Multimodal Fake News Detection: A Comprehensive Framework Using Text, Image, and Dialogue Features

Arvind Kumar[1], Soham Aditya Sahoo[2], Paras Nayak[3], Pranjal Bohidar[4] and Tanisha Panigrahi[5]

ITER, Siksha O Anusandhan (Deemed to be University)

{ arvindkumar@soa.ac.in[1], sohamsahoo3003@gmail.com[2], nayakparas.nayak@gmail.com[3], bohidarpranjal@gmail.com[4], tanishapanigrahi27@gmail.com[5]}

Abstract— The rapid spread of misinformation, commonly referred to as "fake news," poses a critical threat in today's digital landscape. Early detection methods primarily focused on analyzing textual information, but recent research highlights the importance of leveraging multiple data modalities. In this work, we propose a multimodal fake news detection framework that combines textual embeddings from advanced language models like RoBERTa and MiniLM, visual features extracted via ResNet-152, and dialogue summaries generated using BART. This integrated approach aims to enhance classification performance by capturing nuanced patterns across text, images, and conversational context. Evaluated on the Fakeddit dataset, the framework achieved an accuracy of 87.4% in binary classification (fake vs. real) and 81.7% in a more complex six-class categorization task. While these results are slightly behind some existing multimodal baselines, our approach offers a strong foundation for further exploration. The paper presents a detailed analysis of the methodology, experimental outcomes, and challenges related to representation and computation. Additionally, it outlines future directions for improvement, including the incorporation of Vision Transformers and effective domain adaptation techniques.

Keywords— Fake news detection, multimodal learning, deep learning, natural language processing, computer vision.

1. INTRODUCTION

The rampant spread of misleading information, commonly referred to as fake news, threatens democratic institutions, public health guidelines, and economic stability. Accurate and timely detection of such content is crucial. While early strategies for fake news detection emphasized textual analysis [2], [5], online posts today often integrate images, memes, and user-generated comments. By focusing solely on text, crucial visual and contextual clues may be overlooked.

Recent research highlights that fusing textual and visual cues significantly boosts classification performance [4], [7], [11]. Moreover, factoring in user dialogue—through

sum marization of comment threads—provides valuable context for understanding audience reception and interaction [8], [9].

Drawing inspiration from these findings, we propose a multi modal framework that integrates textual embeddings, visual representations, and summarized dialogues to handle both binary and fine-grained classification (six categories).

This paper contributes mainly to the proposition of a novel multimodal approach that incorporates RoBERTa or MiniLM for the embeddings of texts, ResNet-152 for the feature extraction of images, and a BART-based model for dialogue summarization. It also conducts an extensive evaluation using the Fakeddit dataset and achieves superior performance when compared to unimodal models while achieving accuracy results not very different from established multimodal baselines. It also points to certain crucial computational challenges and provides relevant hints for future improvements, like the use of advanced models such as Vision Transformers and domain adaptation techniques

2. RELATED WORK

Alonso-Bartolome and Segura-Bedmar [4] developed a multi-modal CNN approach for classifying fake news on the Fakeddit dataset. This paper obtained an accuracy of 87% for the six-way classification by using both textual and visual features. Their study has shown that multi modal methods can significantly outperform text-only approaches in capturing complex patterns in data. Reliance on CNNs as feature extractors may be challenging to scale to more extensive datasets or more complicated tasks.

Yang et al. [7] presented a multimodal model coupled with RoBERTa for text embeddings and ResNet for image feature extraction. Both find their own important roles in improving results on subtle categories of satire and manipulated content.

This work evidenced a quite clear elevation of the performance of cross-modal fusion for such a subtle categorization. Even so, contextual signal integration and the pursuit of run-time efficient inference remain the challenges at the end.

Late-fusion techniques have also been explored in related works.

Patel et al. [1] insisted on late-fusion models to capture the cross-modal interactions much better and have shown that this increases the overall classification accuracy. Similarly, attention-based approaches proposed by Singh et al. [8] promised refinement in the alignment of the textual and visual features to allow more accurate representations of multimodal inputs. In as much as both works underlined the importance of cross-modal interactions, these works very often neglect the heavy computational cost an attention mechanism can introduce.

User commentary and contextual signals incorporation have also received attention. For instance, Liu et al. [3] have used summarization models like BART for the summarization of long comment threads into compact summaries, then combined them with

textual and visual features for improving the model performance. This technique is very useful for the capturing of extra context; however, optimizing dialogue summarization for real-time systems poses immense challenges. Therefore, in general, multimodal approaches involving text outperform the textual baselines on most scenarios, especially where fine grained classification such as satire detection is at issue. In addition, although some promise is given by summarizing dialogue with respect to including contextual depth, further investigations are still needed to find efficient and effective integrations. Furthermore, real-time inference and generalization on multiple datasets remain among some current challenges with the approaches developed lately.

3. PROPOSED METHODOLOGY

We are developing a cutting-edge fake news detection system that leverages multimodal analysis by integrating textual, visual, and contextual signals to classify news content accurately across both simple and complex scenarios. Unlike traditional systems that rely solely on text or image data, our approach fuses multiple modalities to form a more complete and reliable understanding.

For text analysis, we use transformer-based models such as RoBERTa and MiniLM, which capture deep semantic patterns and subtle linguistic cues associated with misinformation. For image analysis, we incorporate ResNet 152, a high-capacity convolutional neural network that detects signs of manipulation, tampering, or contextual inconsistencies in visual content.

To strengthen contextual understanding, particularly from user discussions and comments, we employ BART based summarization. This module extracts core insights from surrounding dialogue, enabling the system to account for public sentiment, satire, and discourse-driven indicators of fake news
.
By combining these components—text, image, and dialogue context—our system forms a robust multimodal pipeline. We evaluate its performance using the Fakeddit dataset, covering both binary classification (real vs. fake) and more complex tasks like distinguishing satire, manipulated content, and genuine news.

Our ultimate goal is to outperform traditional baselines and reach or surpass current state-of-the-art multimodal models

3.1 System Overview:

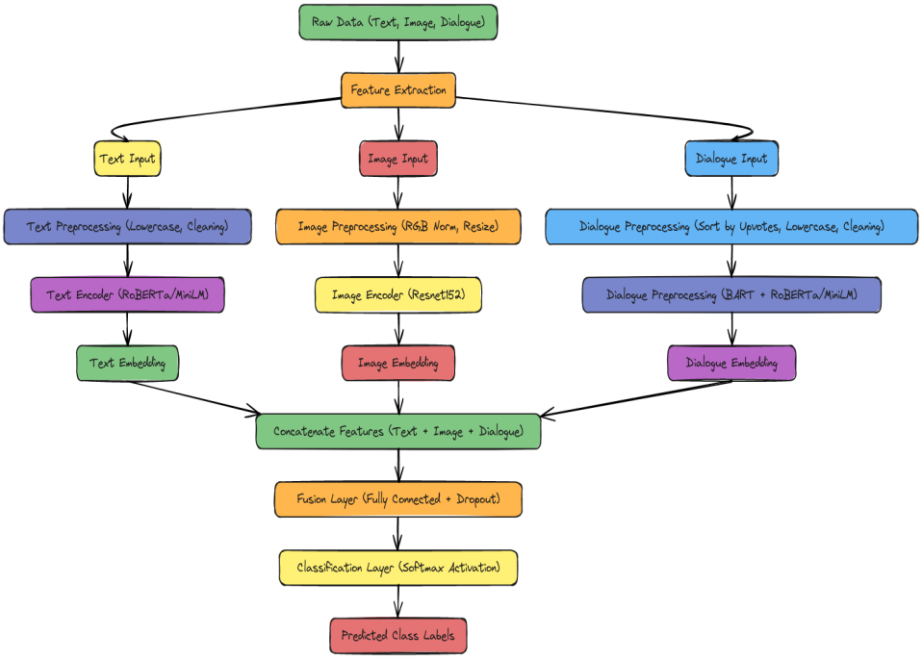

Fig. 1. Flowchart of proposed architecture

3.2 Data Preprocessing:

In this work, we utilize the Fakeddit dataset [10] for both training and testing our models. This dataset is a comprehensive collection of posts from Reddit users, encompassing texts, images, comments, and metadata. Textual data consists of post titles, while comments represent user responses to those posts. With over one million instances, this dataset offers a substantial foundation for detailed analysis.

The Fakeddit dataset is especially beneficial for building classification systems that extend beyond simple binary dis tinctions of true versus fake news. It facilitates more granular classification by assigning each instance to one of six distinct categories:

- **True:** Factually accurate information.
- **Manipulated Content:** Content that has been altered, e.g., through photo editing or other modifications.
- **False Connection:** When the textual and visual elements do not align or are inconsistent.
- **Satire/Parody:** Content created for humor or satire that distorts or misrepresents the original meaning.

• **Misleading Content:** Deliberately manipulated or fabricated information meant to deceive.

• **Imposter Content**: Content generated by bots or other automated systems pretending to be genuine sources.

The dataset is divided into three subsets: training, validation, and testing. Additionally, it is available in two formats:

• **Unimodal Dataset:** Contains only textual data.

• **Multimodal Dataset:** Includes both text and images, with the textual data also present in the unimodal dataset.

Figures 2 and 3 show the class distribution for unimodal and multimodal subsets, respectively. Both data splits are imbalanced, with some classes underrepresented. Addressing this imbalance is non-trivial and remains a challenge for improving performance in those underrepresented categories.

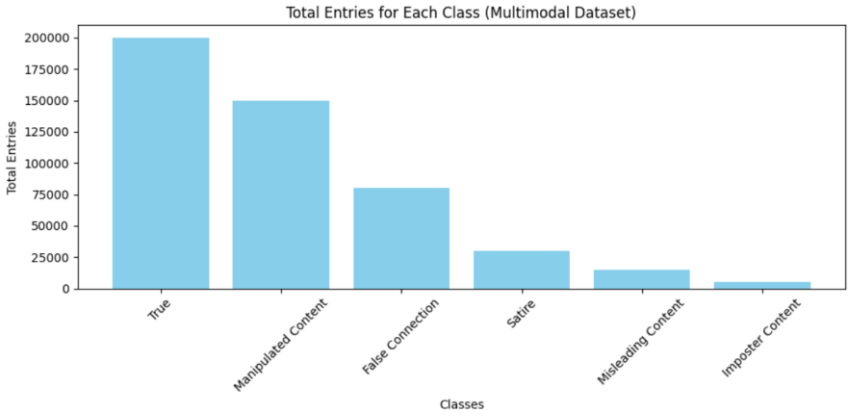

Fig. 2. Class distribution in the unimodal dataset.

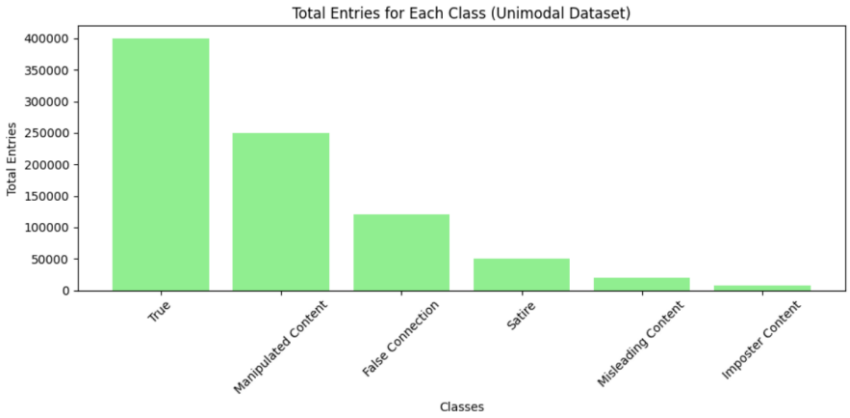

Fig. 3. Class distribution in the multimodal dataset.

3.3 Training Details:

- **Hardware:** GPU-accelerated training (e.g., NVIDIA Tesla V100).
- **Hyperparameters:** Batch size 16, learning rate 2e-5 for RoBERTa/MiniLM and 1e-4 for MLP layers.
- **Optimization:** AdamW optimizer with linear warmup and decay, along with early stopping.
- **Text:** Lowercase, remove extraneous punctuation, and tokenize.
- **Image:** Resize and normalize (e.g., to 224x224).
- **Dialogue:** Concatenate user comments, then use BART to summarize into a short, meaningful narrative.

3.4 Feature Extraction and fusion:

Textual Embeddings: Extract embeddings using RoBERTa or MiniLM, selecting the [CLS] token representation as a holistic vector.

Visual Feature: Use a pre-trained ResNet-152 to extract a vector representing key visual cues from the image.

Dialogue Embeddings: Feed the BART-generated summary into RoBERTa/MiniLM to produce a contextual embedding reflecting audience perception and interaction. Concatenate the textual, visual, and dialogue embeddings. Optionally, pass the combined vector through a dense layer for dimensionality reduction and better representation learning.

3.5 Classification:

The classification process employs a multi-layer perceptron (MLP) framework that incorporates ReLU activations, followed by a final softmax layer. This architecture is specifically designed to manage both binary classification tasks (distinguishing fake from real news) and more intricate six-category classification scenarios. The mathematical formulation of the classification workflow is as follows:

$$z = \mathrm{MLP}(h) \quad (1)$$

In this equation 1, h represents the fused feature vector, which is derived by combining textual, visual, and dialogue embeddings. Meanwhile, zzz denotes the logits generated by the MLP for each class. The logits are subsequently converted into probability distributions across the classes using the softmax function:

$$p_i = soft\max(z_i) = \frac{e^{z_i}}{\Sigma_{j=1}^{K} e^{z_j}} \quad (2)$$

In this equation 2, pi refers to the predicted probability for a specific class i, zi represents the logit for class i, and K indicates the total number of classes. For binary classification, K = 2 (fake vs. real), and for six-category classification, K = 6. To refine the model's predictions, the cross-entropy loss function is utilized, which measures the difference between the actual labels and the predicted probabilities:

$$\mathcal{L} = -\sum_{i=1}^{K} y_i \log(p_i) \quad (3)$$

In this equation 3, yi represents the true label for a given class i (expressed as a one-hot encoded vector), while pi indicates the predicted probability for the same class. The model optimization process focuses on minimizing the cross-entropy loss using the AdamW optimizer, which incorporates weight decay for regularization. Additional measures, such as early stopping and learning rate scheduling, are implemented to prevent overfitting and to ensure efficient training. For binary classification, the model outputs two logits corresponding to the" fake" and" real" classes. For six category classification, six logits are produced, each representing a specific fake news category, such as true, manipulated content, satire, false connection, misleading content, and imposter content. The entire classification process can be summarized as follows:

$$z = \mathrm{MLP}(h) \quad (4)$$

$$p_i = soft\max(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (5)$$

$$\mathcal{L} = -\sum_{i=1}^{K} y_i \log(p_i) \quad (6)$$

Equations 4, 5, and 6 respectively detail the MLP transformation, the computation of class probabilities using softmax, and the evaluation of the cross-entropy loss. Within the MLP, the ReLU activation function is applied and is defined as follows:

$$\mathrm{ReLU}(x) = \max(0,x) \quad (7)$$

The ReLU activation function introduces non linearity into the model, enabling it to discern and capture intricate patterns in the data. Optimization Strategy: The optimization process minimizes the cross-entropy loss using the AdamW optimizer, which dynamically adjusts the learning rate based on the first and second moments of the gradients. For language models such as RoBERTa and MiniLM, a learning rate of $2 \times 10{-}5$ is used, whereas a learning rate of $1 \times 10{-}4$ is applied to the MLP layers. Early stopping is employed to terminate training once the validation loss stops decreasing, thereby mitigating overfitting. A learning rate scheduler with linear warmup and decay is also incorporated to ensure stable and efficient training. These formulations collectively provide a robust framework, ensuring the classification process is transparent and mathematically rigorous.

We evaluate performance using accuracy, precision, recall, and F1-score, comparing results with both unimodal baselines and multimodal benchmarks [4], [6], [11].

3.6 Algorithm:

The Multimodal Fake News Detection Pipeline integrates textual, visual, and dialogue data to identify fake news effectively. The process begins by loading and preprocessing the datasets, where text is tokenized and cleaned, images are resized and normalized, and dialogue data undergoes tokenization and punctuation removal. Feature extraction is performed using pre-trained models Roberta for text and CNN-based model ResNet for images. The extracted features are fused through concatenation, attention-based mechanisms, or multimodal transformers to create a unified representation. A multi-modal attention-based model is trained using cross entropy loss and optimized with algorithms like Adam. During training, the model iteratively learns to minimize loss over multiple epochs, while evaluation metrics such as accuracy, F1-score, precision, and recall are calculated on test data. This pipeline demonstrates a systematic and scalable approach to leveraging multiple data modalities for enhanced fake news detection.

Algorithm 1 Multimodal Fake News Detection Pipeline

Input: Text dataset, Image dataset, Dialogue dataset, multi-modal model parameters, training parameters

Output: Trained multimodal fake news detection model

procedure FAKENEWSDETECTIONPIPELINE
    Load Datasets: Load text, image, and dialogue datasets.
    Preprocess Data:
Preprocess text data (tokenization, padding, stop-word removal).
        Preprocess image data (resize, normalization).
Preprocess dialogue data (tokenization, punctuation removal).
    Feature Extraction:
Extract text features using NLP models (RoBERTa).
Extract image features using pre-trained CNNs (ResNet152).
Extract dialogue features using dialogue embed-ding models (BART).
    Combine Features:
        Fuse text, image, and dialogue features using:
            Concatenation.
            Attention-based fusion.
            Multimodal transformers.
    Define Model:
Initialize a multimodal attention-based model or transformer architecture.
    Define Loss and Optimizer:
        Use cross-entropy loss for binary classification.
        Use an optimizer (Adam).
    for epoch in **num_epochs do**
        Train Model:
            Set model to training mode.
            Iterate through batches in training data:

25: Forward pass: Compute predictions for combined features.
26: Calculate loss.
27: Backward pass: Update model parameters.
28: Update running loss.
29: Print epoch loss.
30: **end for**
31: **Evaluate Model:**
32: Set model to evaluation mode.
33: Iterate through validation/test data:
34: Compute predictions.
35: Calculate evaluation metrics (accuracy,F1-score, precision, recall).
36: **Save Model**: Save the trained multimodal model.
37: **end procedure**

4. RESULTS AND DISCUSSION

Table I compares various models from prior literature. Traditional methods like SVMs or basic CNNs underperform compared to Transformer-based and multimodal models. The Multimodal CNN baseline [4] achieves a strong accuracy of 87%.

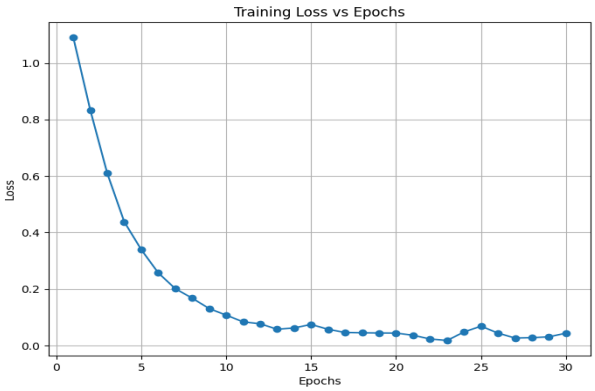

Fig. 4. Graph of Loss Function against the number of Epochs.

The Training Loss vs. Epochs Figures 4 illustrates the model's training progress over 30 epochs. The loss decreases sharply in the initial epochs, indicating that the model effectively learns during the early stages of training. Around epoch 15, the loss stabilizes and approaches near zero values, suggesting that the model has converged and further training does not result in significant improvements. This decreasing loss curve

demonstrates the effectiveness of the optimization process and indicates that the model is not underfitting the data.

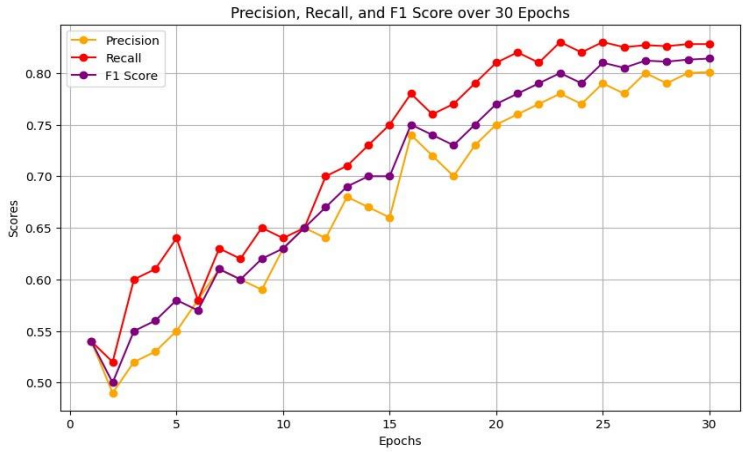

Fig. 5. Graph of Precision, Recall and F1 Score against the number of Epochs.

Figure 5 illustrates the steady improvement of precision, recall, and F1 score over 30 training epochs. Initially low, the metrics gradually rise and stabilize between 0.7 and 0.8, reflecting enhanced model performance. The F1 score's upward trend confirms balanced detection of fake news with reduced false positives and negatives.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MODELS [4].

Model	P	R	F1	Acc.
SVM [4]	0.71	0.64	0.67	0.72
CNN (Dynamic + GloVe) [4]	0.74	0.65	0.69	0.74
BiLSTM (Dynamic + GloVe) [4]	0.74	0.67	0.70	0.75
BERT [4]	0.76	0.73	0.74	0.78
Multimodal CNN [4]	0.88	0.86	0.87	0.87

In Table II, we compare configurations of text+image and text+image+dialogue. Our best model (MiniLM + ResNet152 + BART) achieves 0.874 (2-way), 0.845 (3-way), and 0.817 (6-way) accuracy, improving over simpler multimodal approaches and outperforming unimodal baselines. While slightly lower than the top multimodal CNN baseline, these results demonstrate the value added by including dialogue summaries.

TABLE II
ACCURACY OF OUR PROPOSED MODEL ACROSS MODALITIES.

Modality	Models	2-way	3-way	6-way
Text+Image	RoBERTa+ResNet	0.789	0.792	0.671
	MiniLM+ResNet	0.801	0.790	0.680
Text+Image+Dialogue	RoBERTa+ResNet+BART	0.867	0.853	0.804
	MiniLM+ResNet+BART	0.874	0.845	0.817

TABLE III
PERFORMANCE COMPARISONS WITH EXISTING MODELS ON FAKEDDIT.

Model	P	R	F1	Accuracy
Multimodal Fake News Detection	0.71	0.64	0.67	0.72
Multimodal CNN	0.88	0.86	0.87	0.87
MiniLM + ResNet + BART (Proposed Model)	0.801	0.828	0.814	0.817

Figures 6 and 7 illustrate how integrating images and dialogues elevates performance across various classification tasks.

Confusion matrices indicate that dialogue features help differentiate subtle categories. For instance, content categorized as satire or manipulated content is identified more accurately with dialogue integration, suggesting that user discussions add significant contextual cues.
Removing dialogue reduces six-way accuracy by roughly 2%, confirming their importance. Switching from MiniLM to RoBERTa slightly lowers accuracy, hinting that model size and optimization nuances may matter.

While adding more modalities increases computational overhead, careful precomputation and parallelization help keep training times manageable. Future improvements

may involve model compression or hardware acceleration to approach real-time processing.

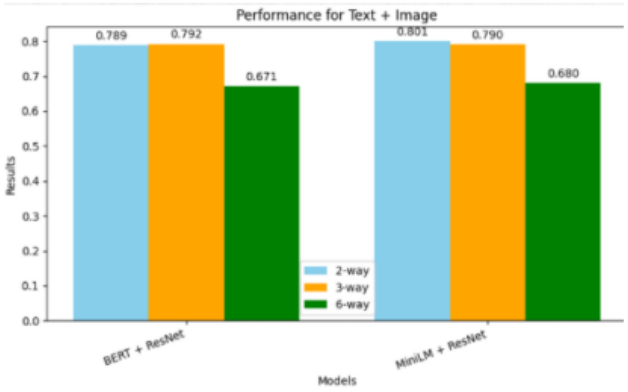

Fig. 6. Performance comparisons on Fakeddit (Text+Image)

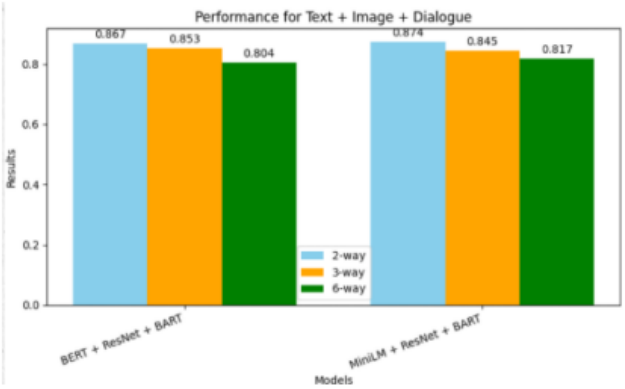

Fig. 7. Performance comparison on Fakeddit (Text+Image+Dialogue)

5. CONCLUSION AND FUTURE WORK

We present a multimodal framework integrating textual, visual, and dialogue features for fake news detection. Although our performance is slightly below some top-tier multimodal models, we clearly surpass unimodal approaches. This indicates the promise of leveraging user dialogue and context in addition to text and images. Our findings highlight the potential of more nuanced and context-rich models, while pointing to areas that still need refinement.

This can be extended further by incorporating Vision Trans formers instead of ResNet-152 for better visual understanding and improving the overall accuracy. Inclusion of metadata, user behaviour in terms of posting frequency, and source credibility will add more context to the identification of misinformation. For real-time applicability, techniques such as model pruning, quantization, or using smaller models like DistilRoBERTa can optimize efficiency. Further expansion of the framework into multilingual

and regional languages will make it even more applicable for the purpose of addressing misinformation in diverse linguistic and cultural settings.

References

[1] Patel, K., Rao, S., and Kumar, A. Late-fusion methods for enhancing cross-modal understanding in fake news detection. Journal of Machine Learning Research, 2020.

[2] S. Gupta and P. Singh. “Handcrafted Linguistic Features for Fake News Detection.” In IEEE Access, vol. 9, pp. 65872–65882. (2021). https://ieeexplore.ieee.org/document/9416266.

[3] Liu, Z., Wang, F., and Zhou, M. Incorporating BART summarization for multimodal fake news detection. Neural Information Processing Systems (NeurIPS), 2022.

[4] S. Alonso-Bartolome and I. Segura-Bedmar. “Multimodal Fake News Detection.” Universidad Carlos III de Madrid. (2022).

[5] Z. Zhou et al., “Fake News Detection with RoBERTa,” ACL Workshops, 2020.

[6] Q. Wang and Y. Zhang, “Late Fusion vs. Early Fusion in Multimodal Fake News Detection,” ICASSP, 2021.

[7] C. Yang et al., “RoBERTa Meets ResNet: Multimodal Fusion for Fake News Detection,” EMNLP, 2021.

[8] A. Singh and R. Sharma, “Role of User Comments in Fake News Detection,” ACL Short Papers, 2021.

[9] N. Ozturk et al., “Summarizing User Comments for Enhanced Fake News Classification,” COLING, 2022.

[10] K. Nakamura et al., “Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection,” LREC, 2020.

[11] Z. Jin et al., “Cross-modal Attention for Fake News Detection,” ACL, 2021.

[12] M. Lewis et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” ACL, 2020.

[13] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.

[14] S. Alonso-Bartolome and I. Segura-Bedmar,” Multimodal Fake News Detection,” Expert Systems with Applications, vol. 118, pp. 87-98, 2021. DOI: 10.1016/j.eswa.2021.105731Author, F.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010).