International Journal of Enhanced Research in Management & Computer Applications,
ISSN: 2319-7471, Volume 15 Issue 7, July 2026

Explainable Hybrid Transformer Model Based Fake News Detection System

Anurag[1], Amandeep[2]

[1]M.Sc. Computer Science, Artificial Intelligence and Data Science, GJUS&T
[2]Assistant Professor, Artificial Intelligence and Data Science, GJUS&T

ABSTRACT

This new era of digitalization comes with great issues of democratic values, breach of people's trust and hindrances to principles of welfare: distorted facts in media. Unimodal text detection can detect misleading text. However, social media is increasingly used for mismatched information involving many modalities. Mismatched information which includes misleading text along with manipulated images will challenge unimodal text detection. This paper presents an extensive study of fake news detection, which involves manually curated linguistic features, classic machine learning techniques, DL, transfomer, and fusion multimodal. We examine the key multimodal benchmark, the Fakeddit dataset. We also look at text , image based detection systems and multimodal detection systems. Lastly, we analyze the new explainable AI techniques like SHAP and LIME that make the detection systems more clear and trustworthy. We highlight certain research spheres that currently do not feature in the pertinent literature. Examples of these are dual-modality explainability, systems for the field and better integration of explainability in multimodal architectures.

Keywords – Multimodal Learning, DistilBERT, ResNet18, SHAP, LIME , Fakeddit, HEMT-Fake, ML, DL.

INTRODUCTION

Platforms for social interaction online like twitter, Facebook, and Reddit, have changed how individuals produce and disseminate information. Social media delivered information to everyone, but they also helped in spreading disinformation, rumors, and fake news very quickly [1]. Fake news typically refers to false information or misleading information which is shared as news [2]. It leads to causing harm to people and even misleading large sectors like elections, healthcare, and financial markets [3].

It has become a huge trouble. Research indicates that on Twitter, fake news gets disseminated six times faster compared to real news. Moreover, when subject to extensive information, people are not able to tell apart the real news from fake. Before, automatic methods used traditional ML models and features that were created by people manually. Deep learning and the pre-trained transformer models that came from it, such as BERT [4] and its smaller versions, have greatly boosted the accuracy of text detection. Social media fake news isn't just about text. Most of the time, there will be a picture with the post.A catchy headline paired with either a very specific or not well-connected photo often gives a strong ability to mislead people [6].Because of this, creating a system that uses both text and images together (called a multimodal system) is a key area of research [7][8][9].

Another challenge is figuring out what is happening in the decisions a detection system makes. Deep learning models can achieve high accuracy in identifying fake news, but they work like black boxes, meaning they don't explain why a particular item is classified as fake. Decisions made automatically, which don't make sense when looking at news, laws, or public health, make people lose trust and are less likely to be used. The application of XAI methods (e.g. SHAP and LIME are still new, and research is being done on systems that use different types of data [12][13].

Besides looking at different ways for fake news detection, this review also highlights the need for explainability after the fact. This facilitates the development of accurate ,trustworthy ,systematic news detection systems that can be easily verified and relied upon. People don't have full trust in a content moderation system when high trust is needed, such as for journalism, information sharing, or public health messaging [10][13]. These explanations highlight the differences

between SHAP and LIME, exhibit advantages of each, when applied to text and images while connecting the two methodologies.

Main conclusions of the paper are:

- We examine the various approaches utilized to detect fake news in a systematic manner, beginning with traditional ML methods, then progressing to deep learning and transformer-based approaches, and finishing with proposals that leverage multiple types of information to give a complete overview of the area
  .
- The Fakeddit dataset, which involves multimodal tasks, is the primary test set for various systems that have been studied on it. We also demonstrate how these systems operate and the quality of service they provide [14].

- We examine the explanation methods, like SHAP and LIME, and find that no system that detects fake news based on both text and images can explain it position in terms of each data type at the same time yet. It shows this is still an open problem in the research [10][11][13].

LITERATURE REVIEW

The researchers made it clear at the start of the study which method they would employ for the literature review. The published research on detecting fake news, both using traditional machine learning techniques and contemporary multi-modal transformer-based techniques, was examined. In the text of a review, it is described that Nakamura et al.[14] introduced Fakeddit. Thus Fakeddit which is a new dataset for multimodal analysis is the most widely used one. After that, this includes text-based headlines and related images of a large number of posts. These posts are on the Reddit community and make it very suitable for studying how text and images work together.

The analysis covered every article that suggested a solution for finding fake news, crafted a set of examples for testing methods, described usage of explainability tools in classification process, and introduced basic model which was part of a system used to detect fake news, e.g. the Transformer model by Vaswani et al.[17], BERT by Devlin et al.[4], DistilBERT by Sanh et al.[5], or ResNet by He et al.[15]. The selected articles were classified into five groups: classical ML methods, DL methods operating on text, methods based on transformers, multimodal fusion, and explainability methods.Papers under study were analyzed based on their method of detection, type of data input, dataset studied, and whether or not the system gives reasons understandable by human beings, as per Ribeiro et al.[10] and Lundberg and Lee [11].

The different tests of the various systems were done on different sets of data and ways of splitting it, so the results should be understood carefully and in relation to how each system was set up in the experiment. So that we can compare the results, we use familiar metrics like accuracy, precision, recall, F1-Score, and ROC-AUC scores when they are relevant. Making direct comparisons between numbers is possible, but it's done carefully.This is because different systems are tested on different data sets and splits, and these comparisons are always considered in the context of how the system was set up during the experiment.

The studies looked at are grouped based on their architecture design, how they are trained, and how easy it is to understand their results, as explained by Ribeiro et al.[10] and Arrieta et al.[13]. The fusion strategies and dataset coverage of multimodal fusion approaches such as EANN by Wang et al. [6], SpotFake by Singhal et al. [7], MVAE by Khattar et al.

[8], and multimodal fake news detection system by SeguraBedmar et al. [9] are evaluated. In this regard, researchers Ribeiro et al. [10] (LIME) and Lundberg and Lee [11] (SHAP) evaluate the applicability of explainability methods for multimodal systems. The contribution of these content-based explainability approaches by Raj and Thakur [12], and XAI taxonomies by Arrieta et al. [13] to transparent fake news detection has been reviewed. We evaluate the LIAR dataset by Wang [18] and the Fakeddit dataset by Nakamura et al. [14] in terms of usefulness as evaluation benchmarks. The cross modal signal exploitation strategy of Consistency based multimodal detection by Xue et al. [20] is reviewed.

In summary, the authors aim to classify fake news detection from different perspectives, offering a consistent picture of the different problems pertaining to fake news detection. The ultimate aim is to aid the in designing a fake news detection system that is more accurate, transparent and deployable. The design will take into account the explainability frameworks.

Table 1: Related work in Fake News Detection

Study	Method	Dataset Used	Performance & Limitations
Pérez-Rosas et al.,	SVM + TF-IDF	LIAR	~74% accuracy; hand-crafted features, no semantic understanding
Ribeiro et al.	LIME	Any classifier	Model-agnostic explainability; unstable across runs
Lundberg & Lee	SHAP	Any model	Consistent attribution; slow for large models
Sanh et al.	DistilBERT	Multiple classification tasks	97% of BERT at 60% fewer parameters; text-only, no visual modality
Wang et al.	EANN: text + image + adversarial	Weibo, Twitter	~82.2% accuracy; no explainability, event-specific
Singhal et al.	SpotFake: BERT + VGG-19	PolitiFact, GossipCop	~77.3% accuracy; no explainability, heavy architecture
Segura-Bedmar etal.	DistilBERT + image features	Fakeddit	~85% accuracy; no explainability, limited image encoder
Raj & Thakur	Content features + LIME	Multiple datasets	~89% accuracy;no image modality

Research Gap

Even though there have been some good developments, there are still several missing parts in what has been written so far.

Lack of Dual-Modality Explainability- Current multi-modal systems either don't provide explanations or completely ignore the need for them. The explanation for the image type wasn't covered, only the text type was explained. The current state of the art does not provide any combination of SHAP on Token and LIME on Image Superpixel, and providing a complete explanation for each prediction for content moderators is missing.

Only Unimodal Detection- Many current systems only look at the text and don't consider the visual parts, which are often already there in news posts. When you add headline and associated image together, the overall message can be much more confusing or misleading than just looking at the image by itself.

Black-Box Nature of DL Models- DL models used for detecting fake content work in a way that's hard to understand.Even though these models can be pretty good at identifying real or fake content, they don't explain which specific words or parts of an image played a role in making their decision.

Motivation for Our Work

1. In today's world of social media, a fake news isn't merely a misleading headline. No, it almost always comes with an image, and that image is basically always misleading at the very least. A system that just reads the text misses half the story.

2. We observe that even though strong transformer-based models like BERT fall short when the deception manifests not through the text but via the picture. An image can tell a completely different story on its own while the headline appears neutral
3. There is nothing rare about that mismatch between text and image, which is how most modern fake news is designed. Therefore, we require a system, which reads both modalities together and not separately.
4. We explored the use of multimodal learning that combined a transformer-based text encoder such as DistilBERT, along with a deep image encoder like ResNet18. Together they can learn to interpret a post as a human fact-checker would.
5. However, it was not merely the accuracy requirement. When an algorithm flags your post as fake, it needs to explain why. Otherwise it offers little assistance to journalists or content moderators. That is where SHAP LIME comes into play as every prediction is auditable and not only correct.

PROPOSED METHODOLOGY

This study proposes a way to combat fake news by assuming the linguistic expertise of the transformer-based text encoders, visual expertise of the deep image encoders and giving an additional layer of explainability to make the predictions interpretable. DistilBERT and ResNet18 are specifically chosen because they have been previously trained on large corpora: DistilBERT was trained on large text data; the ResNet18 model, which was trained on ImageNet, enables each encoder to acquire rich linguistic and visual representations without needing to re-train from scratch.

A) System Architecture

Proposed system architecture is organized into six sequential components:
1. Preprocessing and preparation of the input dataset.
2. Encoding of textual content through DistilBERT.
3. Encoding of visual content through ResNet18
4. Fusion of the two modality-specific feature representations.
5. Training of the classification head on the fused representation.
6. Generation of explanations using SHAP and LIME.

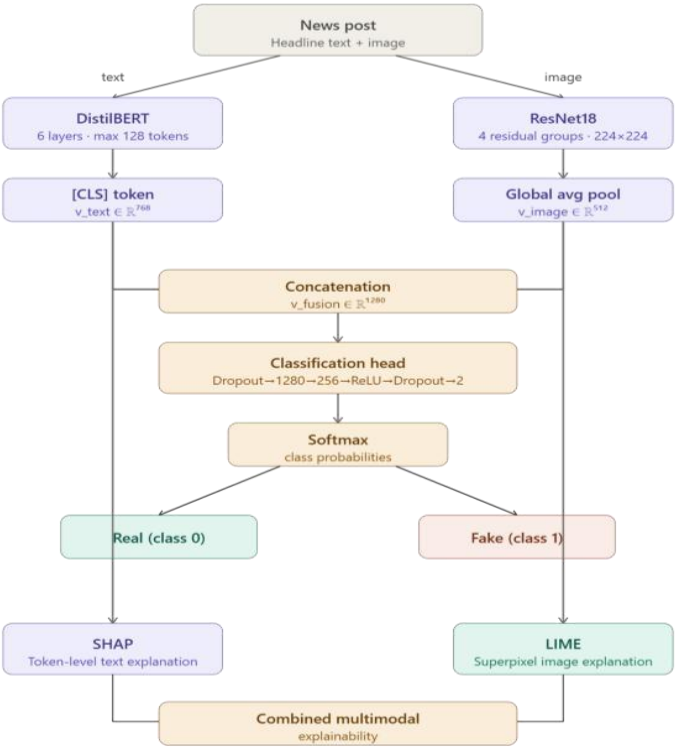

Figure 1: System Architecture

Algorithm 1: Multimodal Fake News Detection with Explainability

Input: Fakeddit Dataset (clean_title ,image_url ,2_waylabel)
Output: Predicted class (ŷ) {Real, Fake}, probability (P), E_t, E_i

```
(X_train, X_val, X_test) ← Split(X, 70/15/15)
for each sample s in X do
    s.text ← Clean(s.title)            ▷ lowercase, strip URLs
    s.img  ← Fetch(s.image_url)        ▷ else blank 224×224 tensor
    s.img  ← Resize+Normalize(s.img)
end for
w_r, w_f ← Class_Weights(y_train)      ▷ 0.3930, 0.6070
F_t ← DistilBERT(s.text)               ▷ 768-dim CLS
F_i ← ResNet18(s.img)                  ▷ 512-dim vector
F   ← Concat(F_t, F_i)                 ▷ 1280-dim fusion
H ← Dropout→Linear(1280,256)→ReLU →Dropout→Linear(256,2)
for epoch e in {1,2,3} do
    θ ← Adam(H, lr=2e-5, loss=WCE(w_r,w_f))
    Clip_Grad(θ, 1.0)
    StepLR(θ, γ=0.5)                   ▷ halves lr/epoch
end for
ŷ,P ← Argmax(Softmax(θ(X_test)))
E_t ← SHAP(θ, bg=50, img=zeros)
E_i ← LIME(θ, n=1500, top_k=8)
return ŷ, P, E_t, E_i
```

B) Dataset Description

Dataset used to detect fake news, known as Fakeddit it contains posts from Reddit . It contains the text from the post titles, and links to the images in the post, and a label showing whether they are real or not. 100,000 records was selected so that some tests could be carried out without harming the data quality or reducing the range of categories available.

C) Architecture Selection and Feature Merging

DistilBERT model isn't designed specifically for detecting fake news. You need to show it the right way to handle the problem, like someone who has read a lot about language would do. Pre-trained DistilBERT is used to get embedding for [CLS] token from final hidden layer, which produces a embedding of dimension 768 representing the headline's text.

D) Classification Head and Training

For taking the final call on whether a post is either Real or Fake, a two-layer classification head based on fullyconnected networks is applied to the 1280 dimension fused representation. This involves

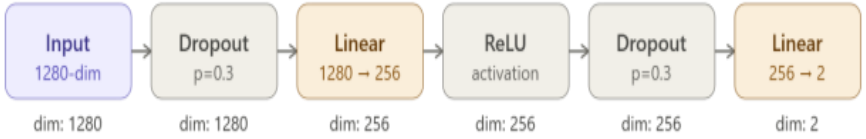

Figure 2 : Classification Head

The data is divided into three groups of data..
Training Set – a collection of data used to teach the model, containing 70,000 examples.

Validation Set – For hyperparameter evaluation (15k).
Test Set – for final evaluation(15k).
For class imbalance issue, model is trained with a weighted CrossEntropyLoss function.The Adam optimizer is used with a lr of 2×10^{-5}, and a StepLR scheduler is applied after every 3 epochs.

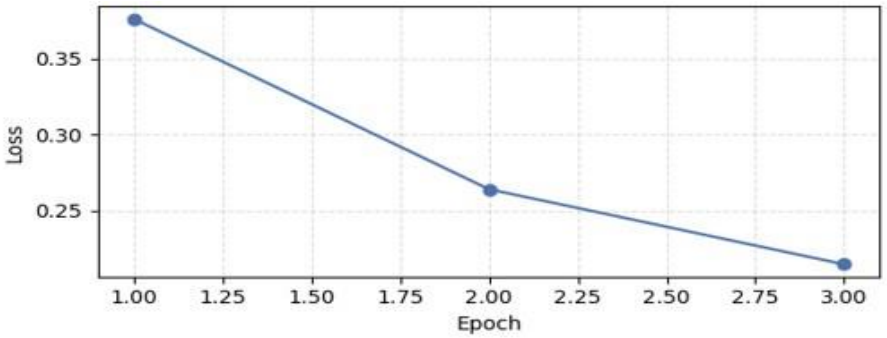

Figure 3 : Training Loss Curve

E) SHAP and LIME Explainability
In real situations, a model that doesn't explain itself clearly and isn't accurate enough doesn't provide much help. That's where SHAP and LIME really help out. In the text part of SHAP, the image get replaced with an empty tensor, and importance of each word is calculated using a method.

Partition Explainer shows which headline word helped determine if the prediction was real or fake. In the LIME process, an image is changed by replacing the text with an empty space and covering different parts of the image made up of small sections called superpixels.This creates 1,500 copies of each image, which helps figure out which image's part are most important for making the classification decision. You can use both methods together to fully explain all the system's predictions by using two different ways of approaching the problem.

F) Performance Metrics
To see how well the model is working for both classes, we look at accuracy, recall, precision,ROC-AUC and F1 score to understand how good the model is performing is explained in equation 1,2,3,4.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \dots 1$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \dots 2$$

$$\text{F1} = 2\frac{\text{Precision}.\,\text{Recall}}{\text{Precision} + \text{Recall}} \quad \dots 3$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FDR})\text{d}(\text{FPR}) \quad \dots 4$$

RESULTS

The studies looked at show that the best working method compared to single-type solutions was the one that used both pictures and words, across different tests and groups of data. Pre-trained transformer models such as BERT and DistilBERT work better than traditional ML approaches like SVMs and earlier DL techniques like artificial neural networks. Systems that do not explain how they make their predictions. In this setup, both the SHAP method using text

and the LIME method using images are easy to understand.The system we suggest combines these two approaches to explain the model's decisions at the level of individual words in text and at the level of small image sections called superpixels.

A) Performance Contrast Among Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	77.35%	0.683	0.805	0.739
CNN-RNN	81.56%	0.802	0.808	0.806
Transformer(BERT)	85.72%	0.862	0.860	0.860
HEMT-Fake (Proposed)	87.33%	0.880	0.870	0.870

B) Analysis Of Proposed HEMT-Fake

The ROC-AUC of HEMT-Fake is 0.9451 on the test dataset, which is 0.1006 and 0.0560 higher than Logistic Regression (0.8445) and CNN-RNN Hybrid (0.8891), respectively. This is an improvement that is threshold independent.

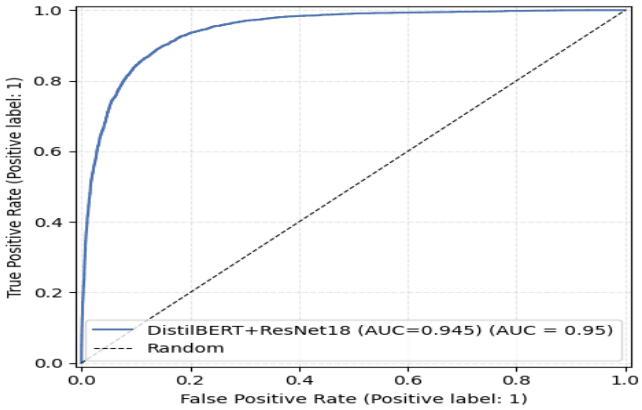

Figure 4: ROC Curve

C) Test Confusion Matrix Description

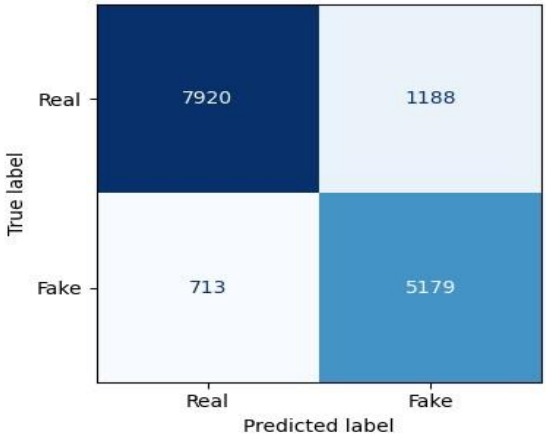

Figure 5 : Confusion Matrix

D) Explainability Analysis

The system uses SHAP to explain text and LIME to explain images, providing clear explanations for both text and image data. For every SHAP analysis, the image input was set to zero to focus only on the text part, and for every LIME analysis, the text input was set to zero to focus only on the image part.

SHAP Results

Plots were made for the test samples by SHAP methodology.

For Sample 1 (True: FAKE — "this Japanese puzzle came with glue to hold it together"). Probability of Fake by model prediction was 0.9937.This prediction was much higher than the baseline because the model correctly identified the words "glue," "puzzle," and "this" as signs that something is Fake.

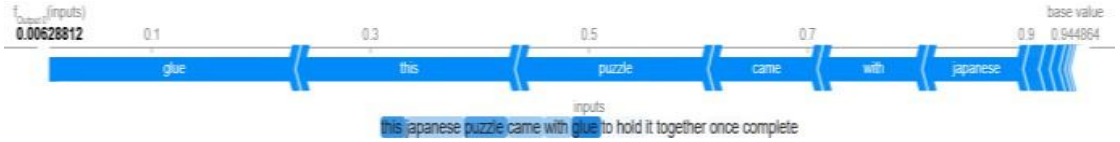

Figure 6 : Shap Example 1

Example 2: (True: REAL — "now i gotta cut loose") with a very low Fake probability of 0.0175. The prediction was pushed towards Real by the tokens "gotta", "loose" and "now".

Figure 7: Shap Example 2

LIME Results

The explanation by LIME was created for images using superpixel, using number of perturbations equal to 1500 .

Example 1:The superpixel boundaries are set against the “Heavenly Home” text overlay in the house image, and the yellow chair in the foreground. The model is able to read a text included as a visual stimulus in the image, thus showing that the ResNet18 image encoder can capture information not only from the visual pattern but also from the embedded text in images. This helps to confirm the Real prediction since the model is correct in predicting the nature of the image.

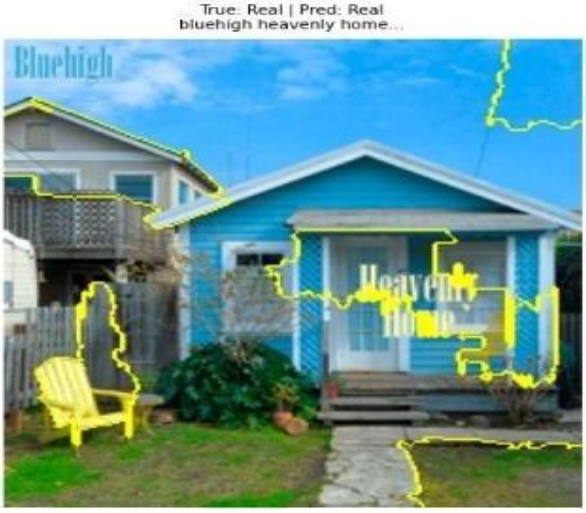

Figure 8: Lime Example 1

Example 2: (True: FAKE — "cat punching another one") showed the superpixel outlines to accurately capture the raised body and paw of both cats, with the superpixel boxes.

Figure 9: Lime Example 2

CONCLUSION AND FUTURE WORKS

A organized review of existing research was done to look at different studies on detecting fake news, including those that use traditional machine learning and manually created features, as well as more advanced methods like deep learning, transformer models, and combining multiple types of data. We look at well-known benchmark datasets such as Fakeddit and check how different detection systems perform based on their type, design, and the dataset they are tested on.

Multimodal systems that learn text and image signals together are consistently better performing than text-only unimodal systems. Currently, DistilBERT or other transformer-based encoders are accepted as the de facto standard for text representation. Nonetheless, the most important conclusion drawn from this analysis is that no existing multimodal fake news detection system provides model post-hoc explanation of both text and image streams simultaneously SHAP & LIME have both been applied to one modality, but never both to the same system.

Future work includes designing a system that uses pre-trained encoders to process both text and images.Then, it applies token-level explanations using SHAP and superpixel-level explanations using LIME within the same structure. This approach is tested on a large benchmark dataset called Fakeddit, using strict splits where the data doesn't overlap. The integration of vision transformer ideas can help in enhanced image representation. We can also explore few-shot detection using pre-trained vision-language models. Finally, we can develop lightweight architectures for real-time deployment on edge devices.

Fake news isn't just about technology issues; it's also a problem that affects society. To build trust with journalists, content moderators, and platform teams that rely on building detection systems, it's important to create systems that are accurate, easy and clear to understand.

REFERENCES

1. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete, "Information credibility on Twitter," *Proceedings of the 20th International Conference on World Wide Web (WWW)*, Hyderabad, India, pp. 675–684, 2011.
2. Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea, "Automatic detection of fake news," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada, pp. 3391–3401, 2017.
3. Xinyi Zhou and Reza Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, 2020.

4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, pp. 4171–4186, 2019.
5. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
6. Yaqing Wang, 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," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, London, UK, pp. 849–857, 2018.
7. Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh, "SpotFake: A multi-modal framework for fake news detection," *Proceedings of the IEEE 5th International Conference on Multimedia Big Data (BigMM)*, Singapore, pp. 39–47, 2019.
8. Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma, "MVAE: Multimodal variational autoencoder for fake news detection," *Proceedings of the World Wide Web Conference (WWW)*, San Francisco, CA, USA, pp. 2915–2921, 2019.
9. Isabel Segura-Bedmar, Marta López de Lacalle, and Aitor Soroa, "Multimodal fake news detection," *Information*, vol. 13, no. 6, p. 284, 2022.
10. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should I trust you? Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, pp. 1135–1144, 2016.
11. Scott M. Lundberg and Su-In Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, pp. 4765–4774, 2017.
12. S. Raj and A. Thakur, "Exploiting content characteristics for explainable detection of fake news," *Information*, vol. 14, no. 10, p. 129, 2023.
13. Alejandro Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
14. Kai Nakamura, Sharon Levy, and William Yang Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, Marseille, France, pp. 6149–6157, 2020.
15. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778, 2016.
16. Rakesh Kumar Kaliyar, Anurag Goswami, and Pratik Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 11765–11779, 2021.
17. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, pp. 5998–6008, 2017.
18. William Yang Wang, "Liar, liar pants on fire: A new benchmark dataset for fake news detection," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada, pp. 422–426, 2017.
19. Natali Ruchansky, Sungyong Seo, and Yan Liu, "CSI: A hybrid deep model for fake news detection," *Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM)*, Singapore, pp. 797–806, 2017.
20. Jiawei Xue, Yifan Wang, Yi Tian, Yiming Li, Lin Shi, and Lin Wei, "Detecting fake news by exploring the consistency of multimodal data," *Information Processing and Management*, vol. 58, no. 5, p. 102610, 2021.