Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Tri-FND: Multimodal Fake News Detection Using Triplet Transformer Models

Engy Ehab[1,*], Nahla Belal[1], Yasser Omar[2]

[1] Department of Computer Science, College of Computing and Information Technology Arab Academy for Science, Technology and Maritime Transport, Smart Village, Cairo, Egypt
[2] School of Library and Information Studies, University of Oklahoma

ARTICLE INFO

Keywords:
Fake News; Multi-Model Learning; Transformer Models; Image-Text Matching

ABSTRACT

The prevalence of fake news accompanied by multimedia content on the internet presents a significant challenge for users attempting to discern its authenticity. Automatically identifying and classifying fake news is a crucial way for combating misinformation and maintain the integrity of information dissemination. This paper proposes a fake news detection approach that exploits multimodality's potential and integrates textual and visual data to improve the fake news classification system. The novel multimodal learning approach to fake news detection, which has been termed Tri-FND, uses triplet transformers for fake news detection. This approach utilizes state-of-the-art language and vision transformers with Contrastive Language-Image Pretraining (CLIP) to improve feature representation and textual and visual semantic alignment. This technique significantly enhances the capability of identifying fake news by analyzing both text and images. Experiments were conducted on two linguistic datasets: the English dataset is sourced from Twitter, while the Chinese dataset is sourced from Weibo. The proposed approach can achieve an overall accuracy of 0.90 on the Twitter dataset and 0.93 on the Weibo dataset.

1. Introduction

Fake news represents a significant societal problem, leading to misinformation, confusion, and the manipulation of public opinion. The rapid dissemination of misinformation through social media and other online platforms can have profound societal impacts [1], including the distortion of public opinion and the erosion of trust in legitimate news sources. The primary challenge [2] lies in effectively detecting and mitigating the spread of misinformation to maintain the integrity of information. Conventional approaches [3, 4] to identify fake news often struggle to capture the nuanced cues present in multimodal data, such as images, videos, and audio. Consequently, there is a growing interest in utilizing multimodality to enhance fake news classification systems.

* *Corresponding author.*
E-mail address: enjyehab@aast.edu

https://doi.org/10.37934/araset.63.1.255270

Deep learning, specifically using transformers [5], has brought about a significant revolution in both language processing and computer vision tasks. This research explores the transformative power of transformers in textual and vision domains, focusing on their potential to drive advancements in machine intelligence. Transformers utilize self-attention mechanisms to capture relationships between input tokens, enabling parallelized computation and reduce sequential processing limitations. Vision Transformers (ViTs), as demonstrated in [6] exhibit proficient modeling of long-range dependencies and contextual relationships. ViTs, applied to images, eschew the grid-based processing of CNNs in favor of a token-based approach, capturing global contextual information and long-range dependencies within images.

Contrastive learning-based multimodal pre-training techniques have shown promising results in multimodal representation learning. The Contrastive Language-Image Pretraining model [7], (CLIP) a representative dual-stream model, is a groundbreaking approach to image-text matching, utilizing large-scale pre-training on diverse images and text. This model facilitates cross-modal understanding by embedding visual and textual representations into a common feature space, enabling accurate association with text prompts. CLIP learns to align images and text semantically, allowing robust performance across various tasks. The most recent cutting-edge approaches for detecting fake news are highlighted in the literature.

Jin *et al.* [8], introduced Att-RNN, a recurrent neural network designed for rumor detection, which incorporates textual, visual, and social contexts. It employs LSTM for textual analysis and pre-trained VGG19 for image processing. Wang *et al.* [9], introduced The EANN model, a GAN-based method, that identifies fake news by learning event-invariant features across multiple modalities. Khattar *et al.* [10], introduced MVAE as a system capable of discerning fake news. It achieves this by extracting textual and visual features, converting them into sampled multimodal representation, and subsequently learning a unified representation through joint training. Singhal *et al*., [11] propose a SpotFake a multimodal framework that uses the BERT language model for text processing and a pre-trained VGG-19 model for visual feature extraction. Singhal *et al*., [12] an improved version of SpotFake, called SpotFake+, extracts text features using pre-trained XLNet models.

Song *et al*., [13] utilize CARMN a framework for detecting multimodal fake news, which utilizes multichannel convolutional neural networks to produce fused features from word and image embeddings. It considers spatial and frequency-domain information, cross-model attention to capture image-text relationships, and self-attention to derive feature vectors for determining fake news, ensuring meaningful fusion across modalities while minimizing noise impact. Wu *et al*., [14] proposed that MCAN aims to enhance fake news identification by leveraging the interdependence among multimodal features. It overlooks shallow features. It utilizes multiple co-attention layers to effectively combine textual and visual features.

Chen *et al*. [15], proposed CAFÉ model that evaluates cross-modal ambiguity by combining cross-modal correlations and unimodal features. It uses CLIP cosine similarity to weight multimodal features, guiding classifier learning. CAFE compresses image and text data, minimizes KL divergence, and adjusts multimodal feature weights. However, alignment is not guaranteed due to limited input data. Ghorbanpour *et al*. [16], Introduced FNR, a method that evaluates the similarity between the textual and visual content of a news item to authenticate news articles. Ying Guo *et al*. [17], advanced models like Bert and ResNet are combined for feature extraction, enhancing news detection accuracy through multimodal bilinear pooling and self-attention mechanisms. Fangfang Shan *et al*. [18], This research presents EANBS, a model that utilizes BERT and Text-CNN for extracting text features, VGG-19 on ImageNet for local features, similarity representation learning, inference, event-based adversarial networks, and multimodal networks.

Previous studies, constrained by traditional techniques, often failed to identify the most effective feature representations, limiting model performance. In this study, an analysis was conducted on each modality independently, utilizing various types of vision and language transformers to evaluate their respective impacts. The best-performing models from these evaluations were subsequently identified and integrated into a multimodal framework. The CLIP model was employed to enhance the semantic alignment between the two modalities. Finally, all components were combined into a fully connected layer to achieve comprehensive integration and analysis. This approach improves the accuracy and robustness of models while offering a deeper understanding of data characteristics.

This paper presents Tri-FND, a novel multimodal learning approach using triplet transformer models for fake news detection. By leveraging transformer-based models and the CLIP model, Tri-FND achieved an overall accuracy of 0.90 on the Twitter dataset and 0.93 on the Weibo dataset.

The major contributions are highlighted as follows:

i) Various types of language and vision transformer models are employed.
ii) A comparative study is conducted on different language and vision transformers for each modality.
iii) The proposed model integrates two encoders with the purpose of enhancing feature representation by separately extracting feature representations from each modality.
iv) Contrastive Language-Image Pretraining (CLIP) is utilized to enhance the semantic alignment between the two modalities.
v) Extensive experiments are conducted on commonly used datasets: Twitter (in English) and Weibo (in Chinese).

The remainder of the paper is organized as follows: Section 2 describes the proposed architecture, Section 3 details the dataset, experimental configurations, evaluation criteria, results analysis, discussion, and ablation study, and Section 4 concludes the paper and outlines prospects for future work.

2. Methodology

This section highlights the methodologies utilized in the Tri-FND model. An architectural overview of the model is first presented, followed by a detailed explanation of the network.

2.1 Model Architecture

The Tri-FND model uses four sub-modules to identify fake news. The first sub-module handles the preprocessing for the two modalities. The second utilizes pre-trained language and vision models for textual and vision feature representation. The third uses CLIP to align text with images, Lastly, the fourth sub-module incorporates a multimodal model fusion followed by a fully connected layer for classification. Figure 1 illustrates the comprehensive architecture of the model.

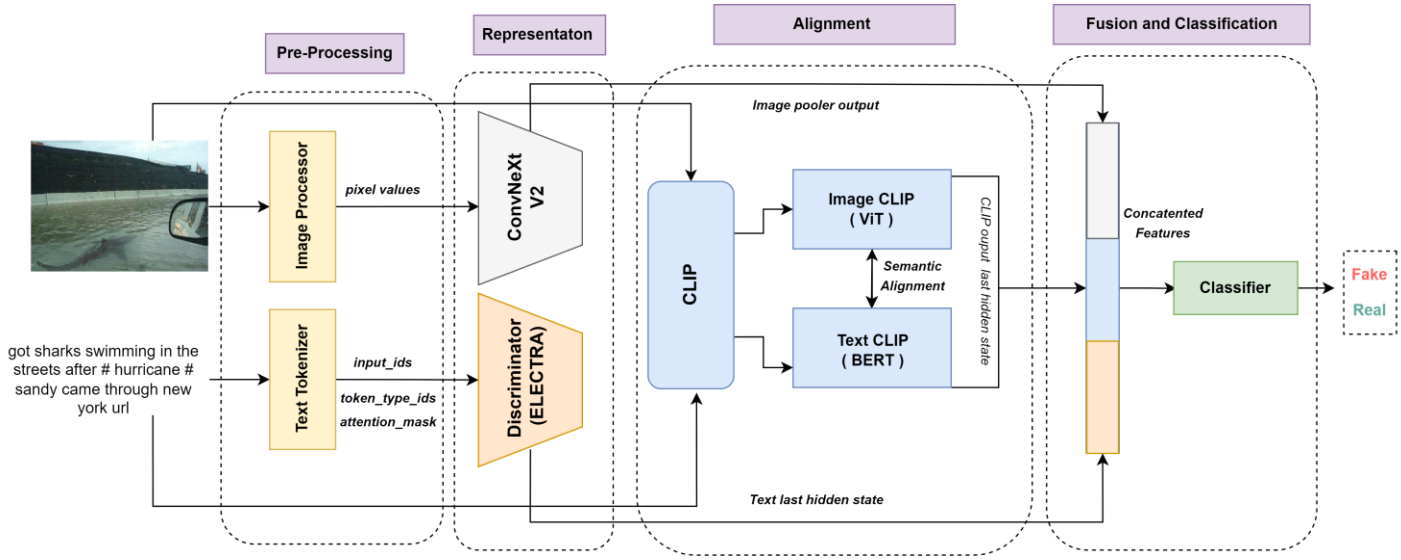

Fig.1. Tri-FND Model Architecture

2.2 Data Preprocessing

In the first module, the Tri-FND model processes tweet text and images using a text tokenizer and an image processor. The tokenizer converts the text into tokens, producing input_ids, token_type_ids, and attention_masks for linguistic analysis. Concurrently, the image processor resizes, normalizes, and transforms the image into a tensor.

2.3 Representation

In the second module, six pre-trained models are evaluated to identify the most suitable text and image representation encoders.

2.3.1 Model selection for the text modality

After preprocessing the text, word vector representations were generated using Bidirectional Encoder Representations from Transformers (BERT) [19] and its various versions. BERT is bidirectional and generates contextualized word embeddings by capturing information from both the left and right contexts of input tokens using the Masked Language Model (MLM) approach, which differs from conventional language models that predict the next word in a sequence. The following transformer-based models have been implemented; RoBERTa [20], XLM [21], ConvBERT [22], BART [23], and ELECTRA [24]. Among these models, ELECTRA demonstrated superior performance, achieving higher accuracy compared to the others. ELECTRA is short for "Efficiently Learning an Encoder that Classifies Token Replacements Accurately". This model, ELECTRA, enhances transformer network pre-training efficiency compared to BERT by introducing a task called replaced token detection. Instead of masking input parts, ELECTRA distorts them and trains a discriminative model to identify replaced tokens, allowing it to learn from every token, unlike BERT. This method increases computational efficiency and performance across various data types.

Notably, ELECTRA differs from BERT by omitting contrastive learning techniques and a pooling projection layer. In the Tri-FND model, ELECTRA is utilized as the text encoder, which uses a discriminator similar to a Generative Adversarial Network (GAN) [25], which is optimal for this task

as the dataset used contains both real and manipulated (fake) text. The ELECTRA model processes the input tokens and attention masks as in Eq. (1). The hidden state representation, H_{text}, is obtained by processing the input token matrix X_t through the ELECTRA model, conditioned on an attention mask matrix M:

$$H_{text} = Electra(X_t, M) \tag{1}$$

where X_t is the input token matrix with dimensions (B, T), where B denotes the batch size and T represents the sequence length. The matrix M is the attention mask, also of shape (B, T). Given that H_{text} returns the hidden states with dimensions (B, T, H), extracting the hidden state corresponding to the [CLS] token from H_{text}, as described in Eq. (2)

$$H_{txt} = H_{text}[:,0,:] \tag{2}$$

This formulation captures the essential representation needed for further downstream tasks. Table 3, which will be discussed in the results section, displays the outcomes of training various language transformer-based models on the proposed benchmark datasets.

2.3.2 Model selection for the image modality

Given the notable performance of Vision Transformers (ViTs) in image classification tasks, pre-trained ViT models are employed as the backbone of the network for fake news detection. Various ViTs, including the original ViT, Swin Transformer [26], PVT [27], NesT [28], DeiT [29], and ConvNeXt [30], are evaluated based on their performance on specific datasets. Among these, the Swin Transformer Base and ConvNeXt v2 demonstrate superior results, making them optimal choices for the backbone of Tri-FND.

ConvNeXt improves upon the classic ResNet [31], by incorporating modern techniques from the Swin Transformer, resulting in enhanced classification performance. While attempts to combine ConvNeXt with self-supervised learning methods like masked autoencoders (MAE) were unsatisfactory, ConvNeXt v2 [32], offers significant enhancements. The key improvement in ConvNeXt v2 involves the Fully Convolutional Masked Autoencoder (FCMAE) architecture, which processes visible pixels with an encoder and uses a decoder to reconstruct images from encoded pixels and mask tokens. Additionally, the introduction of a Global Response Normalization (GRN) layer helps the model better differentiate image features, especially in partially obscured images, greatly enhancing image recognition and processing performance. As described in Eq. (3), the hidden state representation, H_{image} , is derived by processing the input image matrix X_i through the vision model I

$$H_{image} = I(X_i) \tag{3}$$

where, I denotes the vision model, and X_i represents the input image matrix. The final hidden state representation is derived from extracting the pooler output from the last layer of the sequence in H_{image} as in Eq. (4):

$$H_{img} = H_{image}.pooler_output \tag{4}$$

2.4 Alignment

The third module employs CLIP, a multimodal dual vision and language model leveraged to assess cross-modal correlation and align textual and visual features, thereby enhancing the effectiveness of identifying fake news. CLIP was utilized to encode both textual and visual information into a shared space, facilitating semantic alignment across modalities as described in Eq. (5).

$$H_{clip} = C(C_t, C_v) \tag{5}$$

CLIP encodes visual and textual data through separate encoder networks: a Vision Transformer (ViT) for images and BERT for text. The key concept is embedding both images and text into a shared feature space where similar semantic concepts are closely located. This facilitates direct comparisons and semantic associations without explicit alignment guidance. The hidden states generated encapsulate the learned representations of both modalities, encoding the semantic information of the input text and images to enable cross-modal comparisons and understanding. As described in Eq. (6,7) the last hidden state outputs for both text and image from the CLIP model are extracted.

$$H_{txtclip} = H_{clip}.text_{model}.output.last_hidden_state[:,0,:] \tag{6}$$

$$H_{imgclip} = H_{clip}.vision_{model}.output.last_hidden_state[:,0,:] \tag{7}$$

CLIP's hidden states achieve semantic alignment across modalities, enabling meaningful associations between related image and text content without needing explicit point-to-point correspondences.

2.5 Multimodal Classification

In the last module, encoder vectors are concatenated to create a comprehensive feature representation. For the English dataset, this includes a 256-dimensional vector from Electra-S, a 1024-dimensional vector from ConvNeXt v2-base, a 768-dimensional vector from Text CLIP (BERT-B), and a 512-dimensional vector from Image CLIP (ViT-B), resulting in a 2560-dimensional vector. For the Chinese dataset, Electra-L generates a 1024-dimensional vector and CLIP Image Model resulting in a 768-dimensional vector leading to a concatenated 3584-dimensional vector. as described in Eq (8). These concatenated vectors are then processed through a fully connected layer, followed by a sigmoid activation function as in Eq (9).

$$H_{Concat} = Concat(H_{txt}, H_{img}, H_{txtclip}, H_{imgclip}) \tag{8}$$

$$Output = \sigma(H_{concat}W_{top} + b_{top}) \tag{9}$$

where σ is the sigmoid function, $W_{top}$is the weight matrix, and b_{top} is the bias term of the final linear layer.

3. Results

This section will present a comprehensive description and the experimental results of applying Tri-FND using two real-world datasets. Furthermore, we will compare this approach with the baseline models.

3.1 Dataset

The effectiveness of the Tri-FND model was comprehensively evaluated through experiments conducted on English and Chinese datasets as outlined in [33, 34] and [35]. The Twitter dataset, aimed at the Verifying Multimedia Use task, consists of tweets containing text, images, and social context details. The Weibo dataset, sourced from Xinhua News Agency and the Weibo platform, includes posts with text, images, and social information, collected from May 2012 to January 2016 via Weibo's official fake news debunking system. These datasets serve as valuable resources for assessing the model's capability in detecting fake news and multimedia content on diverse social media platforms. Their varied events facilitate the effective generalization of the model to different scenarios. Visual representations of event distribution percentages will be included (Figure 2 and 3), while Table 1 lists dataset training and testing distributions.

Table 1
Datasets class distributions

Dataset	Train		Test		Total
	Fake	Real	Fake	Real	
Twitter	6649	4599	545	444	12,237
Weibo	3748	3758	999	995	9500

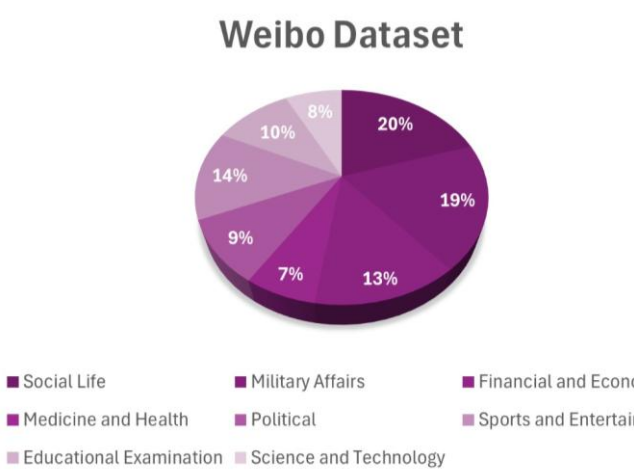

Fig. 2. Weibo events distribution

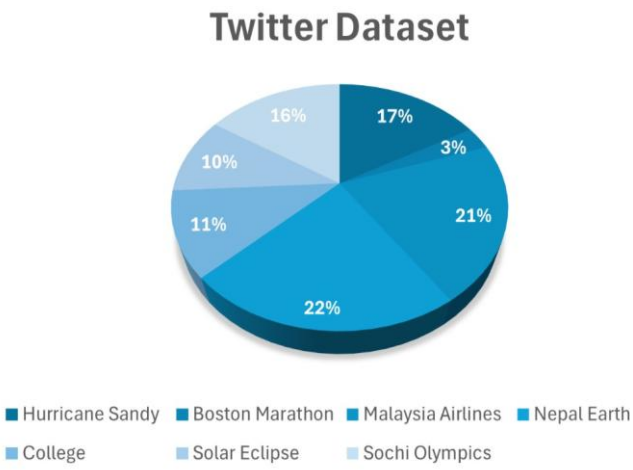

Fig. 3. Twitter events distribution

3.2 Experimental Setup

Through extensive experiments and repeated adjustments, the optimal parameter configurations were identified. The test set was used to determine the most suitable hyperparameters for the training set, aiding in model generalization and preventing overfitting. These configurations are detailed in Table 2.

Table 2
Hyperparameters Setting

Hyperparameter	Twitter	Weibo
Optimizer	AdamW	AdamW
Epsilon	1e-8	1e-8
Batch Size of uni-modals	32	32
Batch Size for multimodal	8	4
The text learning rate for uni modals	2e-5	1.6e-5
The text learning rate for Multimodal	2e-6	2e-6
Image Learning rate	1e-4	1e-4
Text maximum Sentence length	75	285
Image size	224x224	224x224
Number of epochs	20	20
Text encoder dimension	256	1024
Image encoder dimension	1024	1024
Clip text encoder dimension	768	768
Clip image encoder dimension	512	768

3.3 Implementation Detail

The experiments were performed on the Google Colaboratory platform using NVIDIA A100-SXM4-40GB. All pre-trained models utilized in the study were acquired from the Hugging Face library [36].

3.4 Evaluation Metrics

To evaluate the performance of the Tri-FND model, traditional metrics such as accuracy, F1 score, recall, precision, ROC, and AUC were utilized, which were computed as follows in Eqs. (10)-(13):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{F1-Score} = \frac{2 * (Precision\ x\ Recall)}{Precision + Recall} \tag{13}$$

i) Receiver Operating Characteristic (ROC): The curve illustrates a classifier's capability to differentiate between fake and real news instances by plotting the true positive rate compared to the false positive rate.
ii) The Area Under the Curve (AUC): It serves as an indicator of a classifier's performance in distinguishing between fake and real news.

3.5 Baseline Models

To validate the effectiveness of the proposed model, a comparison with various baseline models was conducted. Table 3 outlines the contributions, methods used, and future scope for each model.

Table 3
Review of literature baseline models

Model	Contribution	Text encoder	Image Encoder	Fusion	Future Scope
Att-RNN (2017)	Combines textual and visual features with social context using an attention mechanism	Word2Vec	VGG19	Attention network	To improve the effectiveness of the model
EANN (2018)	Utilizes an event discriminator to identify event-specific data	Text-CNN	VGG19	Concatenation	To enhance the proposed model fusion network
MVAE (2019)	Leverages VAE to discovers correlations between different modalities	Bi-LSTM	VGG19	Concatenation	Leveraging tweet dissemination and user attributes.
Spotfake+ (2020)	The main novelty of the proposed model lies in its utilization of the pre-trained language model XL-Net.	XL-Net	VGG19	Concatenation	To Include meta-levels of different modalities.
MCAN (2021)	Proposes multiple co-attention layers aimed at integrating and learning inter-modality relationships.	BERT	VGG19	Multiple co-attention layers	To Expand the fusion process using a co-attention network
CARMN (2021)	The model preserves unique properties while reducing noise introduced during the fusion of different modalities.	Word level embedding	VGG19	Concatenation +Attention	Enhancing Event-based multimodal fake news
CAFÉ (2022)	This model evaluates cross-modal ambiguity by combining cross-modal correlations and unimodal features.	Bert	ResNet	Cross-model Fusion	Enhanced Cross-Modal Fusion Techniques
FNR (2023)	Proposes similarity between news images and text by examining the correlation among news articles associated with a particular event	BERT	ViT-base	Concatenation + Similarity	To construct and utilize user network graphs to leverage relationships between users and their shared news
MBPAM (2023)	A technique for identifying fake news by combining textual and visual data, initially employing a two-branch approach to capture hidden layer details of each modality for extracting more valuable features	BERT	ResNet	Bilinear pooling method	To incorporate social subject information and merge textual data with multiple images.
EANBS (2024)	This model integrates event-based adversarial networks with multimodal networks to learn the associations between modal features and events.	Bert, TextCNN	VGG19	Concatenation	To investigate the incorporation of video information features.

3.6 Results Analysis

This section presents a comprehensive analysis of the experimental results for the Tri-FND framework, highlighting its performance across various models. The results are shown in Tables 4, 5, and 6, focusing on both single-modality (textual and visual) and multimodality experiments.

Table 4 summarizes the performance of different language models on both the Twitter and Weibo datasets. BERT-base achieves moderate accuracy on the Twitter dataset, while Conv-Bert-base shows improvement with higher accuracy and recall for fake news. Electra-small stands out with the highest accuracy, precision, and F1 score on Twitter, indicating notable performance.

For Weibo Dataset, BERT-Chinese achieved the highest performance with an accuracy of 0.90. Conv-BERT-base achieved a lower performance with an accuracy of 0.75. XLM-base achieved an accuracy of 0.78. RoBERTa-base matches XLM-base's performance with an accuracy of 0.79. BART-base shows strong results with an accuracy of 0.81. Electra-large achieved the highest accuracy, following the Bert model for this dataset with an accuracy of 0.87. Additionally, it demonstrated a strong F1-score, attaining 0.86 for fake news and 0.88 for real news.

Table 5 evaluates the performance of vision models on both the Twitter and Weibo datasets. ViT-base achieves an accuracy of 0.60 on the Twitter dataset, with the highest recall for real news. Swin-base demonstrates significant improvement with an accuracy of 0.62. DeiT-small achieves the highest accuracy of 0.72 and exhibits a better F1-score for both fake and real news. ConvNeXtV2 emerges as the best performer with an accuracy of 0.74, the highest recall for fake news, and notable F1 scores for both fake and real news.

For the Weibo dataset, the ViT-base shows moderate performance with an accuracy of 0.82. Swin-base is the best performer in this dataset with an accuracy of 0.89 and the highest F1-score of 0.87 and a recall of 0.89 for fake news. PVT-tiny achieved an accuracy of 0.84 and an F1-score of 0.83 for fake news and 0.85 for real news, NesT-base achieved an accuracy of 0.88 but the highest precision with 0.91 for fake news and the highest precision and recall with 0.89 and 0.90 for real news. DeiT-small achieved the highest recall for real news with 0.94. ConvNeXtV2 exhibited the lowest performance, with an accuracy of 0.77, and balanced precision, recall, and F1 score for both real and fake news. Despite this, its potential as an effective encoder for multimodal learning will be discussed in subsequent sections.

Table 6 presents the multi-modal experimental methods used in Tri-FND. The performance of multimodal models that combine text and image data using various encoders on both the Twitter and Weibo datasets. Electra-small + ConvNeXtV2 + CLIP emerges as the top performer on the Twitter dataset, achieving an accuracy of 0.90 and excelling across other metrics. Integration of the CLIP model notably enhances semantic alignment between the two modalities. On the Weibo dataset, Electra-large + ConvNeXtV2 + CLIP stands out as the top performer, achieving an accuracy of 0.93 and balanced F1-scores for both fake and real news, indicating optimal performance. For experimental comparison, results are listed in Table 7 and the baseline models used for comparison are discussed in the literature review section and Table 3.

Table 4
Language models result for the Twitter and Weibo test set

Text Modality	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
Twitter Dataset	BERT-B	0.63	0.66	0.67	0.66	0.59	0.59	0.59
	Conv-Bert-B	0.64	0.67	0.68	0.67	0.60	0.58	0.59
	XLM-B	0.58	0.63	0.58	0.60	0.53	0.57	0.55
	RoBERTa-B	0.58	0.65	0.54	0.59	0.53	0.64	0.58
	BART-B	0.54	0.64	0.40	0.49	0.50	0.72	0.59
	Electra-S	0.65	0.75	0.61	0.67	0.65	0.61	0.63
Weibo Dataset	BERT-Chinese	0.90	0.93	0.86	0.89	0.87	0.93	0.90
	Conv-Bert-B	0.75	0.78	0.70	0.74	0.73	0.81	0.76
	XLM-B	0.78	0.82	0.73	0.77	0.75	0.84	0.79
	RoBERTa-B	0.79	0.86	0.70	0.77	0.74	0.89	0.81
	BART-B	0.81	0.84	0.76	0.80	0.78	0.86	0.82
	Electra-L	0.87	0.91	0.82	0.86	0.83	0.92	0.88

Table 5
Vision models result for the Twitter and Weibo test set

Image Modality	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-Score	Precision	Recall	F1 - Score
Twitter Dataset	ViT-B	0.60	0.75	0.42	0.54	0.54	0.83	0.65
	Swin-B	0.62	0.76	0.46	0.57	0.55	0.82	0.66
	PVT-T	0.56	0.67	0.41	0.51	0.51	0.75	0.61
	NesT-B	0.57	0.70	0.40	0.51	0.52	0.79	0.62
	DeiT-S	0.72	0.77	0.70	0.73	0.66	0.74	0.70
	ConvNeXtV2	0.74	0.75	0.79	0.77	0.72	0.67	0.77
Weibo Dataset	ViT-B	0.82	0.78	0.83	0.81	0.86	0.82	0.84
	Swin-B	0.89	0.85	0.89	0.87	0.91	0.87	0.89
	PVT-T	0.84	0.79	0.86	0.83	0.88	0.82	0.85
	NesT-B	0.88	0.91	0.83	0.86	0.89	0.92	0.90
	DeiT-S	0.83	0.90	0.68	0.78	0.79	0.94	0.86
	ConvNeXtV2	0.77	0.76	0.80	0.78	0.79	0.74	0.77

Table 6
Multimodal model for fake and real class for the Twitter and Weibo test set

Multimodal	Text Encoder	Image Encoder	CLIP	Accuracy	Fake News			Real News		
					Precision	Recall	F1-Score	Precision	Recall	F1-Score
Twitter Dataset	Bert	Swin-B	-	0.78	0.76	0.86	0.81	0.80	0.68	0.73
	ConvBert	Swin-B	-	0.83	0.80	0.92	0.86	0.88	0.73	0.80
	Electra-S	Swin-B	CLIP	0.86	0.83	0.95	0.86	0.92	0.76	0.85
	Electra-S	ConvNexTV2	CLIP	0.90	0.90	0.91	0.91	0.89	0.88	0.89
Weibo Dataset	Bert	Swin-B	-	0.89	0.91	0.87	0.89	0.87	0.91	0.89
	Electra-L	Swin-B	-	0.90	0.89	0.91	0.90	0.90	0.89	0.90
	Electra-L	Swin-B	CLIP	0.93	0.93	0.93	0.93	0.93	0.93	0.93
	Electra-L	ConvNexTV2	CLIP	0.93	0.93	0.94	0.93	0.94	0.93	0.93

Table 7
Presents a comparison with baseline models for the Twitter and Weibo test set

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score
Twitter	Att-RNN	0.68	0.78	0.61	0.68	0.60	0.77	0.67
	EANN	0.69	0.75	0.58	0.65	0.62	0.76	0.69
	MVAE	0.74	0.80	0.71	0.75	0.68	0.77	0.77
	SpotFake++	0.79	0.79	0.82	0.81	0.78	0.74	0.76
	MCAN	0.80	0.88	0.76	0.82	0.73	0.87	0.79
	CARMN	0.74	0.85	0.61	0.71	0.67	0.88	0.76
	CAFE	0.80	0.80	0.79	0.80	0.80	0.81	0.80
	FNR	0.78	0.78	0.85	0.82	0.79	0.71	0.75
	MBPAM	0.86	0.83	0.75	0.79	0.88	0.92	0.90
	EANBS	0.86	0.85	0.88	0.86	0.88	0.84	0.86
	Tri-FND (Ours)	0.90	0.90	0.91	0.91	0.89	0.88	0.88
Weibo	Att-RNN	0.78	0.86	0.68	0.76	0.73	0.89	0.81
	EANN	0.81	0.89	0.66	0.76	0.77	0.93	0.85
	MVAE	0.82	0.85	0.76	0.80	0.80	0.87	0.83
	SpotFake++	0.87	0.88	0.84	0.86	0.85	0.89	0.87
	MCAN	0.89	0.91	0.88	0.90	0.88	0.90	0.89
	CARMN	0.85	0.89	0.81	0.85	0.81	0.89	0.85
	CAFE	0.84	0.85	0.83	0.84	0.82	0.85	0.83
	FNR	0.87	0.87	0.89	0.88	0.88	0.87	0.88
	MBPAM	0.90	0.94	0.87	0.90	0.86	0.94	0.90
	EANBS	0.89	0.87	0.91	0.89	0.90	0.88	0.89
	Tri-FND (Ours)	0.93	0.93	0.94	0.93	0.94	0.93	0.93

3.7 Discussion

The results demonstrate that integrating multimodal data significantly improves fake news detection compared to single-modality approaches, with models utilizing the CLIP architecture showing superior accuracy and balanced precision across both Twitter and Weibo datasets. Weibo dataset performance is generally higher, indicating potential benefits from its specific characteristics. Insights from language and vision models suggest Electra-S and BERT-Chinese perform well on Twitter and Weibo, respectively, while ConvNeXtV2 excels on Twitter and Swin-base and ConvNeXtV2 leads on Weibo. Combining Electra with ConvNeXtV2 achieves the best overall performance, underscoring the advantage of integrating text and image data. Advanced transformer-based models combined with powerful image encoders offer superior fake news detection across platforms and modalities. Figures 4 and 5 illustrate the ROC curve for both datasets on the test set, achieving a score of 0.91 for the English dataset and 0.98 for the Chinese dataset. Furthermore, the results derived from the confusion matrix for the test sets, depicted in Figures 6 and 7, indicate 496 false positives (FP), 53 false negatives (FN), 49 true negatives (TN), and 391 true positives (TP) for the English dataset. For the Chinese dataset, the figures are 935 false positives (FP), 75 false negatives (FN), 64 true negatives (TN), and 920 true positives (TP).

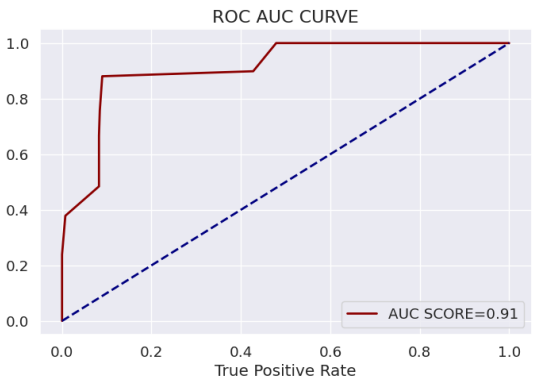

Fig. 4. Illustrates the ROC Curve on the Twitter test set

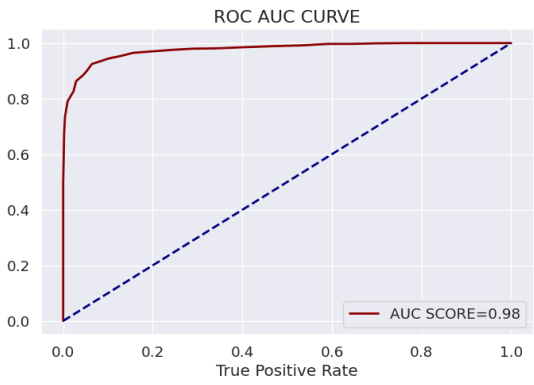

Fig. 5. Illustrates the ROC Curve on the Weibo test set

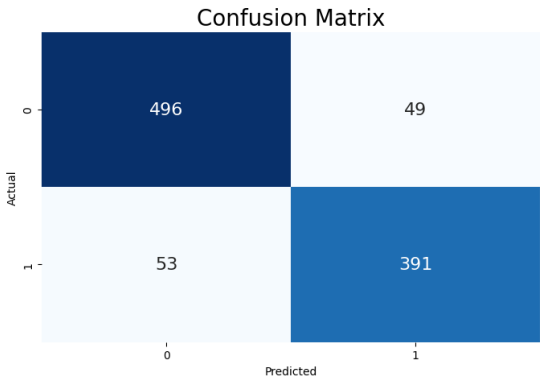

Fig. 6. Confusion matrix of the Twitter dataset on the test set

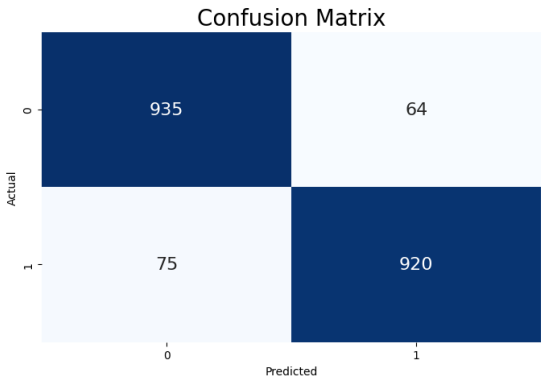

Fig. 7. Confusion matrix of The Weibo dataset on the test set

3.8 Ablation Study

An ablation study was conducted to determine the most beneficial data modality, justify the use of a multi-modal approach, and evaluate the inclusion of CLIP in the Tri-FND model. Initially, each modality was utilized and tested independently. Subsequently, the modalities were combined, and the effectiveness of a multi-modal approach for fake news detection was analyzed.

The data presented in Table 8 on the Twitter dataset highlights that due to the concise, imprecise, and often inappropriate language used in tweets, individual tweet analysis alone is less accurate, with image analysis performing better. However, combining both modalities leads to improved outcomes, indicating that they compensate for each other's weaknesses. Further enhancements are observed when integrating the Clip model, which examines the text-image semantic alignment. In contrast, Table 8 indicates that images on Weibo lack expressiveness and are susceptible to manipulation, which facilitates the dissemination of fake information. Consequently, relying exclusively on visual analysis proves ineffective for detecting fake news. In contrast, text analysis demonstrates superior performance in this context. Nonetheless, the integration of both modalities results in a significant improvement in detection performance, highlighting their complementary nature. Furthermore, considering the relationship between text and images using the CLIP model further enhances the accuracy.

Table 8
Ablation study on the Twitter and Weibo datasets

Dataset	Method	Accuracy	AUC
Twitter	Text	0.65	0.70
	Image	0.74	0.79
	CLIP	0.75	0.82
	Text + Image	0.74	0.83
	Text + Image + CLIP	0.90	0.91
Weibo	Text	0.87	0.91
	Image	0.77	0.84
	Chinese CLIP	0.92	0.96
	Text + Image	0.89	0.95
	Text + Image + CLIP	0.93	0.98

4. Conclusion

In conclusion, this paper introduces Tri-FND, a model designed to detect fake news on social media by leveraging triplet transformer models. By utilizing state-of-the-art language and vision transformers with Contrastive Language-Image Pretraining (CLIP), Tri-FND improves the alignment between text and image representations. Experimental results on Twitter and Weibo datasets demonstrate Tri-FND's effectiveness, achieving high-performance scores compared to baseline models. Future advancements in fake news detection could involve incorporating additional modalities and leveraging network graphs of users for improved accuracy. Building public trust in fake news detection requires reliable and interpretable machine-learning methods. Therefore, employing explainable approaches and providing users with explanations is crucial for enhancing detection accuracy. The next steps involve implementing such approaches to further enhance fake news detection.

Acknowledgment
This research was not funded by any grant.

References
[1] Ghani, M. M., W. A. Mustafa, D. L. S. Bakhtiar, and M. Khairudin, “A Comprehensive Study: AI Literacy as a Component of Media Literacy,” *Journal of Advanced Research in Applied Sciences and Engineering Technology* 53, no. 2, (2025): 112–121. https://doi.org/10.37934/araset.53.2.112121
[2] Ghani, M. M. *et al.*, “Current Approaches of Artificial Intelligence (AI) in Leading Behavioural Change: The Latest Review,” *Journal of Advanced Research in Applied Sciences and Engineering Technology* 35, no. 1, (2024):143–155. https://doi.org/10.37934/araset.34.3.143155
[3] Krishna, M. M., Midhunchakkaravarthy, and J. Vankara, “Detection of Sarcasm using Bi-Directional RNN Based Deep Learning Model in Sentiment Analysis,” *Journal of Advanced Research in Applied Sciences and Engineering Technology* 31, no. 2, (2023): 352–362. https://doi.org/10.37934/araset.31.2.352362
[4] Shohan, M. H. *et al.*, “Use of Natural Language Processing for the Detection of Hate Speech on Social Media,” *Journal of Advanced Research in Applied Sciences and Engineering Technology* 51, no. 2, (2025):86–96. https://doi.org/10.37934/araset.51.2.8696
[5] Ashish, Vaswani. "Attention is all you need." *Advances in neural information processing systems* 30 (2017): I.
[6] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
[7] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748-8763. PmLR, 2021.

[8] Jin, Zhiwei, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. "Multimodal fusion with recurrent neural networks for rumor detection on microblogs." In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 795-816. 2017. https://doi.org/10.1145/3123266.3123454

[9] Wang, Yaqing, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. "Eann: Event adversarial neural networks for multi-modal fake news detection." In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 849-857. 2018. https://doi.org/10.1145/3219819.3219903

[10] Khattar, Dhruv, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. "Mvae: Multimodal variational autoencoder for fake news detection." In *The world wide web conference*, pp. 2915-2921. 2019. https://doi.org/10.1145/3308558.3313552

[11] Singhal, Shivangi, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. "Spotfake: A multi-modal framework for fake news detection." In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pp. 39-47. IEEE, 2019. https://doi.org/10.1109/BigMM.2019.00-44

[12] Singhal, Shivangi, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 10, pp. 13915-13916. 2020. https://doi.org/10.1609/aaai.v34i10.7230

[13] Song, Chenguang, Nianwen Ning, Yunlei Zhang, and Bin Wu. "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks." *Information Processing & Management* 58, no. 1 (2021): 102437. https://doi.org/10.1016/j.ipm.2020.102437

[14] Wu, Yang, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. "Multimodal fusion with co-attention networks for fake news detection." In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pp. 2560-2569. 2021. https://doi.org/10.18653/v1/2021.findings-acl.226

[15] Chen, Yixuan, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. "Cross-modal ambiguity learning for multimodal fake news detection." In *Proceedings of the ACM web conference 2022*, pp. 2897-2905. 2022. https://doi.org/10.1145/3485447.3511968

[16] Ghorbanpour, Faeze, Maryam Ramezani, Mohammad Amin Fazli, and Hamid R. Rabiee. "FNR: a similarity and transformer-based approach to detect multi-modal fake news in social media." *Social Network Analysis and Mining* 13, no. 1 (2023): 56. https://doi.org/10.1007/s13278-023-01065-0

[17] Guo, Ying, Hong Ge, and Jinhong Li. "A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism." *Frontiers in Computer Science* 5 (2023): 1159063. https://doi.org/10.3389/fcomp.2023.1159063

[18] Shan, Fangfang, Huifang Sun, and Mengyi Wang. "Multimodal Social Media Fake News Detection Based on Similarity Inference and Adversarial Networks." *Computers, Materials & Continua* 79, no. 1 (2024). https://doi.org/10.32604/cmc.2024.046202

[19] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171-4186. 2019.

[20] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

[21] Conneau, Alexis, and Guillaume Lample. "Cross-lingual language model pretraining." *Advances in neural information processing systems* 32 (2019).

[22] Jiang, Zi-Hang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. "Convbert: Improving bert with span-based dynamic convolution." *Advances in Neural Information Processing Systems* 33 (2020): 12837-12848.

[23] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019). https://doi.org/10.18653/v1/2020.acl-main.703

[24] Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555* (2020).

[25] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

[26] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012-10022. 2021. https://doi.org/10.1109/ICCV48922.2021.00986

[27] Wang, Wenhai, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568-578. 2021. https://doi.org/10.1109/ICCV48922.2021.00061

[28] Zhang, Zizhao, Han Zhang, Long Zhao, Ting Chen, Sercan Ö. Arik, and Tomas Pfister. "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, pp. 3417-3425. 2022. https://doi.org/10.1609/aaai.v36i3.20252

[29] Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. "Training data-efficient image transformers & distillation through attention." In *International conference on machine learning*, pp. 10347-10357. PMLR, 2021.

[30] Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and S. A. Xie. "ConvNet for the 2020s. arXiv." *arXiv preprint arXiv:2201.03545* 10 (2022). https://doi.org/10.1109/CVPR52688.2022.01167

[31] Targ, Sasha, Diogo Almeida, and Kevin Lyman. "Resnet in resnet: Generalizing residual architectures." *arXiv preprint arXiv:1603.08029* (2016).

[32] Woo, Sanghyun, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. "Convnext v2: Co-designing and scaling convnets with masked autoencoders." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16133-16142. 2023. https://doi.org/10.1109/CVPR52729.2023.01548

[33] Boididou, Christina, Symeon Papadopoulos, Yiannis Kompatsiaris, Steve Schifferes, and Nic Newman. "Challenges of computational verification in social multimedia." In *Proceedings of the 23rd international conference on world wide web*, pp. 743-748. 2014. https://doi.org/10.1145/2567948.2579323

[34] Larson, Martha, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. "The benchmarking initiative for multimedia evaluation: MediaEval 2016." *IEEE MultiMedia* 24, no. 1 (2017): 93-96. https://doi.org/10.1109/MMUL.2017.9

[35] Wu, Ke, Song Yang, and Kenny Q. Zhu. "False rumors detection on sina weibo by propagation structures." In *2015 IEEE 31st international conference on data engineering*, pp. 651-662. IEEE, 2015. https://doi.org/10.1109/ICDE.2015.7113322

[36] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface's transformers: State-of-the-art natural language processing." *arXiv preprint arXiv:1910.03771* (2019). https://doi.org/10.18653/v1/2020.emnlp-demos.6