can effectively identify signs of manipulation, compression artifacts, or inconsistencies to spot falsified or manipulated images[11]. Despite their ability, they are suboptimal as fake news can often combine misleading text and visuals. As a result, image-only models may fail when authentic-looking images are paired with misleading captions or narratives.

In this study, all NLP experiments use only clean title field from Fakeddit as input. The clean title is a normalized version of the original headline to remove any noise related punctuation and formatting inconsistencies. Consistency in data was prioritized by excluding all the original titles, comments, and metadata from the analysis.

The text-only unimodal systems used NLP techniques and transformer architectures such as BERT (Bidirectional Encoder Representations from Transformers)-based models. While these models have the ability to detect hidden textual clues, context-based sentiment and inconsistencies in semantics, they still can be limited in situations where visual and text data contradict each other[32,33]. This could include scenarios where a true caption and false image are paired together. Given that unimodal techniques have limitations to detect multimodal fake news, further efforts are directed to use both text and images together[34].

These multimodal approaches integrate various data modalities (typically text and images) and utilize their complementary information which a single modality fails to provide. A model combining textual encoders and image feature extractors strengthened its ability to analyze both the content of headline and its corresponding image[7,16]. These combined models are thus enabled to detect inconsistencies across multiple modalities with improve performance for binary real/fake classification tasks.

In order to address dataset imbalance issues, we trained all the models in our study using standard class weighted cross-entropy loss and relied on standard argmax-based decision thresholds. Early stopping was also applied based on validation performance. This was done to prevent overfitting. Dropout was the primary form of regularization. This approach ensured that the models were compared fairly while minimizing overfitting and improving class imbalance.

Testing happened using uniform training setups to ensure reproducibility across models. The Adam optimizer was used for all models. The learning rate for NLP models was 1e-5 and it was 1e-4 for computer vision models. This was without learning rate scheduling. NLP models were trained for up to 3 epochs with a batch size of 32. Computer vision and multimodal models were trained for up to 5 epochs with a batch size of 16. Gradient accumulation was used when memory

constraints required smaller batch sizes. Early stopping was used to know the final epoch based on validation performance. This setup helped to have efficient training while maintaining consistency across models.

Computer Vision Algorithms

In fake news detection, computer vision (CV) algorithms are widely used to process and analyze visual content. Both CNNs and transformer-based architectures have performed strongly in extracting image features relevant to misinformation. In unimodal approaches, these models operate solely on images. This is done to identify cues of manipulation, falsification, or visual inconsistency. In multimodal approaches, CV models first encode the images into meaningful feature representations. Then, these are fused with textual features to jointly inform the detection of fake content.

Pretrained transfer learning models were the backbone of our pipeline. They were combined with CNNs and transformers. Initially, we used models already trained on ImageNet, a large image dataset[27]. Later, we fine-tuned them on Fakeddit dataset to recognize patterns of misinformation. For example, CNNs like ResNet, EfficientNet, and ConvNeXt can transfer strong object and scene-level feature extraction directly to this task. Transformer-based models like ViT, DeiT, and Swin Transformers have their own strengths and work uniquely by understanding how different parts of the image relate to each other globally[35–38]. Beyond general ImageNet-initialized architectures, we also incorporated domain-relevant pretrained models designed for identifying fake media. For example, EfficientNet-based models were fine-tuned on the DFDC dataset[12]. By combining the general-purpose visual learning with task-specific pretraining strengthened our models performance further. Additionally, we used pretrained models designed specifically for deepfake detection. For example, we evaluate EfficientNet models fine-tuned on the DFDC dataset. Combining broad visual representation with targeted domain pretraining improved our models in several ways including feature extraction, that had positive impact on both our unimodal and multimodal benchmarks.

CNN-Based Transfer Learning

This study employs a range of CNN architectures which have been pretrained on the ImageNet dataset[27]. Then, they have been fine-tuned on Fakeddit to investigate their efficacy in fake news detection[16]. Models were selected based on architectural diversity and proven performance in image classification. Table 3 summarizes the evaluated CNN and ViT models. It includes key attributes such as parameter count, model size, and depth.

Transfer learning is motivated by evidence from CNNs trained on large-scale datasets. These models learn transferable visual features that can be adapted to new tasks with limited labeled data[35,39]. This approach improves several factors compared to training from scratch. These factors include convergence, generalization, and computational efficiency[36]. Evaluating multiple architectures also enables a more methodological comparison of design choices. This supports a robust assessment of visual representations for misinformation detection.

Vision Transformer-Based Transfer Learning

Transfer learning has been adopted in computer vision. This is driven by extensive empirical evidence from large-scale pre-trained vision transformers. These models capture generalizable spatial and semantic patterns. These can be effectively adapted to specialized downstream tasks including image classification and multimodal fake news detection[38,46]. ViT models were pre-trained on large-scale image datasets such as ImageNet-21k or JFT-300M. Following that, fine-tuning was performed on Fakeddit to improve its generalization and quicker convergence. These features minimized the dependency on data when compared to models trained from scratch. Some of these selected models were DeiT, EfficientViT, ViT, and Swin Transformer, representing a diverse set of transformer innovations[38,45–47]. Among these, ViT is a foundational transformer architecture, DeiT is a data-efficient approach, EfficientViT is a computationally optimized variant, and Swin Transformer has a very ordered, locally aware design. All of these unimodal assessments formed the benchmark for the next steps with multimodal integration. In particular, we separated the impact of visual features on overall model performance.

NLP Algorithms

This study evaluates a diverse set of transformer-based NLP models for unimodal fake news detection. This happened through fine-tuning pre-trained architectures on Fakeddit. These NLP models were pre-trained on large corpora such as Wikipedia and BookCorpus. This allowed them to leverage transfer learning to improve text understanding and classification performance[29]. This setting establishes a baseline by isolating the contribution of linguistic features prior to multimodal integration.

Transfer learning is central to this study's NLP pipeline as models are not trained from scratch. Instead, the pre-trained transformer-based models are fine-tuned for fake news detection. This enables the transfer of rich semantic and syntactic knowledge[29,43]. Prior work shows that such models generalize well, converge faster, and require less data than training from scratch[48]. The models selected were ALBERT,

Table 3 Pre-trained computer vision CNN-based and transformer-based transfer learning models used in our study (Depth of "-" is derived from `https://keras.io/api/applications/`)

Reference	Model	Size (MB)	Parameters	Depth
Chollet 2017[40]	Xception	88	22.9M	81
Simonyan and Zisserman	VGG16	528	138.4M	16
He et al. 2015[35]	ResNet50V2	98	25.6M	103
He et al. 2015[35]	ResNet101V2	171	44.7M	205
He et al. 2015[35]	ResNet152V2	232	60.4M	307
Chollet 2017[41]	InceptionV3	92	23.9M	189
Chollet 2017[42]	InceptionResNetV2	215	55.9M	449
Sandler et al. 2018[43]	MobileNetV2	14	3.5M	105
Chollet 2021[44]	DenseNet201	80	20.2M	402
Tan and Le 2021	EfficientNetB7	256	66.7M	438
Tan and Le 2021	EfficientNetV2B3	59	14.5M	–
Tan and Le 2021	EfficientNetV2L	479	119.0M	–
Liu et al. 2022[37]	ConvNextSmall	192.29	50.2M	–
Liu et al. 2022[37]	ConvNeXtXLarge	1310	350.1M	–
Liu et al. 2023[45]	EfficientViT	120	24.0M	16
Touvron et al. 2021[46]	DeiT	330	86.0M	12
Liu et al. 2021[47]	Swin Transformer	340	88.0M	12
Dosovitsky et al. 2021[38]	ViT	330	86.0M	12

BERTweet, DeBERTa, USE, ELECTRA, FBERT, Longformer, DistilBERT, and RoBERTa. This selection represents a broad and methodologically diverse group. These models are transformer innovations which have defined modern NLP research[49]. RoBERTa, DeBERTa, and ELECTRA are general-purpose architectures. They provide strong contextual representations for identifying misleading language. Efficient models like ALBERT and DistilBERT support lightweight experimentation. In addition, BERTweet is a domain-specific model[35,50–52]. It captures the informal linguistic patterns common in social media misinformation[53].

Fusion Approach

We combined text and image information to build a multimodal fusion model. Textual content is first processed through a pretrained NLP model discussed above. This converts each news title into a dense feature vector. Simultaneously, the corresponding image is processed through a pretrained CV model. This can be CNN-based or transformer-based as discussed above. These unimodal feature vectors (textual and visual) are then concatenated. This occurs along the feature dimension to form a joint representation of the post. The fused representation is passed through a fully connected neural network classifier. There is at least one hidden layer and nonlinear activations which allows the model to learn from the cross-modal correlations between the text and the image. Given that the model was trained end-to-end on multimodal data, it was able to use complementary data and dynamically weight the impact of different features from these modalities to improve accuracy for identifying fake news compared to unimodal models.

Evaluation Metrics

For evaluation of our fake news models, we used standard performance metrics: precision recall and the F1 score. Precision measures the accuracy of the model to flag fake content while minimizing false alarms. Recall metric measures overall ability to spot actual fake instances. The F1 score is a harmonic mean of the two and thus helps balance these metrics, which is even more important when handling class imbalances. Similar to prior benchmark studies on datasets like Fakeddit and FakeNewsNet, we also report both the macro and class-wise F1 values which allows us to directly compare different models[16,25].

Results and Discussion

Unimodal Approach

We first show the performances of fine-tuned unimodal models. This starts with image-based CV architectures and then text-based NLP models. This happens to understand the individual impacts of visual and textual information. We then compared these results with prior literature. Figure 5 summarizes the test and validation F1 scores of unimodal CV models. Table A1 can be seen for the full metrics. Overall,

image-only models have a high margin of error with test F1 scores between 0.45 and 0.60. This shows that visual content has limited signals for fake news detection. Many misleading posts rely on context or hidden cues not captured by images alone. Performance also varies significantly across models. This highlights the importance of model design.

We evaluated a group of CNN-based transfer learning models. Of the group, ConvNeXt-XL has the strongest and most consistent performance. It has a test F1 score of 0.78 for fake images and 0.69 for real images. Other models are from the ResNet, Inception-based, and EfficientNet families. These models have much lower and less stable F1 scores. They often have near-chance performance for one class. This is usually for real images. Lightweight models like MobileNetV2 and DenseNet201 are weaker at prediction. This further proves the difficulty of detecting misinformation only from images. Among ViT models, EfficientViT has the strongest and most consistent performance. It has test F1 scores of 0.82 for fake images and 0.75 for real images. ViT and DeiT perform slightly lower for fake content (0.79-0.80 test F1 score). However, they score much lower for real images (0.62-0.63 test F1 score). This shows difficulty with true content. Swin Transformer performs the worst with test F1 scores of 0.61 for fake images and 0.56 for real images. Validation scores have a similar pattern. This shows consistent performance on real data.

Overall, ViTs slightly outperform several CNNs though both groups have inconsistent performance. EfficientViT has the highest test performance of all the unimodal CV models in this study. It has test F1 scores of 0.82 on fake images and 0.75 on real images. These findings agree with prior fake news detection research. Both show that image-only methods have limited prediction abilities. As shown in Figure 5, the results from this study are consistent with prior studies. Both report F1 scores of 0.50-0.75 for image-only models. While newer models like ConvNeXt improve over older models such as VGG16 and ResNet, the overall performance remains limited demonstrating the challenges of only using visual features for fake news detection.

Figure 6 shows that text-only models consistently outperform image-only ones with unimodal NLP models having test F1 scores ranging from 0.84 to 0.90. We noted that even lightweight models like DistilBERT and ALBERT perform well demonstrating that textual features alone have strong signals within Fakeddit. This data is shown in table A2 with evaluation metrics for training, validation, and testing. Among all the models, DeBERTa achieves the best performance with test F1 scores of 0.90 for fake content and 0.86 for real content. While other transformer models including ELECTRA, RoBERTa, and BERTweet perform close to DeBERTa (0.88-0.89 test F1 score), they fall slightly short. Similar pattern was noted with validation scores which demonstrated good performance on real data. Our results are in alignment with prior

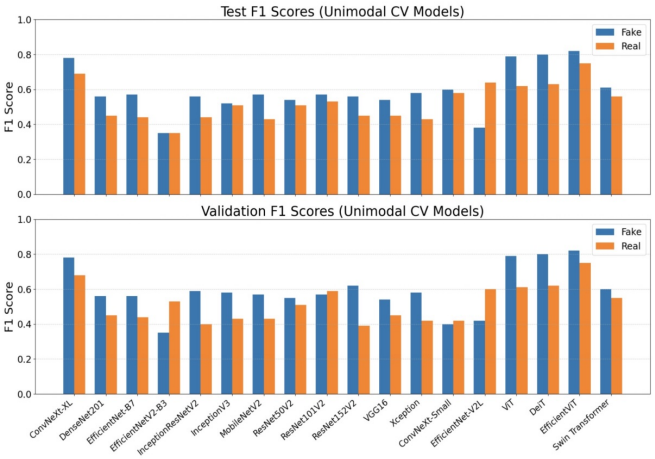

Figure. 5 (Top) Test F1 scores and (bottom) validation F1 scores for classifying real and fake images using fine-tuned unimodal CV models.

studies as they both have similar F1 scores of 0.85-0.92 when using transformer-based NLP models on large datasets. While these strong performances with NLP models emphasize the power of textual cues on Fakeddit, their performance may fall when these textual signals are more hidden showing the need to build multimodal systems.

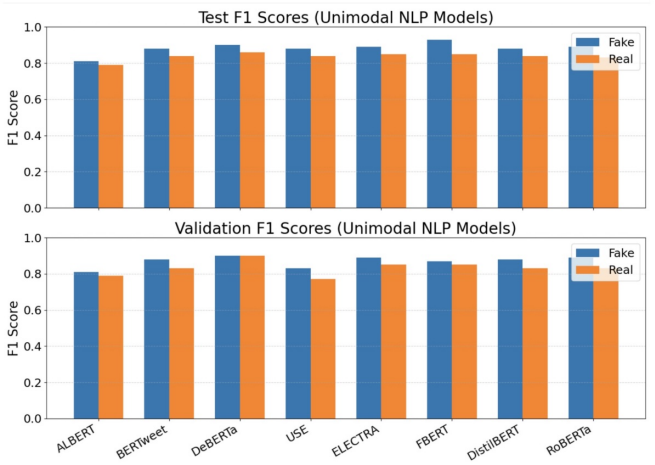

Figure. 6 (Top) Test F1 scores and (bottom) validation F1 scores for classifying real and fake text using a fine-tuned unimodal NLP approach.

When we compared unimodal NLP with CV models, a consistent and clear performance deficit was noted with text-based models outperforming image-only approaches. Figure 6 shows NLP models achieving test F1 scores of 0.85-0.90

while Figure 5 shows CV models achieving test F1 scores of 0.45 to 0.82. However, most CNNs fall between test F1 scores of 0.45 and 0.60. ConvNeXt-XL reaches test F1 scores of 0.78 for fake images and 0.69 for real images. ViTs perform better overall. EfficientViT achieves the strongest results with test F1 scores of 0.82 for fake images and 0.75 for real images. This consistently low performance of image-only models in comparison to text-based approaches demonstrates the stronger signal of NLP in Fakeddit.

An important finding to note was that the model prediction patterns further showed these differences. While the large ViTs achieve near-perfect training F1 scores ($\sim$1.0), they have large drops on test data (e.g. $\sim$0.62 for real images) because of overfitting. Smaller CNNs have lower training performance but still have weaker prediction. This is because of capacity limitations. Two large NLP models are DeBERTa and DistilBERT. These two achieve high training F1 scores (0.97-0.98). Also, they both have lower test performance (0.84-0.86 for real content). This is because of memorization. In contrast, there is BERTweet, which is a lighter model. This model has more consistent prediction at slightly lower performance. Overall, there are the factors of model capacity and limited regularization. These items contribute to overfitting in larger models. And they build performance ceilings in smaller ones.

This imbalance is because of the nature of misinformation. Fake news is built on language manipulation and emotional triggers. These cues are effectively captured by pretrained language models. In contrast, images are usually true. However, they are used out of context without clear visual features. Because of this, image-only models must rely on indirect cues. This makes detection much more difficult without textual context.

Multimodal Approach

We then performed multimodal analysis and present the scores from fine-tuned multimodal models which use both text and images for binary (real/fake) classification. Our results demonstrate impact of both language and image encoders on how well the model can make predictions. We also assess if using both modalities (image and text) have a clear advantage over unimodal methods. To do so, we list the model pairings and training setup and then look at multimodal results and compare them with past studies to better understand impact on learning from two different content formats for fake news detection.

Figure 7 shows the results from fine-tuned multimodal models on Fakeddit for binary classification and the x-axis shows different combinations of NLP and CV models, and also Table A3 contains the exact scores for training, validation, and testing.

We used a methodical approach to pair models including both capacity and efficiency. Given their strong results on Fakeddit, high performing NLP models like DeBERTa, RoBERTa, and DistilBERT were used a baseline. They were each then paired with visual encoders from different model families like MobileNetV2 (lightweight CNN), DenseNet201 and InceptionResNetV2 (deeper CNNs), EfficientViT and Swin Transformer (transformer-based models). By doing so, we ensured that both high-capacity and efficient models were evaluated in our study for controlled comparisons. Pairing DeBERTa + InceptionResNetV2 and DeBERTa + EfficientViT allowed for direct comparison between CNN and transformer-based visual models as visual models were under the same textual conditions. Across all combinations, test F1 scores range from 0.87 to 0.91 with DeBERTa + InceptionResNetV2 having the strongest performance. Validation scores have a similar pattern and demonstrate consistency on real data. Overall, combining strong text and image models is effective. However, gains depend on the choice of visual encoder.

Our results are consistent with multimodal works on Fakeddit. Previous studies report multimodal F1 scores in the 0.88-0.92 range. This shows that the models evaluated here perform similar to existing benchmarks[17,18,21]. However, the multimodal results are not much higher than those of text-only models. Image encoders thus had minimal role in improving F1 scores for few model combinations, as textual data was the main driver.

The multimodal models all had the combination of textual and visual encoders. Across them, we observed that training stability improved and overfitting reduced. This was in comparison to the large-capacity unimodal models that we evaluated. However, the performance results were very close in number. For example, two pairings were DenseNet201 + ALBERT and VGG16 + BERTweet. These were high-capacity text models paired with smaller CNNS. These pairings had consistent F1 scores across fake and real classes. They only had slight drops between the training and test sets. This shows better prediction than vision-only models. In contrast, there were multimodal pairings with large ViTs. Examples of this include DeBERTa + EfficientViT and DistilBERT + DeiT. These pairings sometimes had near-perfect training F1 scores (approximately 0.97-1.0). However, these scores were much lower for the real class on the test set (0.77-0.79). This is because of overfitting. Overall, multimodal models usually outperformed their unimodal parts. This was especially the case for harder real-class examples. However, score increases over strong NLP models only were limited to 1-2 F1 points. Multimodal fusion led to a small increase in accuracy. Also, these models were more consistent. Overall, these patterns show that the strongest signals within Fakeddit are textual.

We compared unimodal and multimodal results. They showed that text-based models alone had high results. DeBERTa alone has test F1 scores of 0.90 for fake news and

0.86 for real news. DeBERTa + InceptionResNetV2 only improves to 0.91 overall. This shows a big factor in multimodal fake news detection. The usefulness of text or images depends on several items. These are the nature of the dataset and the type of misinformation. In datasets like Fakeddit, text has the strongest clues, like punctuation or errors with real-world facts. In these cases, strong NLP models can find most of the clues they need to predict real and fake posts. Sometimes, text alone is enough for high accuracy. Adding image features through multimodal fusion only gives small improvements.

In the real world, misinformation is not always shown using just text. Many posts have captions that are neutral or correct in a hidden way. These posts become misleading only with context, when looking at the full picture. In these cases, the images become more important. Misleading images, deepfakes, or memes then have the stronger clues. In these cases, models like EfficientNet, InceptionResNetV2, or Swin Transformer can be used. Multimodal approaches also help with prediction when the images and text do not match. Social media posts often have text and images with different clues. Text is stronger in some datasets. Multimodal fusion is stronger in others. These include when the clues are hidden, when the pictures have the main clues, or when there are errors between the text and image.

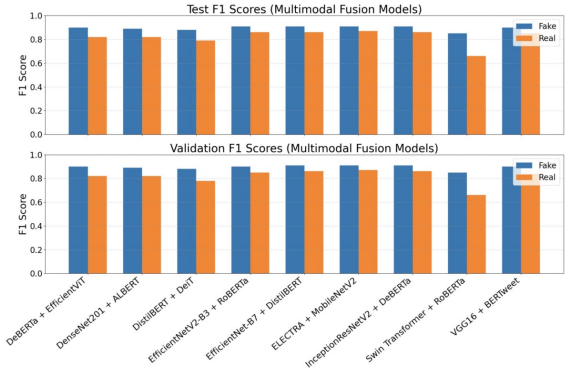

Figure. 7 *(top)* Test F1 scores and *(bottom)* validation F1 scores to classify real and fake context (images and text) using the fine-tuned multimodal approach

Limitations

In our study, we test various unimodal and multimodal models for fake news detection. However, our study has several limitations, Firstly, the models in our study were only trained on a subset of Fakeddit because of computational limits which limits the impact of models to find less common or hidden clues in large-scale misinformation. Secondly, we have limited our analysis to only binary classification where data is classified as fake or real. This limits data analysis simpler but does not show how complex misinformation really is. While Fakeddit has multi-class labels like satire, manipulation, and imposter content, these are not explored, as we only do testing using binary classification.

Our study also has limitations related to posts that come mixed content where text and images can have different clues. In these cases, binary labeling would force these posts into just one category while hiding small and hidden errors which can especially hurt unimodal models. The models are then at risk to predict a post wrong when only the text or only the image has strong clues. While multimodal models are overall in a better spot to solve these cases, we have not separated or tested these cases on their own. Because of this, we cannot measure exactly how big this advantage is. Similarly, we did not build a test group just for text-and-image mismatches and made inferences about how strong multimodal models are in these scenarios. These are however based on overall scores instead of a specific test. Studies in the future should prioritize and test these specific groups to understand their impact on model accuracy across text and images.

A large problem is that we only used the Fakeddit dataset. This dataset is mostly made of Reddit posts and is a large, well-labeled benchmark. However, using only one dataset makes it hard to know if the models work well in general. Models may learn patterns from one social media platform instead of clues from overall misinformation. Because of this, their scores might just show that the models got used to Fakeddit. It might not mean that the models are strong enough to work on data from other platforms. We did not test the models across other datasets, which limits our testing in a few ways. For example, we have not analyzed the accuracy of the models in other social media platforms and the impact of data bias and site changes. In the future, work should focus on testing across different datasets to help models learn from variety of platforms and include testing on other benchmarks like Twitter-based and LIAR. Additionally, our study does not test the models on data distribution changes as the tests were done on fixed data. Specifically, we did not check how the models work for damaged pictures or changed text and thus, the scores in our study only show how the models do in our lab conditions, not in the changing real world.

The practical challenges related to training these models was clear in our study with a variety of issues related to both text and image encoders. Both of these are very sensitive to hyperparameter choices which made tuning difficult. As GPU memory constraints limited us to use smaller batch sizes, we depended on gradient accumulation which further limited our ability to search different hyperparameters and test larger architectures. Hence, our analysis was limited to computationally feasible limits and not the highest performance. Finally, fine-tuning the valuable pre-trained models could have introduced issues related to training-data bias and thus needing

strict training rules to avoid overfitting and forgetting.

The other limitations were related to memory and runtime testing as the EfficientNet-based image models and transformer-based NLP models significantly increased compute cost. Hence, we lowered the number of epochs and ablation tests in our study. We also used two lightweight models - MobileNetV2 and BERTweet which used under 8 GB of GPU memory and trained in under an hour per epoch. They were effective and worked with over 150 samples per second. Additionally, we analyzed high-capacity models - EfficientViT and Swin Transformer paired with DeBERTa or RoBERTa. These pairings needed 24-32 GB, several hours per epoch, and they could only work with 20-30 samples per second. Larger models usually had higher results. However, they would be too slow and bulky to be used greatly in high-volume, real-time online spaces. This demonstrates another significant trade-off between efficiency and high score. Model size as expected was another determinant of model performance with larger models both image and text having drops on test data. This mostly happened for the real class, which is usually because of overfitting. Multimodal models only had a small improvement from this. But gains were very small and mostly caused by text. This shows that we need stricter model training or much more data.

Finally, the math checks are a limitation. Many models only score 0.01-0.02 points different from each other in F1 score. Yet, each model was only trained once without testing different starting points or math checks. Because of this, small performance differences may be just due to random predictions instead of real improvements. This makes it hard to be confident in which model is truly better. Future work should include running the tests multiple times, sharing the score spreads, and using math checks to prove if the models are consistent.

Conclusions

Fake news is a growing problem and in our study, we evaluated the impact of text and image models for fake news detection. This was performed as both unimodal models when they were tested separately or combined multimodal models. A variety of pretrained and benchmarked models were used for analysis. These included CNNs, ViTs and transformer-based language models. The Fakeddit dataset provides binary classification as fake or real. Given that it has text and image data, we were able to compare their relative effects to spot fake news using both unimodal and multimodal models.

Unimodal image models demonstrate that images alone though limited have still useful clues. Most CNNs get average test F1 scores ranging from 0.45-0.60. ConvNeXt-XL performs the best with test F1 scores of 0.78 for fake images and 0.69 for real images. In comparison, ViTs perform better as EfficientViT has test F1 scores of 0.82 for fake images and 0.75 for real images. However, image-only models have significant limits given that misleading posts often use completely real images in hidden or difficult ways. Our study shows the NLP models are consistently accurate in distinguishing real from fake posts suggesting the stronger role of text-based content in Fakeddit. Among transformer-based models with F1 scores of 0.85-0.90, DeBERTa was leading with highest test F1 scores of 0.90 for fake content and 0.86 for real content. Our results are in agreement with prior studies where text-based clues align with the framing of the story, the emotions and the variable differences between text and images even within the same post. These clues have been found to be great for fake news detection. The baseline of text-only methods is much higher compared to image-only options. In comparison, when we used multimodal models, the F1 scores were high ranging from 0.87-0.91. While these scores were high, the marginal increase was small compared to text-only models. This highlights that combining DeBERTa with image-only models leads to marginal improvement and the results were likely dominated by text-based clues with images that only give a few extra clues. The multimodal fusion models will likely be valuable in scenarios where text and images provide contradictory data.

Our study has practical implications for building fake news detection systems in real world as they emphasize the impact of simple text-only models with high accuracy. We however cannot ignore the limitations of systems which have only unimodal data like images or text as the information in them can sometimes be neutral. Multimodal models are particularly valuable to learn across variable data formats which would retain high value in the era of deepfakes. Future studies should focus on models that could use fusion methods to accurately classify fine grained multi-class classification. They should also include tampered data in different platforms. Overall, our study evaluated benchmarked unimodal and multimodal models and highlights their strengths and limits for fake news detection.

References

1 E. Aïmeur, S. Amri, G. Brassard. Fake news, disinformation and misinformation in social media: a review. Social Network Analysis and Mining. Vol. 13, pg. 30, 2023, https://doi.org/10.1007/s13278-023-01028-5.

2 S. Vosoughi, D. Roy, S. Aral. The spread of true and false news online. Science. Vol. 359, pg. 1146–1151, 2018, https://doi.org/10.1126/science.aap9559.

3 BSA 42 — politics and social media. National Centre for Social Research https://natcen.ac.uk/publications/bsa-42-politics-and-social-media 2025.

4 K. Ognyanova, D. Lazer, R. E. Robertson, C. Wilson. Misinformation in action: fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School

Misinformation Review. 2020, https://doi.org/10.37016/mr-2020-024.

5 N. Anstead, L. Edwards, S. Livingstone, M. Stoilova. The potential for media literacy to combat misinformation: results of a rapid evidence assessment. International Journal of Communication. Vol. 19, pg. 23–23, 2025,.

6 M. Westerlund. The emergence of deepfake technology: a review. Technology Innovation Management Review. Vol. 9, pg. 39–52, 2019, https://doi.org/10.22215/timreview/1282.

7 S. Yakkundi, R. Patil, S. Sangani, R. H. Goudar, S. I. Goudar, A. Qazi. Exploring machine learning for fake news detection: techniques, tools, challenges, and future research directions. Discover Applied Sciences. Vol. 7, pg. 873, 2025, https://doi.org/10.1007/s42452-025-07548-3.

8 J. Rout, M. Mishra, M. J. Saikia. Towards reliable fake news detection: enhanced attention-based transformer model. Journal of Cybersecurity and Privacy. Vol. 5, 2025, https://doi.org/10.3390/jcp5030043.

9 K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu. Fake news detection on social media: a data mining perspective. SIGKDD Explor. Newsl. Vol. 19, pg. 22–36, 2017, https://doi.org/10.1145/3137597.3137600.

10 W. Y. Wang. "Liar, liar pants on fire": a new benchmark dataset for fake news detection. in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (eds R. Barzilay & M.-Y. Kan) pg. 422–426, Association for Computational Linguistics, Vancouver, Canada, 2017. https://doi.org/10.18653/v1/P17-2067.

11 D. Ghiurău, D. E. Popescu. Distinguishing reality from ai: approaches for detecting synthetic content. Computers. Vol. 14, 2024, https://doi.org/10.3390/computers14010001.

12 M. Upkare, T. Gogawale, V. Hadoltikar, O. Gurao, V. Hajari, S. Goski. DeepFakeGuard- a deepfake detection website using machine learning. in pg. 626–640, Atlantis Press, 2026. https://doi.org/10.2991/978-94-6463-948-3_44.

13 S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh. SpotFake: a multi-modal framework for fake news detection. in 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) pg. 39–47, 2019. https://doi.org/10.1109/BigMM.2019.00-44.

14 S. Qian, J. Wang, J. Hu, Q. Fang, C. Xu. Hierarchical multi-modal contextual attention network for fake news detection. in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval pg. 153–162, Association for Computing Machinery, New York, NY, USA, 2021. https://doi.org/10.1145/3404835.3462871.

15 Y. Jiang, T. Wang, X. Xu, Y. Wang, X. Song, D. Maynard. Cross-modal augmentation for few-shot multimodal fake news detection. Preprint at http://arxiv.org/abs/2407.12880 2024 https://doi.org/10.48550/arXiv.2407.12880.

16 K. Nakamura, S. Levy, W. Y. Wang. Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection. in Proceedings of the Twelfth Language Resources and Evaluation Conference (eds N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis) pg. 6149–6157, European Language Resources Association, Marseille, France, 2020.

17 S. Kumari, M. P. Singh. A deep learning multimodal framework for fake news detection. Engineering, Technology & Applied Science Research. Vol. 14, pg. 16527–16533, 2024, https://doi.org/10.48084/etasr.8170.

18 K. Saha, Z. Kobti. DeBERTNeXT: a multimodal fake news detection framework. in Computational Science – ICCS 2023 (eds J. Mikyška, C. De Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra & P. M. A. Sloot) Vol. 14074 pg. 348–356, Springer Nature Switzerland, Cham, 2023. https://doi.org/10.1007/978-3-031-36021-3_36.

19 S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh. VQA: visual question answering. in 2015 IEEE International Conference on Computer Vision (ICCV) pg. 2425–2433, IEEE, Santiago, Chile, 2015. https://doi.org/10.1109/ICCV.2015.279.

20 L. Y. H. Low, Y.-T. Wu, Y.-H. Liu, J.-H. Wang. Multimodal fake news detection combining social network features with images and text. in Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025) (eds K.-W. Chang, K.-H. Lu, C.-K. Yang, Z.-R. Tam, W.-Y. Chang & C.-C. Wang) pg. 266–276, Association for Computational Linguistics, National Taiwan University, Taipei City, Taiwan, 2025.

21 S. K. Uppada, P. Patel, S. B. An image and text-based multimodal model for detecting fake news in osn's. Journal of Intelligent Information Systems. Vol. 61, pg. 367–393, 2023, https://doi.org/10.1007/s10844-022-00764-y.

22 R. Ștefan Emil, B. Remus. A review of automatic fake news detection: from traditional methods to large language models. Future Internet. Vol. 17, 2025, https://doi.org/10.3390/fi17100435.

23 Y. Shao, J. Sun, T. Zhang, Y. Jiang, J. Ma, J. Li. Fake news detection based on multi-modal classifier ensemble. in Proceedings of the 1st International Workshop on Multimedia AI against Disinformation pg. 78–86, Association for Computing Machinery, New York, NY, USA, 2022. https://doi.org/10.1145/3512732.3533583.

24 L. Feng, F. Tung, H. Hajimirsadeghi, M. O. Ahmed, Y. Bengio, G. Mori. Attention as an rnn. Preprint at http://arxiv.org/abs/2405.13956 2024 https://doi.org/10.48550/arXiv.2405.13956.

25 X. Zhou, R. Zafarani. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surv. Vol. 53, pg. 109:1-109:40, 2020, https://doi.org/10.1145/3395046.

26 M. Hashemi. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation. Journal of Big Data. Vol. 6, pg. 98, 2019, https://doi.org/10.1186/s40537-019-0263-7.

27 J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei. ImageNet: a large-scale hierarchical image database. in 2009 IEEE Conference on Computer Vision and Pattern Recognition pg. 248–255, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

28 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. Preprint at http://arxiv.org/abs/1706.03762 2023 https://doi.org/10.48550/arXiv.1706.03762.

29 J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at http://arxiv.org/abs/1810.04805 2019 https://doi.org/10.48550/arXiv.1810.04805.

30 S. Bird, E. Klein, E. Loper. Natural Language Processing with Python. 2009.

31 Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, L. Shang. Cross-modal ambiguity learning for multimodal fake news detection. in Proceedings of the ACM Web Conference 2022 pg. 2897–2905, Association for Computing Machinery, New York, NY, USA, 2022. https://doi.org/10.1145/3485447.3511968.

32 R. Oshikawa, J. Qian, W. Y. Wang. A survey on natural language processing for fake news detection. Preprint at http://arxiv.org/abs/1811.00770 2020 https://doi.org/10.48550/arXiv.1811.00770.

33 I. Ali, M. N. B. Ayub, P. Shivakumara, N. F. B. M. Noor. Fake news detection techniques on social media: a survey. Wireless Communications and Mobile Computing. Vol. 2022, pg. 6072084, 2022, https://doi.org/10.1155/2022/6072084.

34 M. Zidan, A. Sleem, A. Nabil, M. Othman. Multimodal fake news detec-

tion: a survey of text and visual content integration methods. International Journal of Computers and Informatics (Zagazig University). Vol. 7, pg. 13–25, 2025,.

35 K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. Preprint at http://arxiv.org/abs/1512.03385 2015 https://doi.org/10.48550/arXiv.1512.03385.

36 M. Tan, Q. V. Le. EfficientNet: rethinking model scaling for convolutional neural networks. Preprint at http://arxiv.org/abs/1905.11946 2020 https://doi.org/10.48550/arXiv.1905.11946.

37 Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie. A convnet for the 2020s. Preprint at http://arxiv.org/abs/2201.03545 2022 https://doi.org/10.48550/arXiv.2201.03545.

38 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at http://arxiv.org/abs/2010.11929 2021 https://doi.org/10.48550/arXiv.2010.11929.

39 J. Yosinski, J. Clune, Y. Bengio, H. Lipson. How transferable are features in deep neural networks? in Advances in Neural Information Processing Systems Vol. 27 Curran Associates, Inc., 2014.

40 F. Chollet. Xception: deep learning with depthwise separable convolutions. Preprint at http://arxiv.org/abs/1610.02357 2017 https://doi.org/10.48550/arXiv.1610.02357.

41 K. Team. Keras documentation: inceptionv3. https://keras.io/api/applications/inceptionv3/.

42 K. Team. Keras documentation: inceptionresnetv2. https://keras.io/api/applications/inceptionresnetv2/.

43 M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen. MobileNetV2: inverted residuals and linear bottlenecks. Preprint at http://arxiv.org/abs/1801.04381 2019 https://doi.org/10.48550/arXiv.1801.04381.

44 K. Team. Keras documentation: densenet. https://keras.io/api/applications/densenet/.

45 X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, Y. Yuan. EfficientViT: memory efficient vision transformer with cascaded group attention. Preprint at http://arxiv.org/abs/2305.07027 2023 https://doi.org/10.48550/arXiv.2305.07027.

46 H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jegou. Training data-efficient image transformers & distillation through attention. in Proceedings of the 38th International Conference on Machine Learning pg. 10347–10357, PMLR, 2021.

47 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo. Swin transformer: hierarchical vision transformer using shifted windows. Preprint at http://arxiv.org/abs/2103.14030 2021 https://doi.org/10.48550/arXiv.2103.14030.

48 Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov. RoBERTa: a robustly optimized bert pretraining approach. Preprint at http://arxiv.org/abs/1907.11692 2019 https://doi.org/10.48550/arXiv.1907.11692.

49 H. Zhang, M. O. Shafiq. Survey of transformers and towards ensemble learning using transformers for natural language processing. Journal of Big Data. Vol. 11, pg. 25, 2024, https://doi.org/10.1186/s40537-023-00842-0.

50 Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut. ALBERT: a lite bert for self-supervised learning of language representations. Preprint at http://arxiv.org/abs/1909.11942 2020 https://doi.org/10.48550/arXiv.1909.11942.

51 V. Sanh, L. Debut, J. Chaumond, T. Wolf. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. Preprint at http://arxiv.org/abs/1910.01108 2020 https://doi.org/10.48550/arXiv.1910.01108.

52 K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. Preprint at http://arxiv.org/abs/2003.10555 2020 https://doi.org/10.48550/arXiv.2003.10555.

53 D. Q. Nguyen, T. Vu, A. T. Nguyen. BERTweet: a pre-trained language model for english tweets. Preprint at http://arxiv.org/abs/2005.10200 2020 https://doi.org/10.48550/arXiv.2005.10200.

Supplementary information

The online version contains supplementary material available at https://nhsjs.com/?p=45087