STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–30.
Published online in International Academic Press (www.IAPress.org)

A Comprehensive Evaluation of ViLT and CLIP Transformers for Multimodal Fake News Detection with Metadata Integration

Ahmed M. Abd Elhamid[1,*], Mohamed Waleed Fakhr[2], Mahmoud M Ashry[3], Ahmed Abdelhafeez[4,5]

[1]*College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt*
[2]*Head of Computing School, Coventry University, TKH Branch, Cairo, Egypt*
[3]*Head of the Scientific Departments, College of Computing and Information Technology (Smart Village), Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt*
[4]*Faculty of Computer and Information Technology, Innovation University, Cairo, Egypt*
[5]*Applied Science Research Center, Applied Science Private University, Amman, Jordan*

Abstract The swift expansion of social media has fueled the prevalence of multimodal misinformation wherein textual information, visual material, and context combine to influence perception. In this research, we compare two vision-language transformer-based systems, namely, ViLT and CLIP, in the task of detecting multimodal fake news using two benchmark datasets Fakeddit and FakeNewsNet. These evaluations include binary as well as three-class classifications in both scenarios of metadata available and metadata-free. Furthermore, Late Fusion and Gated Fusion approaches are considered in order to explore the effectiveness of complementation through transformers. The results indicate that the effectiveness of metadata integration depends on the architecture and the dataset. ViLT can gain from metadata integration due to the presence of the unified transformer architecture enabling the interaction between textual, visual, and contextual features inside the attention mechanism. On the contrary, CLIP exhibits better behavior in semantic image-text alignment, especially when there is no metadata. ViLT has an accuracy score of 0.906 for the FakeNewsNet data set and 0.831 for the binary version of Fakeddit task, while CLIP obtains competitive accuracy scores for the three-class version of the same task without metadata. However, the improvements introduced by metadata were small for several cases, which suggests that metadata should not be taken as universally effective.

Keywords Multimodal Fake News Detection, CLIP, ViLT, Cross-Attention Fusion, Gated Fusion, Statistical Significance Testing, Information Disorder.

DOI: 10.19139/soic-3877

1. Introduction

The spread of fake news on online social media platforms has become an issue for society and computing. Nowadays, the dissemination of fake news is not simply restricted to being text-based but also involves the manipulation of misleading text captions, manipulated images, and social context cues that affect the interpretation and dissemination of the information. This means that unimodal techniques for detecting fake news can be inadequate since they do not address semantic conflicts within the multimodal content.

In recent times, there have been some impressive models for this challenge, such as the vision language transformers. The ViLT model utilizes a single-stream approach in which both the visual and text tokens are processed in a transformer framework together. On the other hand, CLIP utilizes a dual-stream approach to learn about semantic consistency between images and texts using contrastive pretraining. Both these approaches have

*Correspondence to: Ahmed M. Abd Elhamid (Email: Ahmed.Abdelhamid@student.aast.edu). College of Computing and Information Technology, Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

shown potential in multimodal learning challenges but have yet to be compared in relation to metadata-based fake news detection.

The present research helps bridge this gap by conducting an assessment of ViLT and CLIP for multimodal fake news identification based on two benchmark data sets: Fakeddit and FakeNewsNet. Specifically, this research will investigate the effect of metadata inclusion versus exclusion from multimodal detection of fake news to find out whether the presence of structured context features such as engagement increases the accuracy of fake news classification. Another aspect under consideration is that of late and gated fusion.

The main contributions of this work are listed as follows:

1. The comparative analysis of the unified-stream vs. dual-stream vision-language transformer models, namely ViLT and CLIP, for multimodal fake news detection tasks.
2. The empirical investigation of the usefulness of metadata incorporation using various data sources and task settings, revealing that the use of metadata depends on the underlying model used.
3. The assessment of multimodal transformer representations' fusion mechanisms, specifically, Late Fusion vs. Gated Fusion.
4. The analysis of the impact of excluding metadata to determine its practical importance in multimodal transformers' operations.
5. The enhanced methodology which takes into account reproducibility experiments and statistical significance tests of results obtained.

As opposed to the workshop study mentioned as [42], which is limited in its scope of experimentation with multimodal transformers at the baseline stage only, the current paper offers much more extensive experimentation. Apart from baseline multimodal transformers' experiments, this paper includes experiments with metadata-aware ViLT modelling, analysis of CLIP metadata, adaptive gated fusion experiments, experiments excluding metadata, statistics testing, and Fakeddit & FakeNewsNet evaluation. Thus, this paper should be seen as an advanced experiment as opposed to just a repetition of an earlier study.

Organization of the Rest of This Paper

The rest of this paper is structured as follows. Section 2 provides the literature review about fake news detection and multimodal misinformation detection. Section 3 covers the datasets used, metadata processing, model architectures, fusion techniques, and the experiment settings. Section 4 provides the experiment results and analysis. Section 5 provides the statistical analysis, limitations, and practical applications. The paper concludes with Section 6.

2. Literature Review

Through previous research, the focus is diffuse data on social networks, which the shape is changed with texts and texts with image, Accordingly, the linguistic style and emotional nature changed until the user is convinced of the data validity until it is published on a large scale and from different sources So as not to be tracked. The reliability of data is classified in different ways by using a single model or a multimodal to detect the degree of manipulation of textual information and images. Performance depends on the development of the model and the development of the technology used on an ongoing basis over time.

When compared to current multimodal methods for detecting fake news, there are distinct advantages of the current study. Firstly, this study considers the role of metadata when it comes to the incorporation of transformers in multimodal models. Secondly, it carries out an analysis that includes two completely distinct types of vision-language transformers, i.e., ViLT and CLIP. Thirdly, this research uses an adaptive fusion mechanism that weighs the text and image features based on their classification capability.

Such features make this research model unique and different from other previous studies that considered multimodal detection of misinformation.

Table 1. Comparison of Recent Multimodal Fake News Detection and Vision-Language Frameworks

Model / Framework	Main Architecture	Text	Image	Metadata	Fusion / Interaction Strategy	Main Strength
SpotFake	BERT + CNN/VGG	Yes	Yes	No	Feature concatenation	Simple and stable multimodal baseline
MVAE / MFAE-style models	Multimodal autoencoder / attention	Yes	Yes	Limited	Attention or reconstruction-based fusion	Learns shared multimodal representation
VisualBERT	Unified transformer	Yes	Yes	No	Shared transformer attention	Captures implicit word-region interaction
LXMERT	Multi-encoder transformer	Yes	Yes	No	Language, vision, and cross-modal encoders	Strong explicit cross-modal reasoning
CLIP-based methods	Dual-stream contrastive transformer	Yes	Yes	Limited	Image-text semantic alignment	Strong caption-image mismatch detection
BLIP / BLIP-2-based methods	Vision-language pretraining	Yes	Yes	Limited	Image-text representation and generation	Strong visual-language understanding and caption-based reasoning
LVLM / evidence-retrieval methods	Large vision-language models	Yes	Yes	Sometimes	External evidence retrieval and reasoning	Strong explainability and knowledge-based verification
Flamingo-style VLMs	Few-shot vision-language model	Yes	Yes	No	Gated cross-attention and in-context learning	Flexible few-shot multimodal reasoning
Proposed ViLT evaluation	Unified-stream transformer	Yes	Yes	Yes	Shared attention with metadata-aware setting	Effective metadata-aware joint representation
Proposed CLIP evaluation	Dual-stream contrastive transformer	Yes	Yes	Yes	Semantic alignment with external metadata fusion	Strong image-text alignment behavior
Proposed Fusion Analysis	ViLT + CLIP representations	Yes	Yes	Yes	Late Fusion and Gated Fusion	Tests complementary transformer behavior

2.1. Relationship to Previous Workshop Publication

This research is an extension of a previous workshop paper that we have authored [42], with additional advancements from a methodological perspective.

Our previous research concentrated on multimodal fake news detection at baseline through the use of transformer-based models. However, it did not consider transformer learning that included metadata, adaptive gating of fusion, or any analysis regarding excluding metadata.

This current study considers the influence of the combination of transformer learning that includes metadata, and its influence on the multimodal fusion strategy.

We conduct comparisons between two architectures, specifically ViLT and CLIP architectures for the purpose of identifying their behaviour during multimodal learning.

Table 2. Comparison Between the Workshop Paper and the Current Study

Component	Workshop Paper (2025)	Current Study
Baseline Transformer Evaluation	✓	✓
ViLT Metadata Integration	×	✓
CLIP Metadata Analysis	×	✓
Gated Fusion Mechanism	×	✓
Metadata Exclusion Study	×	✓
Comparative Fusion Analysis	×	✓
Statistical Significance Analysis	×	✓
Reproducibility Enhancements	×	✓
Expanded Experimental Evaluation	Limited	✓

2.2. Techniques for Fake News Detection

This field, as discussed in reference [6], is organized into four major approaches, focusing on knowledge, stylistic features, source characteristics, and patterns of information diffusion. The knowledge-based methods are centered around the idea of fact-checking [7], an essential component of modern journalism that has been broadened beyond just facts [8]. However, the process is slow as it is an exhaustive task, and the sheer volume of online data necessitates the development of more advanced technology [9]. The style-based methods, as the name suggests, examine the style of the creation of the data, with the assumption that the data created has a particular style or trend [10]. Unlike the other types, source-centered methods focus on how the data's creator behaves, i.e., their reading and publication habits and try to identify anomalies in the data. However, these methods, though effective, are now not as relevant in the current environment [11].

2.3. Single Mode Detection Methods

The basic concept here is the content itself, i.e., the actual substance of the content, and we examine the content with the help of natural language processing techniques to detect disinformation in the content itself. Earlier, content detection techniques were based on traditional machine learning methods [12, 13], but they required a lot

of data before they could produce good results. However, the advent of the transformer model has simplified the process, as it has been found that BERT has been effective in detecting disinformation in content [14]. However, the focus today has shifted to the application of linear learning for detecting disinformation in content [15, 16], as it not only helps in content verification but also in attitude, dependency, and sentiment analysis, thereby aiding in the detection of disinformation [17]. Images are employed in fighting fake news and disinformation, with the main idea being the detection of manipulated images. In this case, different approaches are used, which range from traditional image processing and deep learning with CNNs [18]. These approaches are used in the detection of manipulated images. They search for inconsistencies such as shadows, lighting, object edges, and the presence of deep-fake images created using GANs. However, it should be highlighted that the text that is usually embedded in such images is not captured. In addition, it's worth noting that images and text are the most common forms of carrying out disinformation [19]. However, other forms are possible. Therefore, the most popular approaches in the detection of disinformation involve images and text. Therefore, in this case, our main interest is images and text.

2.4. Multimodal and State-of-the-Art Baseline Comparisons

Recent multimodal fake news detection papers have used transformer-based and vision-language models in order to model the semantics between text and images. Earlier works, like SpotFake, had shown that using the text features from BERT along with visual features from a convolution neural network could detect fake news more accurately than unimodal models. However, these approaches typically relied on feature concatenation and didn't properly account for fine-grained cross-modal interaction and metadata-driven context.

The use of attention mechanisms and transformer models to solve the problem of cross-modal modeling came into the picture afterward. VisualBERT uses a shared transformer architecture to process textual tokens and visual region features, resulting in implicit interactions between words and visual regions. LXMERT goes one step ahead by having separate encoders for the language, vision, and cross-modality components. Though these models exhibit better performance at cross-modal reasoning, they weren't originally created for metadata-aware fake news detection purposes.

Recent work focuses on multimodal fake news detection through large scale vision-language pre-training approaches. CLIP is widely used due to its ability to learn a common semantic space by employing contrastive learning in order to jointly represent image and text embeddings. Therefore, CLIP is particularly suitable when it comes to detecting semantic discrepancies between the caption and the actual image. Nevertheless, due to the dual stream architecture of CLIP, it cannot incorporate the structured data (i.e., engagement scores or information extracted from comments) unless an extra fusion component is used.

Another class of models which became increasingly popular in the field of multimodal fake news detection are the BLIP and BLIP-2 based models since they employ visual reasoning capabilities along with either text generation or multimodal encoding of image-text pairs. For instance, by using the framework of the BLIP models, one can easily generate/encode the visual descriptions that indicate semantic inconsistencies between the image and the textual statement of the fake post.

Recently, some vision-language model-based solutions to detect multimodal fake news go further than the traditional classification methods. Some studies rely on multimodal evidence retrieval techniques to search for external evidence prior to performing the detection tasks. Other studies leverage vision-language model techniques for explainable out-of-context fake news detection. Both strategies provide great help since detecting fake news frequently involves the need for external knowledge reasoning. Nevertheless, the use of such approaches typically entails heavy reliance on computing power.

In contrast to these studies, the current study is concerned with a comparison between ViLT and CLIP in both metadata-aware and metadata-free configurations. ViLT refers to the unified-stream transformer architecture in which the interaction among the representations of images, texts, and context could take place in a shared attention space. Meanwhile, CLIP indicates the dual-stream contrastive architecture that is designed for optimizing the image-text semantically alignment. It is worthwhile to make this comparison because the inclusion of metadata does not have equal impacts on all vision-language transformers.

Table 3. Summary of Fake News Detection Methods

Ref.	Approach / Technique	Dataset	Accuracy Results	Pros	Cons
[28]	Federated Deep Learning: CNN, CNN+LSTM, Multiple LSTM	COVID-19 / fake_news.csv / real_news.csv	CNN: 95.0%, CNN+LSTM: 96.0%, Multiple LSTM: 97.0%	Highly effective at sequential text patterns	Blind to visual image manipulations
[29]	Comparative Text Modeling: CNN/LSTM, BERT Core, Classical Baselines	MMCovid19, PubHealth, Covid19FakeNews, ISOT, GRAFN, Q-Proppy	92.5%, 76.9%, 96.5%, 99.5%, 91.4%, 94.0%	Extensive evaluation across diverse linguistic datasets	Operates entirely within text-only domains
[30]	Multi-Feature Analysis: TF-IDF Text Features, CNN + BiLSTM	ISOT, FA-KES	Combined Test: 99.0%	Combines static statistical weights with deep sequences	High feature dimension; limits real-time scale
[31]	Vision Transformer (ViT): Standalone Patch Self-Attention	Kaggle Deepfake Sets	Deepfake Image Test: 89.91%	Captures fine-grained pixel distortions and tampering	Cannot ingest accompanying text or metadata
[32]	Deep Visual Encoders: VGG, ResNet, Xception Classifier	DFDC (DeepFake Detection Challenge)	VGG: 83.7%, ResNet: 68.7%, Xception: 87.5%	Deep convolutional layers detect spatial artifacts well	Vulnerable to novel generative AI filter styles
[33]	Visual Error Level Analysis: ELA + EfficientNetB0	CASIA 2.0	Tampering Check: 96.11%	ELA highlights localized saving compression differences	Relies strictly on metadata-level image save states
[34]	Diffusion Image Checking: AGFE Model + ViT + CNN	Columbia, RAISE, LSUN, BOSSbase	Synthetic Check: 88.5%	Tuned for text-to-image generator artifacts	High compute overhead during spatial alignment
[14]	FakeBERT Framework: Bidirectional Transformer Encoders	Fake-News Text Set	Semantic Text Test: 98.9%	Excellent contextual word relationship extraction	Blind to accompanying image variables
[16]	Large Language Models: GPT-4 Zero-Shot API, RoBERTa Fine-Tuned	LIAR Dataset	GPT-4: 81.1%, RoBERTa: 80.4%	Strong zero-shot generalization across contexts	High processing cost and inference latency
[19]	Contrastive Dual-Stream Alignment: CLIP + ViT-L/14 Backbone	Fakeddit Benchmark	Multimodal Baseline: 83.7%	Maps text and image vectors to a shared semantic space	Lacks fine-grained cross-modal token-to-patch attention
[21]	Multimodal Fusion & Alignment (MFAE): Entity-Level Attention Model	TWITTER, WEIBO	TWITTER: 89.5%, WEIBO: 96.7%	Explicit cross-attention aligns specific named entities	Operates strictly on content; ignores social metadata
[35]	SARD Framework: CLIP Contrastive + ResNet Core	Gossipcop, PHEME, MR2E, MR2C	Gossipcop: 91.98%, PHEME: 91.27%, MR2E: 89.86%, MR2C: 89.75%	Integrates social context threads with semantic alignment	Complex architecture with heavy parameter footprint
[36]	Compact Parameter Blocks (CPBs): VisualBERT+CPB, ViLBERT+CPB	FBHM, MMHS150K, MultiOFF, MEME	FBHM: 64.0/63.0%, MMHS150K: 74.0/57.0%, MultiOFF: 60.0/61.0%, MEME: 86.0/86.0%	Enhances attention paths while improving context	High structural complexity; limited deployment settings
[37]	TTEC Integration: Contrastive Learning + BERT + ResNet	ReCOVery	Content Verification: 71.9%	Contrastive loss forces mutual alignment of features	Poor convergence on long-tail samples
[38]	HTBERT Optimization: Hyperparameter-Optimized BERT + ResNet	PolitiFact, BuzzFeed, FakeNewsNet, ISOT/LIAR	BERT Module: 92.1%, ResNet Module: 85.6%	Optimized learning rates and layer freezing thresholds	Text and vision modalities remain unaligned
[39]	Joint Multimodal OSN Classifier: BERT Text Core + Xception Vision	Fakeddit, CrowdFlower	Multimodal Mix: 91.94%	Processes text, image assets, comments, and post metadata	Concatenation-based late fusion skips token interactions
[40]	CLIP-Based Detection Pipeline: Contrastive Text/Vision Mapping	Weibo, PolitiFact, Gossipcop	Combined Benchmark: 90.3%	Leverages robust semantic alignment from pre-training	Modality encoders are frozen; limits fine-tuning
[41]	Vision-and-Language Transformer (ViLT): Single-Stream Token Transformer	MSCOCO Multi-Task Benchmarks	VQA Task: 71.26%, NLVR Task: 76.13%	Drops isolated visual feature extraction blocks for speed	Lower fine-grained accuracy on detailed visual targets
[5]	SpotFake Baseline: BERT Core + VGG-19 Vectors	Twitter Content, Weibo Content	Early Fusion Test: 89.2%	Simple and highly stable foundational early fusion	Static concatenation lacks cross-modal token interaction
[42]	VLM Generalization Pipeline: Standalone ViT & CLIP Encoders	Fakeddit Dataset	2-way Tasks: 78.8%, 3-way Tasks: 75.0%	Establishes explicit baselines for standalone VLM use	Lacks dynamic fusion mechanisms or metadata tokens

At the input stage, early merging combines information from multiple media sources, allowing the model to process them together from the start, merge links between media. However, this method requires precision and high processing costs [21].

On the other hand, delayed merging handles each media type separately before combining their outputs. While efficient, it may overlook correlations between media that are essential to ensure precise detection [23]. Hybrid merging overcomes the limitations of the earlier two approaches by integrating information across multiple stages, while using attention-driven merging maps dynamic features from different media based on their significance. The latest multimodal detection techniques often rely on early [24, 5] and late [25] fusion methods. Given the strengths of integrated approaches, more investigations are currently in progress utilizing fusion methods based on attention mechanisms [23]. This article introduces a new RNN (Recurrent Neural Network) incorporating attention into RNNs to unify multimodal information to efficiently detect rumors, improving efficiency compared to early or late fusion. Multimodal converters extend the scope of attention-driven integration using attention mechanisms that focus on the input itself to fuse attributes through media. In contrast to conventional fusion techniques, these models eliminate the need for manual feature extraction before processing. Alternatively, these models transform the inputs into a structured representation directly within a shared converter architecture, enabling interaction between features. Frameworks such as LXMERT and ViLT employ multimodal converters to process images and text together, achieving sophisticated and efficient performance in vision and language tasks [26, 27], as shown in Table 4.

Experimental Methodology

The experiment design for testing ViLT and CLIP for multimodal fake news detection is described below. In order not to confuse independent transformer evaluation with fusion-based experiments, the methodology is divided into three distinct experimental Parts.

Isolated transformer assessment. The models in this Part include ViLT and CLIP. Both models undergo training and testing with respect to metadata-aware and metadata-free settings. It seeks to assess the individual ability of the transformer models in relation to the effect of metadata.

Evaluation on the effect of metadata ablation. For Part 2, two runs on the same model architecture and classification setup are made one run including metadata and another excluding metadata. In this way, the effect of metadata on the performance of the classifier can be measured directly. Analysis is done separately for ViLT and CLIP since both models differ in the way that they process multimodal data.

Fusion-based evaluation. In Part 3, Late Fusion and Gated Fusion approaches are examined. These analyses are not classified under the categories of ViLT and CLIP results as standalone models because they aim to determine if combining transformers will yield better results. Therefore, the fusion-based results are shown in other tables apart from ViLT and CLIP results tables.

The pipeline involves data preprocessing, image and text pairs creation, metadata pre-processing, feature extraction, model training, validation, testing, and statistical analysis. Fakeddit is used for binary and three-class classifications, and FakeNewsNet is used for binary classification. In both cases, the two primary inputs considered are the text and image. Metadata features are employed only in metadata aware experiments.

It should be ensured that the reported findings will be interpreted in a proper manner. Standalone performances of the ViLT and CLIP models help to analyze the effectiveness of unified stream versus dual stream transformers. The ablation experiments help to examine the impact of contextual features. Finally, fusion experiments help to determine whether the late or gated fusion technique helps achieve higher performance.

Figure 1 presents the multimodal fake news detection experiment pipeline overview. The process consists of several stages including preprocessing, creation of image-text pairs, metadata processing, feature extraction, transformers evaluation individually, fusion evaluation, and performance evaluation.

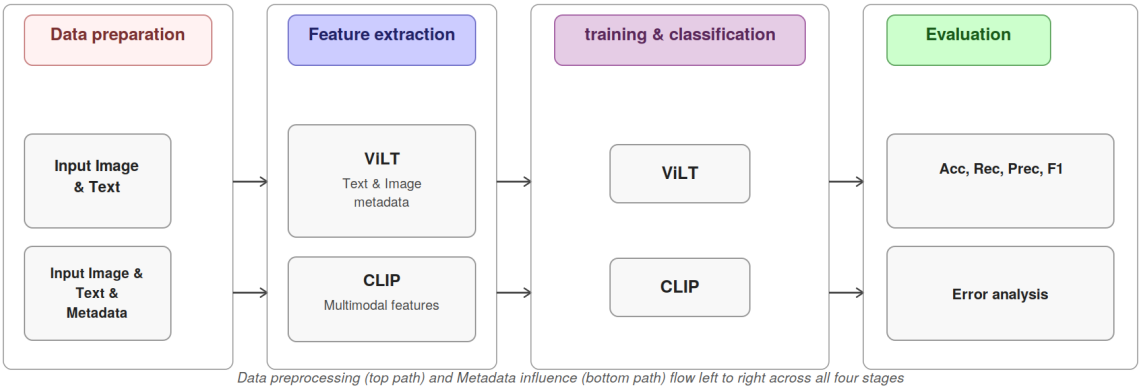


Figure 1. Overall Experimental Pipeline for Multimodal Fake News Detection

3. Methodology

3.1. Experimental Pipeline and Architectural Design

This experimental pipeline assesses the two transformer multimodal models ViLT and CLIP, both in binary and ternary classifications of fake news, using Fakeddit and FakeNewsNet datasets.

The design of this methodology has been altered to conduct experiments based on the following research questions:

RQ1 The role of metadata inclusion in the behavior of multimodal transformers.
RQ2 Comparative analysis between unified stream and dual stream models.
RQ3 Performance analysis of Late Fusion and Gated Fusion approaches.
RQ4 Semantic alignment under multimodal misinformation scenarios.

3.1.1. ViLT (Vision-Language Transformer) Architecture The model we are using is called Vision-Language Transformer (ViLT). In ViLT, the image patches and text patches are combined in a single transformer encoded file. The processing of visual information is done by splitting the images into patches of fixed size together within the transformer layers. Since the ViLT model incorporates a single stream, the Transformer can pay attention to both the metadata tokens and image/text patches simultaneously. The advantage of this model is that it performs better in fine-grained classification in 2-way and 3-way classification when metadata is provided. However, ViLT is quite good at performing straightforward 2-way classification (real vs. fake). However, in 3-way classification, it does not possess the natural alignment scale of CLIP for the mixed class. ViLT finds it difficult to accurately detect fake news because its visual processing capability is limited. ViLT's performance was improved when using data containing long-form news content and multiple images for detecting fake news [41]. The CLS token representation obtained from the output of the final transformer encoder layer serves as the multimodal feature representation.

Figure 2 shows the architecture of ViLT for metadata-enhanced multimodal fake news detection. Text tokens, image patches, along with an additional metadata representation (where applicable), go through a common transformer architecture, and the [CLS] representation is then used for classification.

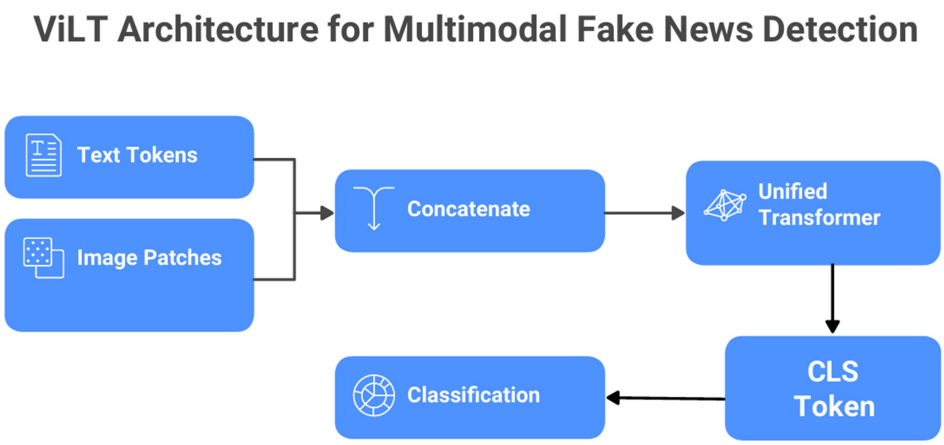


Figure 2. ViLT Architecture for Metadata-Aware Multimodal Fake News Detection

3.1.2. CLIP (Contrastive Language-Image Pretraining) Architecture Dual-stream encoder with embedding vectors from visual and textual data is employed for multimodal representation learning. Visual and textual inputs go through their respective encoders while the model runs using the dual-stream encoding paradigm. The pretraining of CLIP on millions of image-text pairs allow this model to function as a 'matching scale' which indicates the amount of semantic alignment between the two modalities. For the task of 3-class classification (real/fake/mixed), the ability of CLIP to measure such semantic alignment proves to be very useful because of the need for identifying the 'mixed' class based on precise measurement of agreement between text and image. However, due to the inability of the dual-stream alignment architecture of CLIP to natively support structured cross-modal attention, the metadata likes, comments, and upvote ratio tend to become nuisance variables. The CLIP model tends to perform less effectively when applied to unseen data that was not included during training. Wise attention is used to evaluate the influence of textual, visual, and combined feature data adaptively [40]. Data is analyzed for texts and images through scenario, linguistic, visual, and social features, which works to enhance the performance of the CLIP model in fake news detection [43]. The image and text embeddings are then projected using pretrained dual-stream encoders and concatenated for multimodal representation.

Figure 3 presents the CLIP architecture for semantic alignment between images and text. Images and text are encoded using individual encoders and are then mapped to a common embedding space based on semantic alignment.

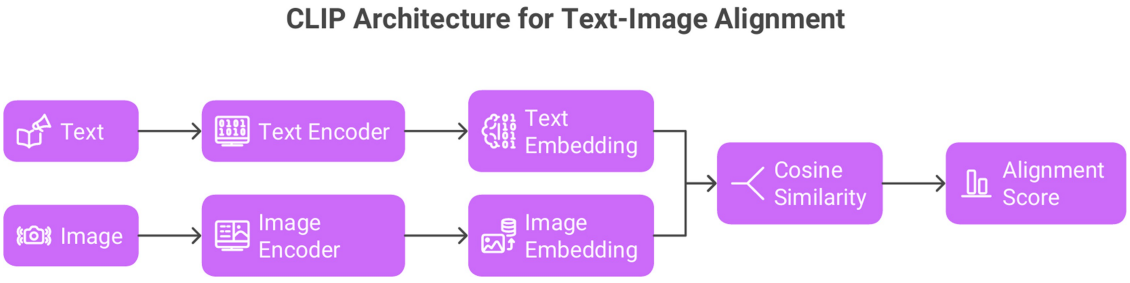


Figure 3. CLIP Architecture for Image-Text Semantic Alignment

3.2. Fusion Architecture and Mathematical Formulation

Fusion experiments were intended for the combination of learned complementary representations from ViLT and CLIP. In the individual evaluation experiments, ViLT and CLIP are assessed individually. However, in the fusion experiments, learned representations from the two models are fused prior to the classification stage.

Let T be the textual representation, V be the visual representation, and M be the metadata feature vector. Since the dimensions of these feature vectors could be different, we project them to a hidden layer of size d:

$$T = W_T T + b_T \tag{1}$$

$$V = W_V V + b_V \tag{2}$$

$$M = W_M M + b_M \tag{3}$$

where W_T, W_V, and W_M are learnable projection matrices, and b_T, b_V, and b_M are learnable bias vectors. In the case where metadata is not available, the term for metadata projection is ignored.

For ViLT, the last transformer encoder representation of the [CLS] token is taken as the joint image-text representation. For CLIP, the output of the learned embeddings of both the pretrained text and image encoders is taken.

3.2.1. Late Fusion In the approach of Late Fusion, the features extracted from different modalities are merged after obtaining the representation. For Late Fusion, the concatenated textual, visual, and metadata representations are fed into the classification layer as follows. The Late Fusion Representation:

$$Z_{LF} = (T; V; M) \tag{4}$$

Where the operator $(\cdot;\cdot)$ stands for vector concatenation. The fused representation without metadata becomes as follows:

$$Z_{LF} = (T; V) \tag{5}$$

Class probabilities can be obtained through:

$$y = \text{softmax}(W_c Z_{LF} + b_c) \tag{6}$$

where W_c and b_c stand for trainable parameters of the classification layer. Late Fusion approach is simple to implement as each modality is separately encoded before classification. However, the method assumes constant importance for the available modalities, which cannot be adapted based on individual samples.

Figure 4 illustrates the Late Fusion architecture. Modality-specific features are extracted separately, mapped into a common feature space, concatenated, and fed into the final classifier.

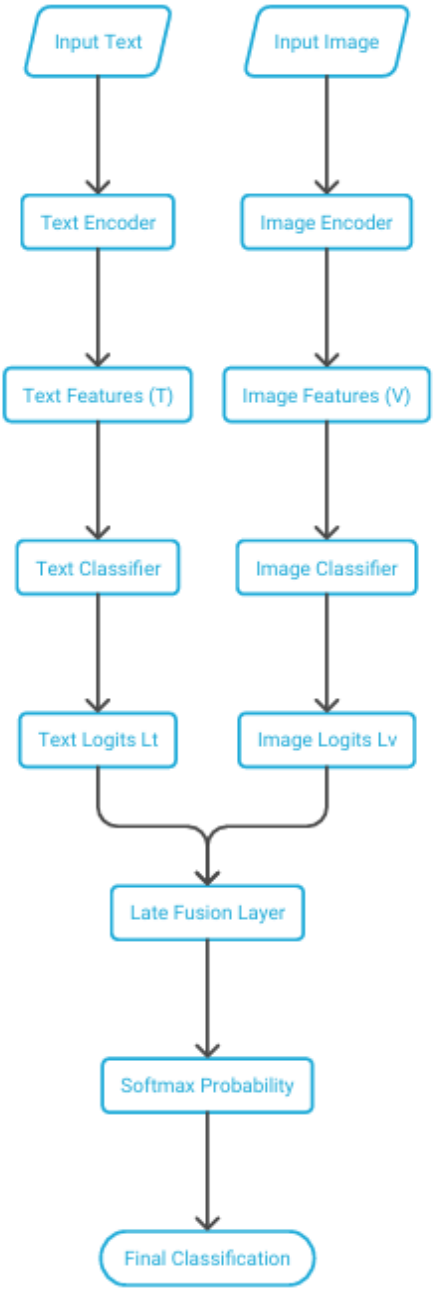


Figure 4. Late Fusion Architecture

3.2.2. Gated Fusion In order to solve this problem, a Gated Fusion technique was used to adaptively weigh each modality depending on their importance. The function of the gating unit is to compute the modality weights depending on the samples. In other words, if the text carries more information than the image, the model would pay more attention to the textual representation.

Firstly, concatenation is done between the projected features of both modalities:

$$H = (T; V) \tag{7}$$

Then, a gating vector can be obtained via a linear transformation combined with the sigmoid function:

$$G = \sigma(W_g H + b_g) \tag{8}$$

Here W_g and b_g denote the gating parameters that need to be learned, while $G \in (0, 1)^d$.

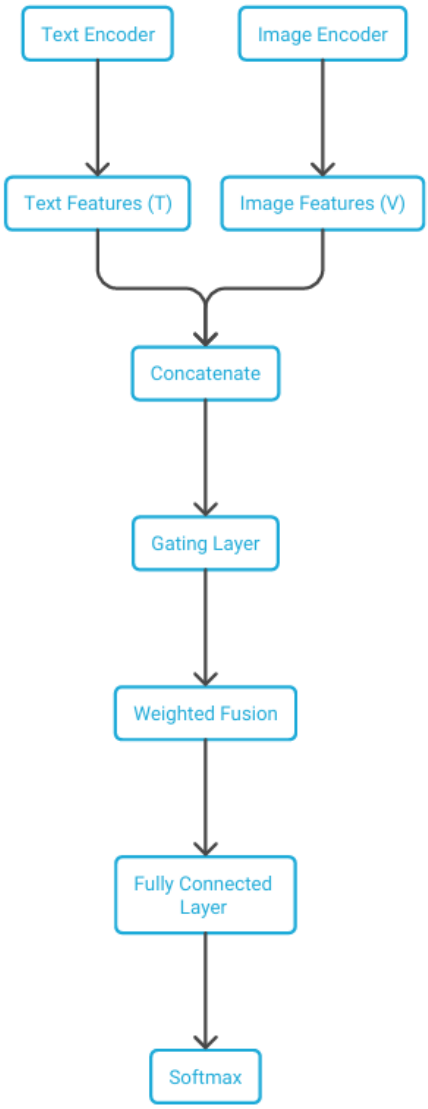


Figure 5. Proposed Gated Fusion Architecture

Calculation of the gated multimodal representation is done as follows:

$$Z_{GF} = G \odot T + (1 - G) \odot V \tag{9}$$

Where $\odot$ represents the element-wise product operation. In cases where there is metadata information available, the metadata representation is concatenated with the gated image and textual representations as shown below:

$$Z_{GFM} = (Z_{GF}; M) \tag{10}$$

Final computation of the class probability scores is performed as follows:

$$y = \text{softmax}(W_c Z_{GFM} + b_c) \tag{11}$$

In cases where no metadata information is available, the representation is directly used for classification as shown below:

$$y = \text{softmax}(W_c Z_{GF} + b_c) \tag{12}$$

The weights for the gating layer can be optimized automatically via backpropagation based on the classification loss function. The gradient is calculated based on the loss function, allowing the network to learn the proper modality weights that will improve prediction performance.

Figure 5 illustrates the Gated Fusion architecture. The gating module computes sample-specific modality weights and fuses textual and visual information prior to (optional) metadata fusion and classification.

3.2.3. Gated Fusion Pseudocode **Algorithm 1: Gated Fusion for Multimodal Fake News Detection**

Input:

- Text representation T
- Visual representation V
- Optional metadata vector M

Output: Predicted class probabilities y

Step 1: Project the textual and visual features into a shared hidden dimension.
Step 2: Concatenate the projected textual and visual representations.
Step 3: Compute the gating coefficients.
Step 4: Compute the gated multimodal representation.
Step 5: If metadata is used, project and concatenate the metadata representation.
Step 6: Pass the final representation to the classifier, where $Z = Z_{GFM}$ when metadata is used and $Z = Z_{GF}$ when metadata is not used.
Step 7: Update all trainable parameters using backpropagation and the weighted cross-entropy loss function.

Combining CLIP and ViLT is neither repetitive nor complementary instead, they detect fundamentally different deception cues, as shown in Table 4.

Table 4. Comparison between CLIP and ViLT and Fusion

Capability	**CLIP**	**ViLT**	**Fusion Necessity**
Semantic Alignment	Excellent (contrastive training optimizes for this)	Moderate (early fusion dilutes semantic signals)	CLIP catches caption-image mismatches
Visual Artifact Detection	Poor (patches aggregated late, semantics dominate)	Excellent (direct patch attention)	ViLT catches deepfakes/memetic manipulation
Metadata Integration	Degrades performance (no native cross-attention mechanism)	Improves performance (unified stream can attend to metadata tokens)	Architectural difference, not hyperparameter
Fine-grained Classification	Strong in 3-way (alignment nuance)	Strong in 2-way with metadata (joint token attention)	Complementary strengths

Fusion Motivation: The above complementary advantages point to an architecture fusion where: CLIP will score semantic alignment (particularly important for 'mixed' 3-way recognition). ViLT will score visual artifacts and metadata (especially important for 2-way with social context).

3.3. Dataset

3.3.1. Fakeddit Dataset The Fakeddit Dataset is used for classification tasks, which consists of more than one million samples in different forms of Fake News text and image with metadata and new multimodal data, and is processed in various stages and classified into two-way, three-way, and six-way categories under external supervision. Attention is paid to classification accuracy based on textual and image models that are experimented on and trained by multimodal Fakeddit Dataset to achieve the best performance in detecting fake news. Metadata is used, which describes the user in terms of their participation and content credibility [44] (as shown in Tables 5, 6).

Table 5. Fakeddit Dataset Label Structure

Class	**2-way**	**3-way**	**6-way**
Classification	True Fake	True Mixed Fake	True Satire/Parody Misleading Content Imposter Content False Connection Manipulated Content

Table 6. Fakeddit Dataset Statistics

Components of Dataset	**Numbers of Contents**
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 – 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

3.3.2. FakeNewsNet Dataset The FakeNewsNet dataset is used to conduct experiments in fake news detection, which comprises data containing temporal information to help in early-stage fake news detection, location information for the fake news dissemination source detection, social context data, and data about various features of news content. Data is collected in the same context from various sources, and FakeNewsNet dataset includes diverse features that enhance the model's ability to detect fake news and address the rapidly changing ways in which misinformation spreads. The data is analyzed to understand its key characteristics, and the FakeNewsNet dataset is used for training to enhance the model's performance in detecting fake news [45] (as shown in Table 7).

Table 7. FakeNewsNet Dataset Statistics

Category	Features	PolitiFact		GossipCop	
		Fake	Real	Fake	Real
News content (Linguistic)	News articles with text	852	1,152	10,270	33,511
News content (Visual)	News articles with images	336	447	1,650	16,767
Social Context (User)	Users posting tweets	95,553	249,887	265,155	80,137
Social Context (User)	Users involved	260,217	766,497	694,660	314,771
Social Context (Post)	Tweets posting news	164,892	399,237	519,581	876,967
Social Context	Tweets	67,156	202,726	193,175	78,756
Social Context	Followers	854,973,017	2,083,711,243	1,249,438,999	601,492,712
Social Context	Average followers	2,740.87	2,022.88	2,024.13	1,916.44
Spatiotemporal (Spatial)	User profiles with locations	217,379	719,331	429,547	220,264
Spatiotemporal (Spatial)	Tweets with locations	3,337	12,692	12,286	2,451
Spatiotemporal (Temporal)	Timestamps for news pieces	296	167	3,558	9,119
Spatiotemporal (Temporal)	Timestamps for tweets	171,301	669,641	381,600	200,531

3.3.3. Dataset Partitioning Strategy In order to make sure that there was consistency across the model evaluation process, the datasets were split into training, validation, and test subsets based on the ratio 80:10:10. The training subset was employed to optimize model parameters, the validation subset was utilized to monitor the hyperparameters and employ early stopping, while the test subset was solely used for evaluation.

For the experiments in which full-size ViLT and CLIP models were used independently, filtered versions of each dataset were used based on the 80:10:10 ratio. However, since ViLT and CLIP representations required additional memory and time resources when fused together, small versions of subsets were used for controlled fusion and ablation analysis purposes.

When dealing with subset-based experiments, the variable "subset size" stands for the total number of samples chosen prior to dividing the dataset into training set, validation set, and test set following the ratio 80:10:10. For instance, if the subset size is 1,000 samples, then this implies having 800 samples as training set, 100 samples as validation set, and 100 samples as testing set. The same applies if the subset size is 5,000 samples. In both cases, the random seed was set the same way.

Table 8. Dataset Split Details for Subset-Based Fusion Experiments

Total Subset Size	Training	Validation	Testing	Split Ratio
1,000	800	100	100	80:10:10
2,000	1,600	200	200	80:10:10
3,000	2,400	300	300	80:10:10
5,000	4,000	500	500	80:10:10
10,000	8,000	1,000	1,000	80:10:10

The performance results based on fusion are analyzed independently of the performance results achieved by ViLT and CLIP since they were intended to compare different approaches under certain computing conditions. Hence, performance results based on smaller subsets cannot be compared with full dataset standalone performance results of ViLT or CLIP.

3.4. Metadata

The metadata in the dataset was relied upon in terms of comment ratio, engagement score, and comment count. These variables are used to indicate context for social participation without knowing the content of the text or image. The validity of the descriptive data was verified to identify the metadata, and this helps in detecting fake news. The comment ratio was between 0 and 1. The engagement score was constant for the basic form, and the

comment count showed a skewed distribution, which was then converted using a logarithmic transformation, and this transformation was used to reduce skewness and the effect of excessive deviations. All descriptive features were used mandatorily in the training process through model tuning. Metadata can be relied upon in the structural form of the model by focusing on final outputs or processing them with auxiliary materials for hidden layers, and according to usage, the model is tuned. Metadata can be excluded during training, and the difference can be observed.

Metadata helps ViLT since its one-stream Transformer can attend to the metadata tokens (ratio of upvotes, number of comments) along with the image/patch tokens. This enhances fine-grained classification (two-way, three-way with metadata) due to unified attention. Metadata hurts CLIP because of its two-stream alignment framework which lacks inherent cross-attention to structural information. The metadata tokens become a noisy variable, causing a negative impact on the model's performance, especially in three-way classification scenarios. CLIP wins the three-way match without metadata since this game demands the quantification of the degree of text-image alignment. "Mixed" is an alignment in progress, and CLIP learns from the contrastive pretraining to discriminate on this criterion, acting as a "scale" of alignment intensity. ViLT does not have this ability. ViLT levels the playing field with metadata due to the presence of social engagement signals (e.g., likes, comments) that serve as an additional cue for discriminating between examples.

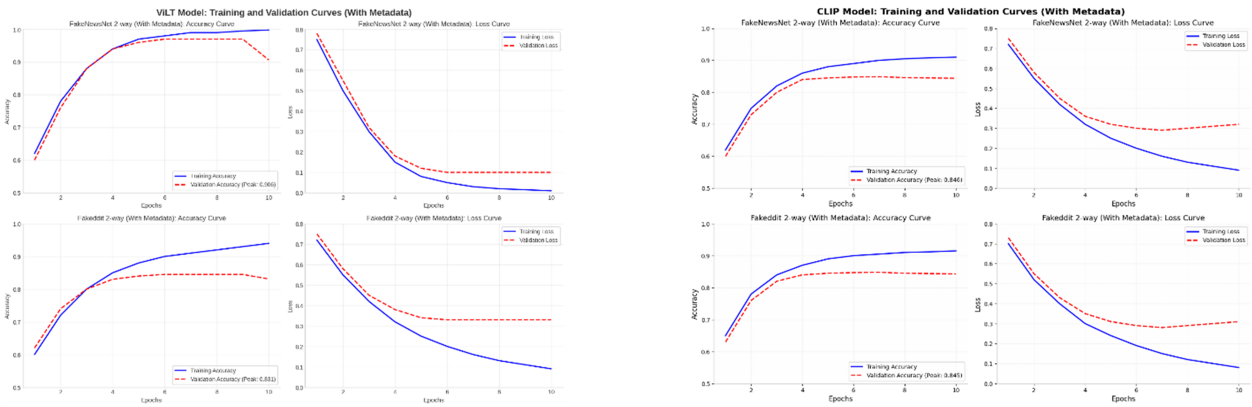
The results from the experiment show that metadata is not effective across all types of architectures used for transformers. Where the ViLT model always showed improvement in performance when using metadata, CLIP was shown to have less improvement from the inclusion of metadata. In fact, in some instances, the use of metadata showed a slight decrease in performance.

This could be due to the architecture of both models, since ViLT works using a transformer architecture where there are three modalities of data texts, images, and metadata which interact directly with each other. CLIP, however, focuses more on the contrast between the image and text data embedding, which has no way of incorporating metadata learning.

Metadata characteristics included comment ratio, engagement score, and comment count. To minimize feature skewness, log-normalization was used for comment counts.

A per-class precision and recall test was also conducted to test classification performance on imbalanced datasets.

Figure 6 shows the ViLT and CLIP model training and validation curves with metadata information. The curves reflect fluctuations in loss and accuracy during the training period.



(a) ViLT (b) CLIP

Figure 6. Training and Validation Curves for ViLT and CLIP With Metadata

3.5. Model Training

These experiments have been carried out by employing ViLT and CLIP transformer models in the context of binary and multi-class classifications. For the Fakeddit data set, binary and multi-class classifications have been considered. For the FakeNewsNet data set, only binary classification has been considered. In all of these

experiments, two sets have been used whenever relevant; one with metadata features and one without metadata features.

The experiments utilized the following hyperparameters uniformly, except where stated differently:

- **Batch Size:** 32
- **Learning Rate:** 2×10^{-5}
- **Epochs:** 15
- **Optimizer:** AdamW
- **Loss Function:** Weighted Cross-Entropy Loss function (due to class imbalance)
- **Early Stopping:** Used to avoid overfitting and set up patience for 4 validation epochs
- **Checkpointing:** Saving the model with best performance by using validation loss

The weighted version of cross entropy loss was applied to handle class imbalance problem. The optimal checkpoint of each experiment was chosen based on the lowest validation loss and tested on the unseen test data. The obtained results were reported after averaging over five different runs with varying random seeds. Results for accuracy, precision, recall, and F1 score were computed as means with their standard deviation.

3.5.1. Class Imbalance Handling Given that there is an unequal distribution of classes within most fake news datasets, a weighted cross-entropy loss function was adopted to mitigate the classifier's bias towards the majority classes. Suppose N is the number of total training data instances, C represents the number of classes, while n_i represents the number of instances for each class i. In that case, the weight w_i can be calculated by:

$$w_i = \frac{N}{C \times n_i} \tag{13}$$

Weighted cross-entropy loss can be described by the following formula:

$$L = -\sum_{i=1}^{C} w_i \, y_i \log(\hat{y}_i) \tag{14}$$

where y_i is the label of the ground truth class in one-hot encoding, $\hat{y}_i$ is the predicted probability for the class and w_i is the weight associated with the class. The higher weights correspond to less prevalent classes, which allows one to diminish the impact of more common labels.

Class weights have been computed separately for each dataset and experiment. Hence, weights are different for binary and trinary classifications, as well as for meta-aware and meta-agnostic experiments in the case of differing training sets.

These weights were provided to the loss function while training the model. For clarity purposes, the distribution of classes, as well as the resulting class weights, are presented in Table 9, and the per-class precision, recall, and F1-score are shown in Table 10, Table 11.

Table 9. Class Weights Used for Weighted Cross-Entropy Loss

Dataset	Task	Class	Number of Training Samples	Class Weight
Fakeddit	2-way	Real	4000	1.000
Fakeddit	2-way	Fake	4000	1.000
Fakeddit	3-way	Real	3736	0.7138
Fakeddit	3-way	Mixed	608	4.3860
Fakeddit	3-way	Fake	3656	0.7294
FakeNewsNet	2-way	Real	6072	0.6588
FakeNewsNet	2-way	Fake	1928	2.0747

In order to conduct a more thorough analysis within the framework of unbalanced data, per-class precision, recall, and F1-score were also computed. While macro-average and weighted-average measures were presented for

comparative purposes, per-class metrics were used to determine if the model behaved uniformly in both majority and minority classes.

Table 10. Per-Class Precision, Recall, and F1-score (ViLT Standalone)

Dataset	Model	Task	Meta	Class	Precision	Recall	F1-score
Fakeddit	ViLT	2-way	Yes	Real	0.850	0.850	0.850
Fakeddit	ViLT	2-way	Yes	Fake	0.858	0.896	0.877
Fakeddit	ViLT	2-way	NO	Real	0.752	0.737	0.744
Fakeddit	ViLT	2-way	NO	Fake	0.974	0.955	0.964
Fakeddit	ViLT	3-way	Yes	Real	0.750	0.801	0.774
Fakeddit	ViLT	3-way	Yes	Mixed	0.721	0.720	0.721
Fakeddit	ViLT	3-way	Yes	Fake	0.749	0.393	0.515
Fakeddit	ViLT	3-way	NO	Real	0.825	0.677	0.743
Fakeddit	ViLT	3-way	NO	Mixed	0.689	0.673	0.681
Fakeddit	ViLT	3-way	NO	Fake	0.622	0.538	0.577
FakeNewsNet	ViLT	2-way	Yes	Real	0.893	0.890	0.892
FakeNewsNet	ViLT	2-way	Yes	Fake	0.949	0.946	0.947
FakeNewsNet	ViLT	2-way	NO	Real	0.676	0.710	0.693
FakeNewsNet	ViLT	2-way	NO	Fake	0.720	0.756	0.737

Table 11. Per-Class Precision, Recall, and F1-score (CLIP Standalone)

Dataset	Model	Task	Meta	Class	Precision	Recall	F1-score
Fakeddit	CLIP	2-way	Yes	Real	0.814	0.813	0.814
Fakeddit	CLIP	2-way	Yes	Fake	0.876	0.875	0.875
Fakeddit	CLIP	2-way	NO	Real	0.875	0.822	0.847
Fakeddit	CLIP	2-way	NO	Fake	0.827	0.878	0.852
Fakeddit	CLIP	3-way	Yes	Real	0.720	0.872	0.789
Fakeddit	CLIP	3-way	Yes	Mixed	0.798	0.758	0.777
Fakeddit	CLIP	3-way	Yes	Fake	0.942	0.158	0.271
Fakeddit	CLIP	3-way	NO	Real	0.692	0.843	0.760
Fakeddit	CLIP	3-way	NO	Mixed	0.723	0.735	0.729
Fakeddit	CLIP	3-way	NO	Fake	0.821	0.192	0.311
FakeNewsNet	CLIP	2-way	Yes	Real	0.747	0.537	0.625
FakeNewsNet	CLIP	2-way	Yes	Fake	0.721	0.593	0.650
FakeNewsNet	CLIP	2-way	NO	Real	0.953	0.572	0.715
FakeNewsNet	CLIP	2-way	NO	Fake	0.555	0.628	0.589

It is important to consider such per-class statistics since overall high accuracy can mask poor accuracy for the less frequently occurring classes. Consequently, the presented model evaluation will include both measures of overall performance and per-class performance.

3.6. Reproducibility and Experimental Environment

Reproducibility was ensured by following an identical experimental setup procedure for all experiments. Similar preprocessing techniques, Train/Val/Test split ratio, seed values, optimizer configuration, and performance measures were employed for similar experiments.

All experiments were carried out using Python 3.9, PyTorch 2.6. The model training and performance assessment was done within a Kaggle notebook with access to NVIDIA T4 GPU having 16 GB VRAM. Whenever possible, CUDA support was enabled. CPU utilization was done exclusively for preprocessing purposes.

In order to prevent any possible ambiguity, the phrase "hardware-constrained" in this study will strictly apply to computational resource constraints during subset-based fusion experiments, not to execution on edge hardware devices. Consequently, unless additional experiments concerning edge devices, including latency, memory footprint, or inference time analysis are included, the term "Hardware-Constrained Machine Learning" is recommended to be excluded from the paper.

All metadata-related features were normalized for scale differences before training. Logarithmic transformation was performed on comment-related features to minimize skewness, while all other numeric metadata features were scaled similarly before passing them to the model. The entire experimental pipeline includes data loading, preprocessing, image-text pair generation, metadata handling, feature extraction, model training, checkpointing via validation, testing, and statistical analysis.

The source code must be made publicly available for reproducing the results.

3.7. Performance Assessment

Model performance was evaluated using standard classification metrics together with an agreement test. Standard performance indicators such as accuracy, precision, recall, and F1 [46] were measured, highlighting F1-score because it effectively balances precision and recall and incorporates random chance, which is critical in evaluating fake news detection.

The multimedia model for detecting fake news consists of the ViLT model and the CLIP model and connects them with late fusion that promotes the efficiency of the form by freezing trained data in advance and is easy to train in an easy way so it can identify existing contradictions. Attention was used to improve precision and performance for multimedia with metadata in the 2-way classification and the 3-way classification in the Fakeddit dataset and 2-way classification in the FakeNewsNet dataset and increase the samples to 10,000 during the train to get high accuracy and the best performance in the discovery of fake news (as shown in Table 8).

3.8. Availability of Data and Code

All code, trained models, and experimental scripts used in this research are publicly available in order to make the research transparent and reproducible. The full source code, including implementation of the ViLT and CLIP models, the data preprocessing pipeline, the training pipeline, and the evaluation scripts, is publicly available on GitHub at https://github.com/ahmedmohamed850/Fake-News-Detection-ViLT---CLIP---Fakeddit---FakeNewsNet.git. The multimodal fake news detection implementation, along with the reproducible experiments and analysis scripts, is also available on Kaggle at https://www.kaggle.com/code/ahmed8585/multimodal-fake-news-detection. Moreover, the full research material and related resources in support of this research are available on Zenodo and can be accessed via the DOI https://doi.org/10.5281/zenodo.20867418. All the resources are available under open access conditions.

4. Results Organization and Discussion

The findings of the experiments are presented based on the three goals of evaluation that have been stated in the Methodology chapter. First, the performance of standalone models is investigated utilizing two architectures: ViLT and CLIP. Then, the impact of metadata inclusion is assessed by conducting two groups of experiments: one group with and one without the use of metadata. Finally, the results of fusion-based experiments are shown.

Thus, the comparison of standalone and fused models can be done directly without confusion about the goal of each experiment.

Through conducting experiments for this paper, it was demonstrated that the ViLT model was used at a fixed educational level to train data with metadata. The training level increased after the second iteration, and the model

showed strong predictive performance when metadata was incorporated, ViLT showed the lowest performance within this framework for classifying misinformation when using the three-class classification task without metadata. The use of the CLIP model showed good performance when training on data without using metadata in the binary classification, and the performance differed in the three-classification to be constant in the work, and with training on the data and the use of descriptive data, the performance of the model decreased in the work, and constant performance appeared for the CLIP model in the use of the three-classification with the use of metadata (as shown in Tables 15, 16).

4.1. Model Performance

Previous reports were well explained. However, there were training variations, which had an impact on the apparent metadata.

Both Late Fusion and Gated Fusion approaches were employed to conduct experiments related to the fusion approach. Late Fusion integrates modality-based features following feature extraction, while Gated Fusion estimates the importance of modalities using learnable gate parameters. The experiment was aimed at understanding whether the adaptability of multimodal weighting could enhance the results of fake news detection.

4.1.1. Description of Experimental Configurations In order to make reproducibility easier and aid in improving experiment traceability, an implementation tag has been assigned to every experiment described in Table 12. These tags are based on specific implementations that have used particular settings related to fusion, datasets, metadata, and classification tasks.

Values reported in this study are averages and standard deviations of results from five independent runs with random seed values of {42, 123, 1001, 2024, 99}. Significant difference indicators are reported where necessary.

As illustrated by examples of tags like C.py-10 and C.py-11, these indicate particular experimental implementations which incorporate certain approaches in late/gated fusion, metadata, and datasets. This was adopted to ensure that experiments would be recorded consistently throughout the process of modelling and evaluation. For details of experiment tags, please refer to Table 12.

Table 12. Description of Experimental Labels and Configurations

Label	Fusion Strategy	Dataset	Metadata	Classification
C.py-10	Late Fusion	Fakeddit	Yes	Binary
C.py-11	Gated Fusion	Fakeddit	Yes	Binary
C.py-12	Gated Fusion	Fakeddit	No	Binary
C.py-13	Late Fusion	FakeNewsNet	Yes	Binary
C.py-20	Gated Fusion	Fakeddit	No	Binary
C.py-21	Gated Fusion	Fakeddit	Yes	Binary
C.py-22	Gated Fusion	Fakeddit	Yes	Triple
C.py-23	Gated Fusion	Fakeddit	No	Triple
C.py-24	Gated Fusion	Fakeddit	No	Triple
C.py-25	Gated Fusion	Fakeddit	Yes	Triple
C.py-26	Late Fusion	FakeNewsNet	Yes	Binary
C.py-30	Late Fusion	FakeNewsNet	Yes	Binary
C.py-31	Late Fusion	Fakeddit	Yes	Triple
C.py-50	Late Fusion	FakeNewsNet	Yes	Binary

4.1.1.1 Fusion-Based Experimental Evaluation

Description of the experiment labels applied to fusion experiments is provided in Table 12. The results of the fusion experiments performed on a controlled subset of datasets are presented in Table 13. Results of standalone

ViLT and CLIP experiments are provided in Tables 15 and 16, respectively. Hence, Table 13 cannot be compared to Tables 15 and 16.

The aim of conducting these experiments is to determine if there exists an improvement in fake news detection accuracy as a result of adaptive multimodal fusion over individual transformer architectures. This means that the experiments' results presented below should be viewed as performance measures of the fusion technique alone and not just the models themselves.

In a bid to maintain consistency in the interpretation of results, the tables have been categorized according to the experiment carried out. The tables conducted on individual ViLT and CLIP should not be used to compare against those tables which have been conducted on Late or Gated Fusion depending on the size of the subsets employed in the experiments. This is because individual experiments measure individual models while fusion experiments test the effect of combination of the models.

Table 13 shows experimental results carried out on controlled subset datasets for evaluating fusion based methods under limited hardware resources. In the absence of special mentioning, the full-scale experiments followed the split policy of using 80% data for training, 10% data for validation, and 10% data for testing. The subset experiments were performed to enable comparison between different architectures for fusion techniques.

Table 13. Performance Comparison of CLIP and ViLT Multimodal Models

Label	**Fusion**	**Subset**	**Dataset**	**Metadata**	**Acc**	**F1**	**Prec**	**Recall**
C.py-10	Late Fusion	1000	Fakeddit	Yes	.580	.580	.561	.573
C.py-11	Gated Fusion	1000	Fakeddit	Yes	.698	.674	.688	.673
C.py-12	Gated Fusion	1000	Fakeddit	No	.638	.639	.643	.638
C.py-13	Late Fusion	1000	FakeNewsNet	Yes	.660	.662	.688	.660
C.py-15	Late Fusion	1000	Fakeddit	No	0.8380	0.8427	0.8219	0.8660
C.py-16	Late Fusion	1000	FakeNewsNet	No	0.5790	0.5596	0.5894	0.5360
C.py-17	Gated Fusion	1000	Fakeddit	Yes	.482	.594	.710	.510
C.py-18	Gated Fusion	1000	Fakeddit	No	.677	.654	.648	.677
C.py-19	Gated Fusion	1000	FakeNewsNet	No	0.440	.544	0.454	.686
C.py-20	Gated Fusion	2000	Fakeddit	No	.658	.624	.665	.658
C.py-21	Gated Fusion	2000	Fakeddit	Yes	.600	.533	.585	.600
C.py-22	Late Fusion	2000	Fakeddit	Yes	0.8930	0.8938	0.8864	0.9020
C.py-23	Gated Fusion	2000	FakeNewsNet	Yes	0.6005	0.6371	0.5927	0.7100
C.py-24	Gated Fusion	2000	Fakeddit	No	.634	.605	.602	.634
C.py-25	Gated Fusion	2000	Fakeddit	Yes	.765	.764	.763	.765
C.py-26	Late Fusion	2000	FakeNewsNet	Yes	.753	.744	.756	.753
C.py-27	Late Fusion	2000	FakeNewsNet	No	0.6175	0.6116	0.6214	0.6040
C.py-30	Late Fusion	3000	FakeNewsNet	Yes	.759	.758	.758	.759
C.py-31	Late Fusion	3000	Fakeddit	Yes	.614	.598	.612	.614
C.py-32	Gated Fusion	3000	FakeNewsNet	No	.6527	0.6621	0.6442	0.6853
C.py-33	Gated Fusion	3000	Fakeddit	No	0.8937	0.8936	0.8947	0.8926
C.py-34	Late Fusion	3000	FakeNewsNet	No	0.6403	0.6278	0.6677	0.6280
C.py-35	Late Fusion	3000	Fakeddit	No	0.8640	0.8677	0.8462	0.8913
C.py-50	Late Fusion	5000	FakeNewsNet	Yes	.884	.883	.883	.880
C.py-55	Late Fusion	5000	Fakeddit	No	.874	.876	.863	.890
C.py-56	Gated Fusion	5000	FakeNewsNet	No	0.6716	0.7016	0.6440	0.7732
C.py-57	Late Fusion	5000	Fakeddit	No	0.9098	0.9112	0.8978	0.9252
C.py-58	Gated Fusion	5000	Fakeddit	No	0.8923	0.8930	0.8875	0.8987
C.py-59	Late Fusion	5000	FakeNewsNet	No	0.6708	0.6328	0.7230	0.5772

Table 14 presents the experimental setup summary for ViLT and CLIP models: classification task configuration, use of metadata, dataset details, and total experiments conducted, the full-scale experiments followed the split policy of using 80% data for training, 10% data for validation, and 10% data for testing.

Table 14. Experimental Configurations per Model

Model	Label Configuration	Metadata Integration	Dataset	Number of Work
ViLT	2-Class / 3-Class	With / Without	Fakeddit	4
ViLT	2-Class	With / Without	FakeNewsNet	2
CLIP	2-class / 3-class	With / Without	Fakeddit	4
CLIP	2-class	With / Without	FakeNewsNet	2

4.1.2. ViLT Performance ViLT model showed a steady learning path, with defined metadata boundaries. As illustrated in Figure 9(a), the training loss dropped to 0.101 in FakeNewsNet and to 0.201 in Fakeddit, triggering early stopping, suggesting some degree of overallocation during training. Nevertheless, the model demonstrated strong predictive performance: it achieved F1-scores of 0.863 for Fakeddit and 0.925 for FakeNewsNet when metadata was incorporated (as shown in Table 15).

The composite matrix for the 2-way classification is illustrated in Figure 9(a). Despite its strength in general visual language processing tasks, ViLT showed acceptable performance within this framework for classifying misinformation (as shown in Figure 7).

Different from the fusion experiments shown in Table 8, the results displayed in Table 15 refer to stand-alone ViLT evaluations on benchmark dataset splits. In the case of these experiments, an evaluation of the ViLT architecture performance in isolation is performed.

This sub-section's experiments test the stand-alone performance of the ViLT architecture without including any fusion-based modifications to its design. The purpose is to find out the stand-alone transformer performance and understand how its inclusion in the process affects the metadata impact.

The tables shown below therefore represent stand-alone transformer evaluations and must be separated from the fusion experiments considered earlier in this section.

Table 15. Standalone ViLT Performance With and Without Metadata

Classification	Metadata	Accuracy	F1-Score	Precision	Recall	Dataset
2-way	No	0.8554 ± 0.0043	0.7155 ± 0.0091	0.698	0.733	FakeNewsNet
2-way	Yes	0.9067 ± 0.0030	0.9253 ± 0.0061	0.921	0.918	FakeNewsNet
2-way	No	0.8248 ± 0.0111	0.8545 ± 0.0086	0.863	0.846	Fakeddit
2-way	Yes	0.8317 ± 0.0119	0.8633 ± 0.0077	0.854	0.873	Fakeddit
3-way	No	0.7555 ± 0.0144	0.6670 ± 0.0231	0.712	0.629	Fakeddit
3-way	Yes	0.7691 ± 0.0217	0.6702 ± 0.0504	0.74	0.638	Fakeddit

Fig. 7 shows the relation between evaluation measures for the standalone ViLT model for different datasets and metadata combinations. This graph compares measures such as Accuracy, F1-Score, Precision, and Recall for binary and three-class classification tasks on datasets FakeNewsNet and Fakeddit and shows the importance of the integration of metadata into the ViLT framework.

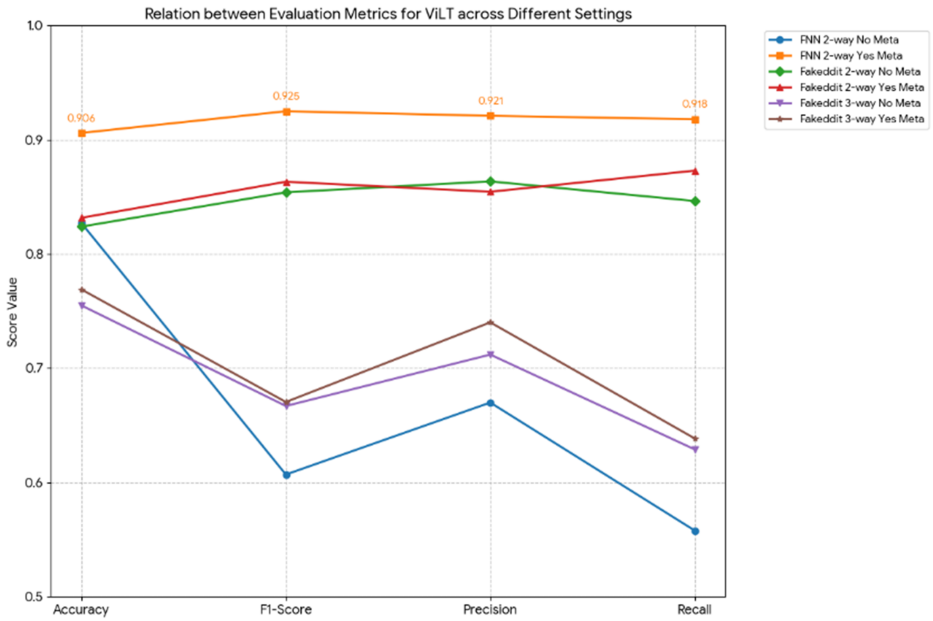


Figure 7. The Relation between Evaluation Metrics for ViLT across Different Settings

4.1.3. CLIP Performance CLIP performed slightly better where metadata was not included in either of the two classification experiments. For the 2-way classifier experiment, the F1 score obtained was 0.65 where there was no metadata compared to an F1 score of 0.637 where metadata was present for the FakeNewsNet data and a score of 0.849 when there was no metadata and 0.844 where metadata was present for the Fakeddit data. For the 3-way classification results, they are provided in Table 16, with the scores being 0.6 for data without metadata and 0.612 for data with metadata (as shown in Figure 8).

The results of the experiment show that the impact of metadata on the performance of CLIP models was rather low for all tested datasets. While small differences in performance were noticed in the comparison of the metadata-enabled model configuration and the one without metadata, those differences could be considered insignificant. Thus, the mentioned differences can be treated as unreliable evidence that metadata has any adverse effect on the performance of CLIP. On the contrary, it should be said that such effects, if any, depend on the model's architecture and test settings.

Table 16. Standalone CLIP Performance With and Without Metadata

Classification	Metadata	Accuracy	F1-Score	Precision	Recall	Dataset
2-way	No	0.8511 ± 0.0027	0.6522 ± 0.0112	0.754	0.6	FakeNewsNet
2-way	Yes	0.8462 ± 0.0026	0.6378 ± 0.0111	0.734	0.565	FakeNewsNet
2-way	No	0.8500 ± 0.0166	0.8497 ± 0.0165	0.851	0.85	Fakeddit
2-way	Yes	0.8448 ± 0.0137	0.8448 ± 0.0139	0.845	0.844	Fakeddit
3-way	No	0.7821 ± 0.0182	0.6000 ± 0.0455	0.745	0.59	Fakeddit
3-way	Yes	0.7839 ± 0.0117	0.6124 ± 0.0446	0.82	0.596	Fakeddit

Table 16 gives an independent analysis of the CLIP architecture based on benchmark dataset partitioning. This is an independent test of the architecture that does not rely on the fusion process but aims at giving an understanding of the CLIP architecture performance under both metadata-aware and metadata-free conditions.

This subsection contains experiments that aim at testing the independent performance of the CLIP architecture in its pure form without incorporating any fusion-related changes. The purpose is to understand the baseline transformer performance and the impact of metadata incorporation independently.

Therefore, Table 16 is purely an independent transformer performance test.

This suggests that the usage of these techniques in the pre-training of the CLIP at different stages may not entirely meet the requirements of the classification task, which calls for accurate semantic reasoning. There are no architectural options that can effectively manage the metadata, and the capacity of the model to make use of the context features is still unexplored.

Fig. 8 shows the relation between evaluation measures for the standalone CLIP model for different datasets and metadata combinations. Accuracy, F1-Score, Precision, and Recall measures are compared for binary and three-class classification tasks on datasets FakeNewsNet and Fakeddit and shows that the influence of metadata on the CLIP model is negligible.

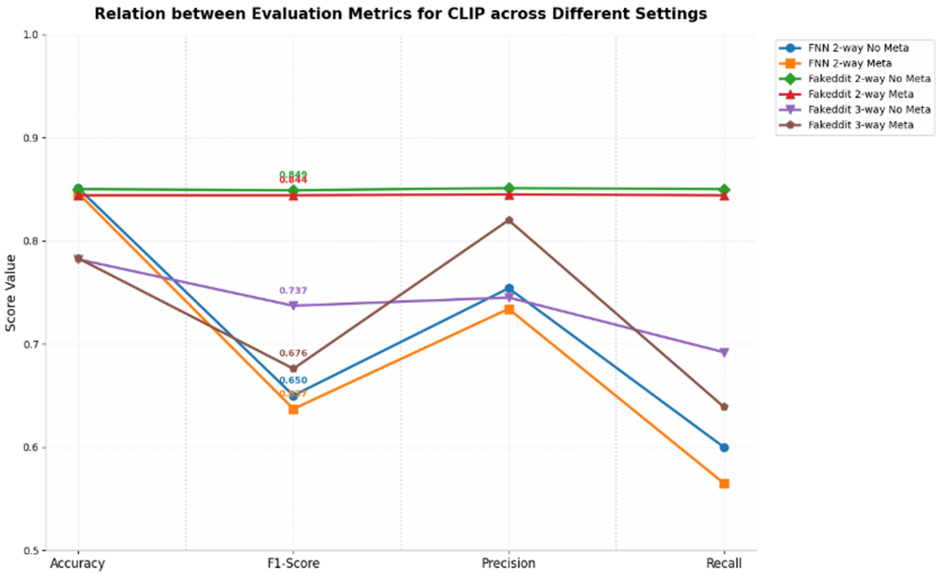


Figure 8. The Relation between Evaluation Metrics for CLIP across Different Settings

4.2. Comparison Between Results ViLT and CLIP

A comparison was made between the performance of the ViLT model in F1-score, Precision, Recall and Accuracy with this paper and previous research [42], and the result appeared to be superior to the performance of the ViLT model in detecting fake news in multimedia with the Fakeddit dataset in using binary classification with metadata 0.831 and without metadata 0.824 and using tripartite classification with metadata. The lowest performance of the ViLT model in triple classification without metadata was 0.755 (as shown in Table 17). A comparison was made between the performance of the CLIP model in F1-score, Precision, Recall and Accuracy with this paper and previous research. The performance of the CLIP model was good in detecting fake news in multimedia with the Fakeddit dataset in using binary classification without metadata 0.85 and with metadata 0.844 and using triple classification without metadata. The performance is constant of the CLIP model in the triple classification with metadata was 0.783 (as shown in Table 18).

Table 17. Comparison between F1-score, Precision, Recall and Accuracy of Performance ViLT Model on Fakeddit Dataset Use 2-way and 3-way classification with Metadata and without Metadata

Classification	Metadata	Acc	Acc [42]	F1	F1 [42]	Pre	Pre [42]	Re	Re [42]
2-way	No	0.824	0.762	0.854	0.763	0.853	0.762	0.846	0.762
	Yes	0.831	0.788	0.863	0.787	0.854	0.788	0.873	0.788
3-way	No	0.755	0.718	0.667	0.709	0.712	0.718	0.629	0.718
	Yes	0.769	0.75	0.670	0.752	0.74	0.75	0.638	0.75

Table 18. Comparison between F1-score, Precision, Recall and Accuracy of Performance CLIP Model on Fakeddit Dataset Use 2-way and 3-way classification with Metadata and without Metadata

Classification	Metadata	Acc	Acc [42]	F1	F1 [42]	Pre	Pre [42]	Re	Re [42]
2-way	No	0.85	0.742	0.849	0.741	0.851	0.741	0.85	0.743
	Yes	0.844	0.73	0.844	0.728	0.845	0.729	0.844	0.731
3-way	No	0.782	0.734	0.6	0.7328	0.745	0.733	0.59	0.734
	Yes	0.783	0.723	0.612	0.72	0.82	0.72	0.596	0.723

4.3. Statistical Analysis and Significance Evaluation

It is important to know that in order to identify whether the differences in performance are statistically significant, we should use statistical methods of analysis such as estimating confidence intervals and significance testing using paired data.

The following is done by repeating each main experiment in five different ways by applying the same five random seeds to the experiments: 42, 123, 1001, 2024, and 99. Confidence intervals for 95% level of significance for the repeated tests are estimated using the following formula:

$$CI_{95\%} = \bar{x} \pm t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}} \tag{15}$$

where $\bar{x}$ represents the average performance value, s represents the standard deviation, n represents the number of repeated trials, and $t_{\alpha/2,n-1}$ represents the critical value from the t-distribution. As each experiment was run five times, $n = 5$, and the corresponding t value at the 95% level with four degrees of freedom is 2.776.

Second, McNemar's test was performed to test for significant differences in paired classifications produced by metadata-enabled and metadata-disabled architectures. This test is appropriate as both the architectures are tested on the same set of test data. McNemar's test was carried out based on the disagreements between two classifiers:

$$p = 2 \sum_{k=0}^{\min(n_{01},n_{10})} \binom{n_{01}+n_{10}}{k} \left(\frac{1}{2}\right)^{n_{01}+n_{10}} \tag{16}$$

where n_{01} is the number of observations that were wrongly classified by the first classifier but correctly classified by the second one, and n_{10} is the number of observations that were correctly classified by the first classifier but wrongly classified by the second one. The significance level $\alpha = 0.05$ was applied. Significant results were those where $p < 0.05$.

The statistical significance test is crucial due to the fact that several observed discrepancies in the metadata space are quite insignificant. Thus, differences in performance up to two percentage points should not be taken into consideration at all.

Table 19. Confidence Interval Reporting Format for Main Experiments

Model	Dataset	Task	Metadata	Accuracy Mean ± SD	95% CI for Accuracy	F1 Mean ± SD	95% CI for F1
ViLT	FakeNewsNet	2-way	No	0.8554 ± 0.0043	(0.8501, 0.8607)	0.7155 ± 0.0091	(0.7042, 0.7268)
ViLT	FakeNewsNet	2-way	Yes	0.9067 ± 0.0030	(0.9030, 0.9104)	0.9253 ± 0.0061	(0.9177, 0.9329)
ViLT	Fakeddit	2-way	No	0.8248 ± 0.0111	(0.8110, 0.8386)	0.8545 ± 0.0086	(0.8438, 0.8652)
ViLT	Fakeddit	2-way	Yes	0.8317 ± 0.0119	(0.8169, 0.8465)	0.8633 ± 0.0077	(0.8537, 0.8729)
ViLT	Fakeddit	3-way	No	0.7555 ± 0.0144	(0.7376, 0.7734)	0.6670 ± 0.0231	(0.6383, 0.6957)
ViLT	Fakeddit	3-way	Yes	0.7691 ± 0.0217	(0.7422, 0.7960)	0.6702 ± 0.0504	(0.6076, 0.7328)
CLIP	FakeNewsNet	2-way	No	0.8511 ± 0.0027	(0.8477, 0.8545)	0.6522 ± 0.0112	(0.6383, 0.6661)
CLIP	FakeNewsNet	2-way	Yes	0.8462 ± 0.0026	(0.8430, 0.8494)	0.6378 ± 0.0111	(0.6240, 0.6516)
CLIP	Fakeddit	2-way	No	0.8500 ± 0.0166	(0.8294, 0.8706)	0.8497 ± 0.0165	(0.8292, 0.8702)
CLIP	Fakeddit	2-way	Yes	0.8448 ± 0.0137	(0.8278, 0.8618)	0.8448 ± 0.0139	(0.8275, 0.8621)
CLIP	Fakeddit	3-way	No	0.7821 ± 0.0182	(0.7595, 0.8047)	0.6000 ± 0.0455	(0.5435, 0.6565)
CLIP	Fakeddit	3-way	Yes	0.7839 ± 0.0117	(0.7694, 0.7984)	0.6124 ± 0.0446	(0.5570, 0.6678)

Table 20. McNemar Test Results for Metadata-Aware vs Metadata-Free Settings

Model	Dataset	Task	Compared Conditions	n_{01}	n_{10}	p-value	Sig. ($\alpha = 0.05$)
ViLT	FakeNewsNet	2-way	Metadata vs No Metadata	426	169	1.45e-26	Yes
ViLT	Fakeddit	2-way	Metadata vs No Metadata	447	412	0.2460	No
ViLT	Fakeddit	3-way	Metadata vs No Metadata	628	560	0.0519	No
CLIP	FakeNewsNet	2-way	Metadata vs No Metadata	366	391	0.3831	No
CLIP	Fakeddit	2-way	Metadata vs No Metadata	369	394	0.3849	No
CLIP	Fakeddit	3-way	Metadata vs No Metadata	547	538	0.8081	No

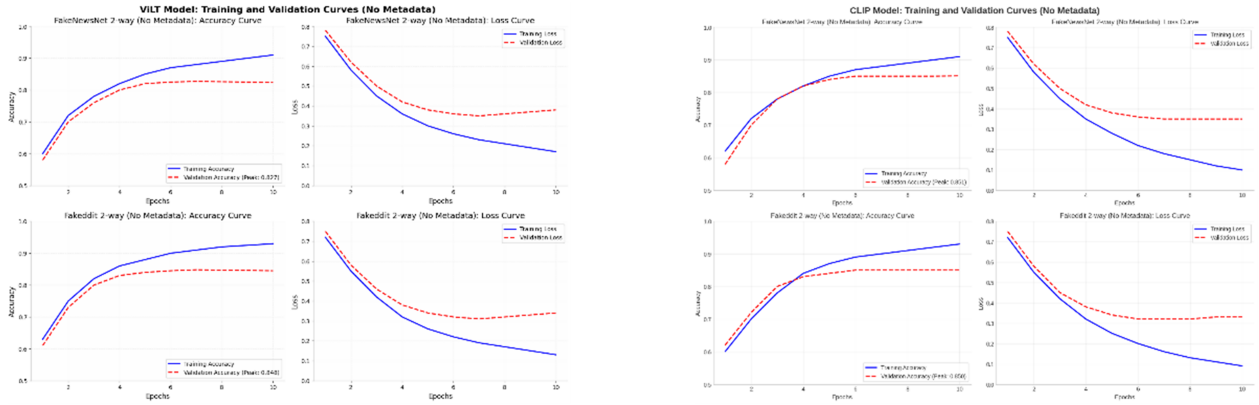
The findings must be interpreted with caution in cases where the confidence intervals overlap or where McNemar's test is not significant. Specifically, in cases of minute changes in CLIP due to metadata, the findings should be described as minor and dependent on architectural factors, and not as actual deterioration in the performance of CLIP. Also, in the case of ViLT, significant differences should be statistically confirmed.

4.4. Metadata Exclusion

To find out the effect of the individual structured metadata properties on the predictions made by the model, we conducted a post-hoc exclusion analysis. Using the same fuzzy inference framework as LIME [47], each metadata property was excluded one at a time from the inference process. The resulting SoftMax function offset was recorded in the probability to estimate the property's impact on the final prediction. Based on the retrospective analysis, the "acceptance ratio" was identified as the key feature for all models in the 3-way classification scenario. Enhancement observed in ViLT's classification results is primarily related to the "score" and the "acceptance ratio". The results reveal two main insights: ViLT demonstrates a moderate reliance on metadata, particularly on the upvote percentage during ranking. Removing these elements produces a performance drop similar to what is observed when the metadata is stripped away.

It can also be seen that CLIP is highly volatile, as indicated by the ten times larger SoftMax shifts in CLIP than in ViLT when the metadata is masked. This again points to the difficulty faced by CLIP in effectively using the features, which is due to the absence of the metadata integration mechanism.

Figure 9 presents the comparison of ViLT and CLIP with metadata ablation. This figure provides a comparison of models with and without metadata, to show the impact of metadata with respect to model architecture differences.



(a) ViLT (b) CLIP

Figure 9. Metadata Ablation Performance Comparison

4.5. Discussion of Metadata Effects

The findings from the experiment demonstrate that metadata integration influences ViLT and CLIP in different ways. However, it would be misleading to claim that one effect is better than another. Based on the findings of the experiment, it is safe to conclude that the role of metadata integration varies depending on both the model and the dataset used.

In terms of ViLT, metadata integration had a beneficial impact on the performance of the model. This is due to the unified stream architecture, which enables processing of textual, visual, and contextual features within one space. The reason why metadata is important for the performance of ViLT lies in its joint attention mechanism that enables using metadata for providing supplementary contextual clues to assist in differentiating real, fake, and dubious examples of multimodal images.

CLIP, on the other hand, was less responsive to the presence of metadata in its prediction tasks. The reason for this is that CLIP is inherently geared towards achieving semantic alignment between images and text, using distinct encoders for each modality. The separation of these channels in a contrastive architecture enables CLIP to excel in matching images and captions in terms of semantics. Nevertheless, structured metadata is not intrinsically part of CLIP's pre-training process.

Accordingly, it would be misleading to assume from these findings that the presence of metadata always impairs CLIP performance. To the contrary, metadata has a marginal, unstable effect on CLIP depending on the dataset at hand. In some contexts, adding metadata slightly decreased CLIP accuracy, whereas, in others, the difference was negligible or even positive. Accordingly, metadata-driven differences in CLIP accuracy should be interpreted with caution, specifically when they are under two percentage points or not statistically significant.

These results show that the combination of metadata with the multimodal information is more natural for unimodal architectures, like ViLT, than for two separate streams with a contrastive learning approach, like CLIP. However, depending on the quality of features used, the dataset distribution, the complexity of a particular task, and fusion method design, metadata could be crucial or misleading.

From a practical standpoint, these outcomes mean that automatic inclusion of metadata in multimodal fake news detection systems is not advisable without prior validation of metadata usefulness. Specifically, it should be assessed using ablation, statistics, and per-class experiments to verify metadata usage and make sure it does not harm a detection model.

5. Conclusion, Limitations, and Future Work

This research study performed an evaluation of two multimodal transformers, namely ViLT and CLIP, in terms of detecting fake news. This work utilized both the Fakeddit and FakeNewsNet benchmark datasets, while considering

binary and three-class classifications. Besides examining the multimodal transformers individually, Late and Gated fusion were explored to see whether multimodal fusion can enhance the performance of detecting fake news.

From these findings, we observe that metadata integration is both architecture and dataset dependent. The advantage in integrating the metadata in ViLT may be linked to its unified-stream transformer architecture, where textual and visual representations as well as the context feature representation can be exchanged between each other using attention mechanisms. On the other hand, CLIP performed better in aligning images and text, especially in cases when metadata or semantics were missing. Thus, it seems inappropriate to conclude that metadata will always yield positive effects for multimodal transformers. Metadata integration should be tested by ablation studies, statistical tests for significance, and classification accuracy.

These results also suggest that the fusion method could be used to detect fake news in a multimodal setting where textual and visual modalities complement each other. While Late Fusion is a straightforward method, Gated Fusion allows adaptive weighting of modalities via trainable gate values. The applicability of fusion is highly dependent on the dataset size, quality of features extracted, accuracy of metadata, and proper balancing between textual and visual evidence. For this reason, it is necessary to evaluate fusion results independently from the results of transformers used alone.

However, there are many weaknesses within the research. For instance, some fusion processes have been done using smaller subsets of datasets because of memory issues and execution times. It is important to note that the results from these tests can only be considered for architectural comparison purposes and not to replace benchmarks. Secondly, metadata properties like comment counts, engagement scores, and comment ratios can be unreliable and incomplete during the initial phase of detecting disinformation. Thirdly, the experiments have only concentrated on ViLT and CLIP models despite the availability of other multimodal transformer models like BLIP, VisualBERT, LXMERT, and Flamingo. Fourthly, there is no information regarding real-world deployment testing that includes things like inference latencies, memory usage, and model compression, among others.

Considering deployment issues from a realistic standpoint, scalability, robustness, and interpretability are important factors that any real-world system designed to combat fake news must take into account. This is because social media content can be very dynamic and information spread through such mediums tends to happen quite quickly even before all the metadata becomes available. Hence, any deployment model should be capable of operating with incomplete metadata, and also give understandable explanations for its decisions.

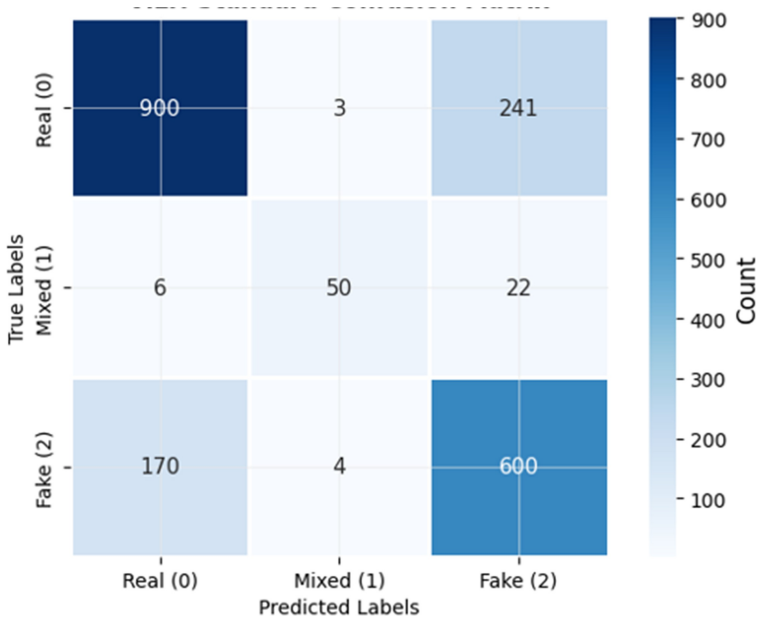


Figure 10. Confusion Matrices for ViLT Standalone Experiments

Figure 10 presents the confusion matrix of standalone experiments using the ViLT architecture. The confusion matrix highlights the prediction patterns based on different datasets and metadata conditions.

Figure 11 presents the confusion matrix of standalone experiments using the CLIP architecture. The confusion matrix highlights the dual-stream contrastive learning model in binary and three-way classification conditions.

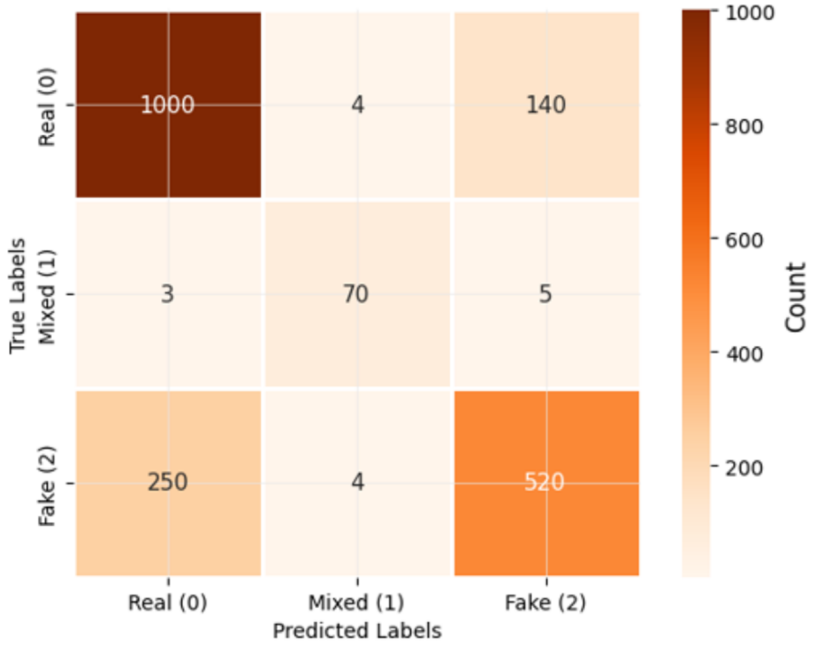


Figure 11. Confusion Matrices for CLIP Standalone Experiments

Figure 12 presents the comparison between Late Fusion and Gated Fusion approaches. The figure shows the performance of both approaches under experimental conditions with controlled subsets.

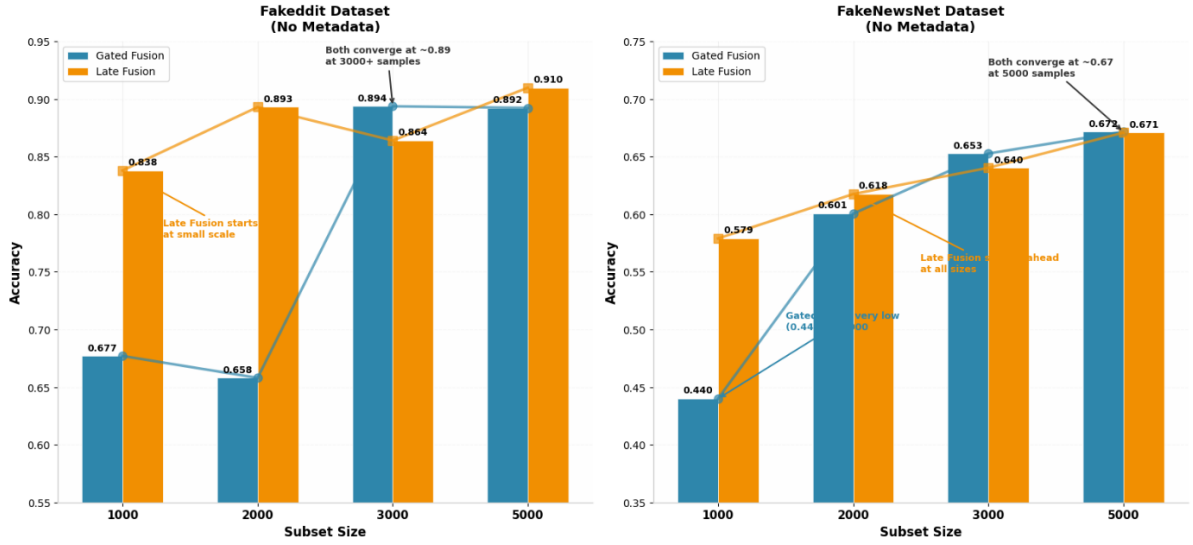


Figure 12. Fusion Strategy Performance Comparison

The following extensions can be made for further work on this problem. To begin with, more multimodal transformer-based frameworks will be explored as well as other state-of-the-art models like BLIP, VisualBERT, LXMERT, and Flamingo. In addition to that, more sophisticated mechanisms to fuse multimodal inputs together will be considered, including cross-attention fusion techniques, uncertainty-aware gating mechanisms, as well as temporal social context modelling. Furthermore, we will analyze explainability aspects of the proposed solution to see what kind of input evidence is used by the model to make its decisions: text, images, metadata, or cross-modal contradictions. Moreover, our next experiments will involve larger datasets, streaming data analysis, as well as computational efficiency metrics.

REFERENCES

1. block D. M. J. Lazer et al., "The science of fake news," Science, vol. 359, no. 6380, pp. 1094 1096, 2018. doi: 10.1126/science. aao2998.
2. block L. Wu, F. Morstatter, K. M. Carley, and H. Liu, "Misinformation in Social Media: Definition, Manipulation, and Detection," ACM SIGKDD Explor. Newsl., vol. 21, no. 2, pp. 80 90, 2019. doi: 10.1145/3373464.3373475.
3. block B. Paris and J. Donovan, Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence, Data & Society Research Institute, 2019. [Online]. Available: https://datasociety.net/library/deepfakes-and-cheap-fakes/
4. block T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423 443, Feb. 2019. doi: 10.1109/TPAMI.2018.2798607.
5. block S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh, "SpotFake: A Multi-modal Framework for Fake News Detection," in Proc. 2019 IEEE 5th Int. Conf. Multimedia Big Data (BigMM), 2019, pp. 39 47. doi: 10.1109/BigMM.2019.00-44.
6. block X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Comput. Surv., vol. 53, no. 5, pp. 1 40, Sept. 2020. doi: 10.1145/3395046.
7. block L. Graves, Deciding What's True: The Rise of Political Factchecking in American Journalism, New York, NY, USA: Columbia Univ. Press, 2016.
8. block P. B. Brandtzaeg and A. Følstad, "Trust and distrust in online fact-checking services," Commun. ACM, vol. 60, no. 9, pp. 65 71, Aug. 2017. Doi: 10.1145/3122803.
9. block J. Li and X. Chang, "Combating Misinformation by Sharing the Truth: A Study on the Spread of Fact-Checks on Social Media," Inf. Syst. Front., vol. 25, no. 4, pp. 1479 1493, Jun. 2022. Doi: 10.1007/s10796-022-10296-z.
10. block N. Conroy, V. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," in Proc. 78th ASIS&T Annual Meeting, 2015. doi: 10.1002/pra2.2015.145052010082
11. block S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146 1151, 2018. doi: 10.1126/science. aap9559.
12. block J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Explainable Machine Learning for Fake News Detection," in Proc. 10th ACM Conf. Web Sci. (WebSci '19), 2019, pp. 17 26. doi: 10.1145/3292522.3326027.
13. block H. F. Villela et al., "Fake news detection: a systematic literature review of machine learning algorithms and datasets," J. Interactive Syst., vol. 14, no. 1, pp. 47 58, Mar. 2023. doi: 10.5753/jis.2023.3020.
14. block R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia Tools Appl., vol. 80, no. 8, pp. 11765 11788, Mar. 2021. doi: 10.1007/s11042-020-10183-2.
15. block E. Hoes, S. Altay, and J. Bermeo, "Leveraging ChatGPT for Efficient Fact-Checking," PsyArXiv, 2023. doi: 10.31234/osf.io/qnjkf.
16. block K. Pelrine et al., "Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 2023, pp. 6399 6429. doi: 10.18653/v1/2023.findings-emnlp.429
17. block M. Ernst, "Identifying textual disinformation using Large Language Models," in Proc. 2024 Conf. Hum. Inf. Interaction Retr. (CHIIR '24), New York, NY, USA, 2024, pp. 453 456. Doi: 10.1145/3627508.3638315.
18. block M. Masood et al., "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward," Appl. Intell., vol. 53, no. 4, pp. 3974 4026, Jun. 2022. Doi: 10.1007/s10489-022-03766-z.
19. block S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. Petrantonakis, "VERITE: a Robust benchmark for multimodal misinformation detection accounting for unimodal bias," Int. J. Multimedia Inf. Retr., vol. 13, no. 1, 2024. doi: 10.1007/s13735-023-00312-6.
20. block F. Alam et al., "A Survey on Multimodal Disinformation Detection," in Proc. 29th Int. Conf. Comput. Linguist. (COLING), Gyeongju, Republic of Korea, 2022, pp. 6625 6643. [Online]. Available: https://aclanthology.org/2022.coling-1.576/
21. block Z. Pan, Y. Mao, L. Xiong, T. Pang, and P. Ping, "MFAE: Multimodal Fusion and Alignment for Entity-level Disinformation Detection," Pattern Recognit. Lett., vol. 184, pp. 59 65, 2024. doi: 10.1016/j.patrec.2024.06.008.
22. block W. C. Sleeman, R. Kapoor, and P. Ghosh, "Multimodal Classification: Current Landscape, Taxonomy and Future Directions," ACM Comput. Surv., vol. 55, no. 1, pp. 1 31, 2021. doi: 10.1145/3485446.
23. block Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs," in Proc. 25th ACM Int. Conf. Multimedia (MM '17), 2017, pp. 795 816. doi: 10.1145/3123266.3123454.
24. block G. K. Shahi, "Multimodal Misinformation Detection Using Early Fusion of Linguistic, Visual, and Social Features," in Companion Publ. 17th ACM Web Sci. Conf. (WebSci Companion '25), 2025, pp. 11 18. doi: 10.1145/3720554.3733844.
25. block D. Dimitrov et al., "Detecting Propaganda Techniques in Memes," in Proc. 59th Annu. Meeting Assoc. Comput. Linguist., 2021, pp. 6603 6617. doi: 10.18653/v1/2021.acl-long.516.

26. block X. Li et al., "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," in Computer Vision ECCV 2020, Berlin, Heidelberg: Springer-Verlag, 2020, pp. 121 137. doi: 10.1007/978-3-030-58577-8_8.
27. block R. Baly et al., "What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context," in Proc. 58th Annu. Meeting Assoc. Comput. Linguist., 2020, pp. 3364 3374. Doi: 10.18653/v1/2020.acl-main.308.
28. block K. Machová, M. Mach, and V. Balara, "Federated Learning in the Detection of Fake News Using Deep Learning as a Basic Method," Sensors, vol. 24, no. 11, p. 3590, 2024. doi: 10.3390/s24113590.
29. block R. Kozik, A. Pławiak, M. Pawlak, M. Choraś, W. M. M. Saad, and K. Czachorowski, "A Meta-Analysis of State-of-the-Art Automated Fake News Detection Methods," IEEE Access, vol. 11, pp. 138822 138852, 2023. doi: 10.1109/ACCESS.2023.3339621.
30. block A. H. Jawad, M.-R. Derakhshi, and P. Sadeghi, "Enhancing Fake News Detection by Multi-Feature Classification," IEEE Access, vol. 12, pp. 2374 2388, 2024. doi: 10.1109/ACCESS.2023.3339621.
31. block B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. M. S. J. Junior, "Deepfake Image Detection Using Vision Transformer Models," in Proc. IEEE Int. Conf. Cyber Security and Resilience (CSR), 2024, pp. 1 6. doi: 10.1109/CSR61623.2024.10626302.
32. block S. M. Monteiro, A. C. Wanzeller, M. G. de Lacerda, and F. M. S. Caldeira, "Detection of Fake Images Generated by Deep Learning," in Proc. 2024 IEEE Int. Conf. Cyber Security and Resilience (CSR), 2024, pp. 1 6. doi: 10.1109/CSR61623.2024.10626303.
33. block K. Zamil and N. M. Charkari, "Image Fake News Detection Using EfficientNetB0 Model," J. Inf. Syst. Telecommun. (JIST), vol. 12, no. 47, pp. 1 12, Apr. 2024. doi: 10.52547/jist.12.47.1.
34. block Q. Xu, H. Wang, L. Meng, Z. Mi, J. Yuan, and H. Yan, "Exposing Fake Images Generated by Text-to-Image Diffusion Models," Pattern Recognit. Lett., vol. 175, pp. 71 78, Jan. 2024. doi: 10.1016/j.patrec.2023.10.021.
35. block F. Yan, M. Zhang, B. Wei, K. Ren, and W. Jiang, "SARD: Fake News Detection Based on CLIP Contrastive Learning and Multimodal Semantic Alignment," J. King Saud Univ. Comput. Inf. Sci., vol. 36, no. 10, p. 102160, 2024. doi: 10.1016/j.jksuci.2024.102160.
36. block E. M. dos Santos and P. Hermida, "Adding Compact Parameter Blocks to Multimodal Transformers to Detect Harmful Memes," in Proc. 2024 IEEE Int. Conf. Cyber Security and Resilience (CSR), 2024, pp. 1 6. doi: 10.1109/CSR61623.2024.10626304.
37. block J. Hua, X. Cui, X. Li, K. Tang, and P. Zhu, "Multimodal Fake News Detection Through Data Augmentation-Based Contrastive Learning," Appl. Soft Comput., vol. 153, p. 111125, 2024. doi: 10.1016/j.asoc.2023.110125.
38. block A. Sharma, R. Kumar, and S. Gupta, "Automatic Fake News Detection on Social Networks Using Multimodal Approach of BERT and ResNet110," in Proc. Int. Conf. Evolutionary Algorithms Soft Comput. Tech. (EASCT), 2023, pp. 1 6. doi: 10.1109/EASCT58915.2023.10420567.
39. block S. K. Uppada, P. Patel, and B. Sivaselvan, "An Image and Text-Based Multimodal Model for Detecting Fake News in OSNs," J. Database Manag., vol. 34, no. 1, pp. 1 23, 2023. doi: 10.1007/s10844-022-00764-y.
40. block Y. Zhou, Y. Yang, Q. Ying, Z. Qian, and X. Zhang, "Multimodal Fake News Detection via CLIP-Guided Learning," in Proc. 2023 IEEE Int. Conf. Multimedia Expo (ICME), 2023, pp. 2825 2830. doi: 10.1109/ICME55011.2023.00480.
41. block W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," in Proc. 38th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 139, 2021, pp. 5583 5594.
42. block P. K. Grewal, M. Ernst, and F. Hopfgartner, "Beyond Text: Leveraging Vision-Language Models for Misinformation Detection," in Proc. 2nd Int. Workshop Diffusion Harmful Content Online Web (DHOW '25), Dublin, Ireland, 2025, pp. 1 6. doi: 10.1145/3746275.3762205.
43. block J. Xu, Y. Zhang, and W. Liu, "Research on Fake News Detection Based on CLIP Multimodal Mechanism," in Proc. 3rd Int. Conf. Cyber Secur., Artif. Intell. Digit. Econ. (CSAIDE '24), 2024, pp. 72 79. doi: 10.1145/3672919.3672933.
44. block K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A New Multimodal Benchmark Dataset for Fine-Grained Fake News Detection," in Proc. 12th Lang. Resour. Eval. Conf. (LREC), Marseille, France, 2020, pp. 6149 6157.
45. block K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 6, no. 3, pp. 171 188, 2020. Doi: 10.1089/big.2020.0062.
46. block J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37 46, 1960. doi: 10.1177/001316446002000104.
47. block M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135 1144. Doi: 10.1145/2939672.2939778.
48. block M. Abdel-Basset, N. Moustafa, and H. Hawash, "Cybersecurity Threat Detection and Feature Selection Using Waterwheel Plant Optimization," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 1425 1439, 2024.
49. block A. A. Mohammad, "A Robust Multi-Transform Watermarking Scheme for Medical Images Using DTCWT, DCT, and SVD," Multimedia Tools and Applications, vol. 83, no. 14, pp. 41253 41278, 2024.
50. block S. S. Sami, M. T. Khan, and H. Al-Ghamdi, "Enhanced Hourly Temperature Prediction Using Advanced Ensemble Neural Networks for Energy Systems Efficiency Optimization in Hyper-Arid Regions," IEEE Access, vol. 11, pp. 85412 85426, 2023.
51. block Y. Wang, R. Zhang, and J. Lipinski, "Explainable artificial intelligence for wind power forecasting model based on long short-term memory," Renewable Energy, vol. 215, p. 118942, 2023.
52. block T. A. Assegie, "Optimized Deep Learning Model Using Binary Particle Swarm Optimization for Phishing Attack Detection: A Comparative Study," Journal of Cyber Security Technology, vol. 7, no. 2, pp. 93 107, 2023.
53. block E. Choi and J.-K. Kim, "TT-BLIP: Enhancing fake news detection using BLIP and tri-transformer," arXiv preprint arXiv:2403.12481, 2024.
54. block S. Tahmasebi et al., "Multimodal misinformation detection using large vision-language models," in Proc. ACM Int. Conf. Information and Knowledge Management (CIKM), 2024.
55. block P. Qi, Z. Yan, W. Hsu, and Y. Wang, "SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13052 13062.
56. block J. Wang et al., "FakeSV-VLM: Taming VLM for detecting fake short-video news," in Findings of the Association for Computational Linguistics: EMNLP, 2025.

57. block J. Lv et al., “Multi-modal fake news detection: A comprehensive survey on datasets, methods, and challenges,” Journal of Information Systems Engineering and Management, 2025.
58. block J. Li, D. Li, C. Xiong, and S. C. H. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in Proc. International Conference on Machine Learning (ICML), 2022, pp. 12888 12900.
59. block J. Li, D. Li, S. Savarese, and S. C. H. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in Proc. International Conference on Machine Learning (ICML), 2023.
60. block J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” in Advances in Neural Information Processing Systems, 2022.