FMD-Mllama at the Financial Misinformation Detection Challenge Task: Multimodal Reasoning and Evidence Generation

Zheyang Luo, Guangbin Zhang, Jiahao Xiao, Xuankang Zhang, Yulin Dou, Jiangming Liu*
Yunnan University
jiangmingliu@ynu.edu.cn

Abstract

This paper presents our system for the Financial Misinformation Detection Challenge Task. We utilize multimodal reasoning, incorporating textual and image information, to address the task. Our system demonstrates the capability to detect financial misinformation while providing comprehensive explanations. Experimental results show that our final system significantly outperforms the baselines and ranks second on the task leaderboard.

1 Introduction

Misinformation is widespread in the financial domain, posing a significant challenge for professionals in the finance industry. Detecting false financial information is crucial for maintaining trust and stability in financial markets. Financial information appears in various forms, including text, images, and videos. Relying on data from a single form is insufficient to capture financial misinformation effectively.

In this paper, we introduce multimodal reasoning method for the Financial Misinformation Detection Challenge Task (Liu et al., 2024). Our approach leverages both image and textual information to address the task. The final system, FMD-Mllama, achieves a score of 79.24 in the shared task and ranks second on the leaderboard.

2 Related Work

2.1 Misinformation Detection

In the field of fake news detection, various models have been employed to tackle misinformation. These models can be broadly categorized into three types: neural network models, pre-trained models, and large language models. For neural network models, Jian et al. (2024) detect media misinformation using Bi-LSTM, while Raja et al. (2022) propose a quantum multimodal fusion-based approach for fake news detection. For pre-trained models, Boissonneault and Hensen (2024) utilize BERT and SKEP to detect fake reviews, and Lu et al. (2023) investigate the effectiveness of models like M-BERT and BERT in detecting fake news. For large language models, Ma et al. (2024) employ GPT-3.5 and Llama2 to construct heterogeneous graphs of news through specific prompts to improve fake news detection. Additionally, Qu et al. (2024) explore the capabilities of ChatGPT and Gemini models for fake news detection using the LIAR dataset (Wang, 2017).

2.2 Multimodal Deep Learning Models

Multimodal models have demonstrated significant potential in tackling complex tasks. These models include CLIP (Radford et al., 2021), Florence (Yuan et al., 2021), LXMERT (Tan and Bansal, 2019), Llama 3.2-Vision (Dubey et al., 2024), GPT-4V (Yang et al., 2023), and KOSMOS-1 (Huang et al., 2023), among others. For the task of detecting financial misinformation, we utilize the FMDID dataset (Liu et al., 2024). To achieve this, we fine-tune the Llama 3.2-11B-Vision-Instruct model.

2.3 Chain of Thought

Chain of Thought techniques (CoT; Wei et al., 2022) are increasingly used to improve model transparency and reasoning quality. Recent studies on fine-tuning with CoT have shown promising results in enhancing model performance. Ho et al. (2022) leverage the capabilities of large models to generate CoT explanations, using these generated CoTs as targets for fine-tuning smaller models. Similarly, Zelikman et al. (2022) employ models to generate both answers and corresponding CoTs. CoTs associated with correct answers are then used as prompts to fine-tune the model.

*Corresponding author

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 277–282
January 19–20, 2025.

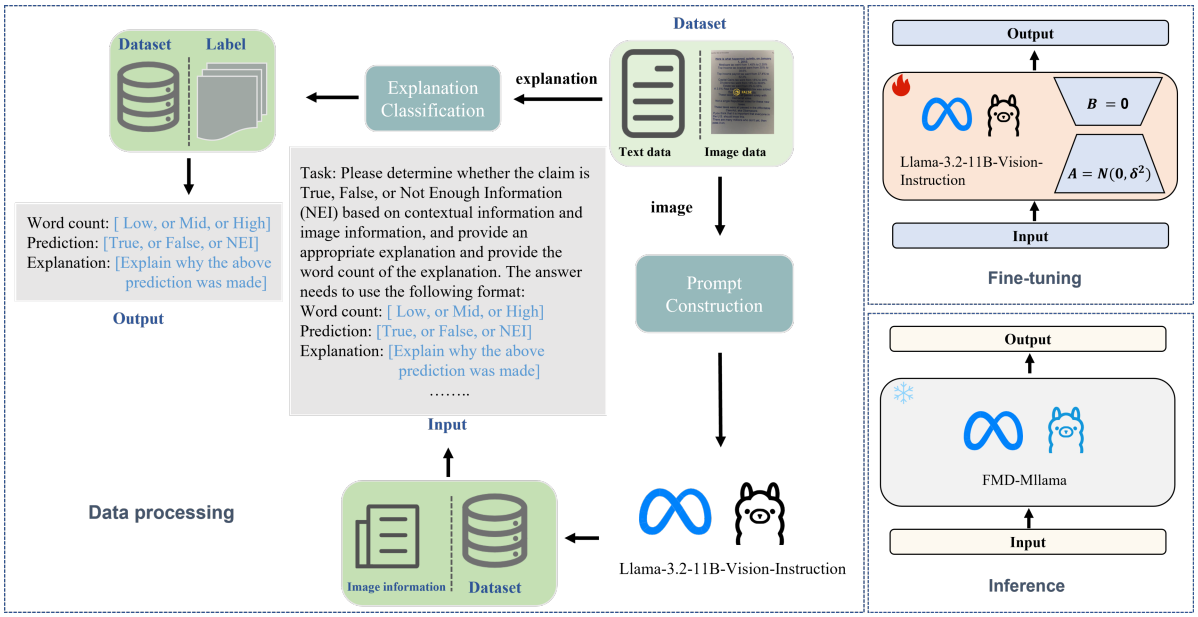

Figure 1: FMD-Mllama.

3 Methods

3.1 Task Formalization

Following the settings of the Financial Misinformation Detection Challenge Task, we aim to train a model that estimates the conditional probability $P_\phi(y \mid x)$, where x represents the given input, such as claims, claim summaries and image links, and the output y corresponds to the judgment category: *True*, *False*, or *Not Enough Information*. Here, *True* indicates the model judges the claim to be true, *False* indicates the claim is judged to be false, and *Not Enough Information* indicates the model finds insufficient information to make a judgment, along with the corresponding explanations.

3.2 FMD-Mllama

Our system consists of data processing, fine-tuning and inference, as shown in Figure 1.

Data Processing The ground truth exhibits significant variation in explanation lengths. We expect the model to learn to generate not only the explanation but also the length of the explanation. We propose classifying explanations by length as additional model outputs. The length distribution is presented in Table 1, categorized into three groups.

To use the set of images provided in dataset, we select the image most relevant to the news content from the available set. We design specialized prompts to enable the Llama 3.2-Vision model to effectively choose the most relevant image and convert it into corresponding textual descriptions, as outlined in Figure 2. These descriptions include *image description*, *contextual information*, and *relevant details*. The *image description* provides basic information about the image, *contextual information* relates to details derived from both the image and the textual content, and *relevant details* include text and image-related information, as shown in Figure 3.

Category	Range	Count
Low	$(0, 151)$	606
Mid	$[151, 286)$	607
High	$[286, \infty)$	607

Table 1: Distribution of lengths over the explanation in training data

{"role": "user", "content": {"type": "image"}, {"type": "text", "text":
Claim: claim_content
Select the image most relevant to the claim provided above and provide a 512-word summary based on that image.
If only one image is available, provide a 512-word summary of that image's content in relation to the claim.
In your summary, please address the following points:
1. Image Description: A clear description of the image's subject matter.
2. Contextual Information: Explain how the selected image relates to the claim.
3. Relevant Details: Include any additional information that enhances understanding of the image and the claim.
Please format your response as follows:
Summary: [the summary will be here].}

Figure 2: The prompt of generating the image information

Model	micro-F1	ROUGE-1	ROUGE-2	ROUGE-L	overall score
FMDllama	71.82	45.02	34.64	37.43	58.42
ChatGPT(gpt-3.5-turbo)	70.12	26.14	09.94	16.32	48.13
FMD-Mllama	**79.55**	**78.92**	**75.17**	**76.63**	**79.24**

Table 2: The final results (%) of our model and the baselines, where the best results are bold.

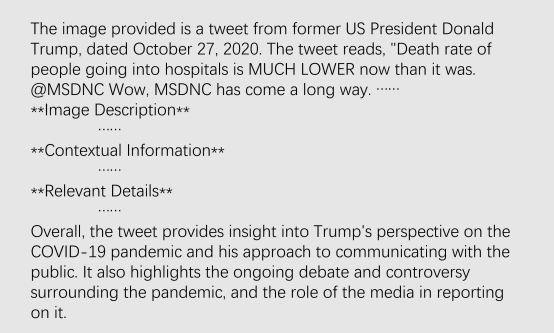

Figure 3: The generating image textual information.

Model	micro-F1	ROUGE-1	overall score
Text	**83.52**	69.28	76.4
Text-image	82.12	70.27	76.2
Text-textual image	83.24	**71.72**	**77.4**

Table 3: The results (%) of the ablation study, where text is the model fine-tuned with textual information, text-image is the model fine-tuned with textual information and image information, text-textual image is the model fine-tuned with textual information and textual image information.

Fine-tuning We fine-tune the Llama-3.2-11B-Vision-Instruct model on the processed data using LoRA (Hu et al., 2021). A specially designed instruction is incorporated, prompting the model to generate three components in its response: the classification of the explanation length, the judgment, and the corresponding explanations, shown in Figure 5. This instruction aims to help the model not only learn the relationship between the input and output but also internalize the required response format. The dual learning objective ensures the model produces outputs that are both contextually relevant and consistently formatted.

Inference Ensuring consistency between the structure and format of the test and training datasets is crucial. This includes aligning the organization of input features, the format of the instructions, and the structure of the expected outputs. By maintaining this consistency, we can evaluate the model under conditions similar to those during training, leading to a more reliable and accurate assessment of its performance.

4 Experiments

4.1 Settings

We use LoRA to fine-tune the models, with a rank of 8, allowing for low-rank decomposition and efficient parameter updates. The scaling factor is set to 32 to maintain an appropriate balance between the pre-trained weights and the LoRA updates. A dropout rate of 0.1 is applied to prevent overfitting during training. The model is trained with a learning rate of 1×10^{-4} over 5 epochs, using a cosine learning rate scheduler and a weight decay of 0.01.

4.2 Metrics

Micro-F1 is used to evaluate the performance of the classification task, while ROUGE-1 is employed to evaluate the performance of explanation generation. The final system performance is evaluated by taking the average of these two metrics.

4.3 Baselines and Results

We take two baselines: FMDllama (Liu et al., 2024) and ChatGPT(gpt-3.5-turbo) provided by the task.

As shown in Table 2, FMD-Mllama significantly outperforms both baseline models across all evaluation metrics, including micro-F1, ROUGE-1, ROUGE-2, ROUGE-L, and overall score. FMD-Mllama achieves a micro-F1 score of 79.55, which is 7.73 points higher than FMDllama and 9.43 points higher than ChatGPT. It also achieves a ROUGE-1 score of 78.92, which is 33.9 points higher than FMDllama and 52.78 points higher than ChatGPT, and a ROUGE-2 score of 75.17, which is 38.53 points higher than FMDllama and 65.23 points higher than ChatGPT. Additionally, FMD-Mllama achieves a ROUGE-L score of 76.63, which is 39.20 points higher than FMDllama and 60.31 points higher than ChatGPT. Finally, FMD-Mllama attains an overall score of 79.24, which is 20.82 points higher than FMDllama and 31.29 points higher than ChatGPT.

Model	micro-F1	ROUGE-1	ROUGE-2	ROUGE-L	overall score
CoT-FMD-Mllama(batch 4)	70.37	68.29	42.07	44.68	69.33
CoT-FMD-Mllama(batch 32)	75.25	50.42	42.07	44.68	62.83
FMD-Mllama	**79.55**	**78.92**	**75.17**	**76.63**	**79.24**

Table 4: The results (%) of CoT fine-tuning.

4.4 Ablation Study

To investigate the role of image information in the task, we conduct ablations with three different data types: textual information, textual information combined with image information, and textual information with both textual and image-related details. Due to the blinded test data, we split the original training dataset into training and test sets to perform these ablation experiments.

As shown in Table 3, the model fine-tuned with both textual information and image-related details achieves the highest ROUGE-1 score and the highest overall score. While this model attains a lower micro-F1 score in judgments, it achieves higher ROUGE scores in explanation generation. This suggests that additional image-related textual information can enhance the model's ability to generate explanations, but it does not improve the model's judgment accuracy.

However, we interestingly find that the model fine-tuned with both textual information and image-related textual details achieves a higher micro-F1 score and overall score than the model fine-tuned with only textual information and image information. This suggests that the model benefits more from the additional textual image information than from the image information alone.

4.5 Discussion on CoT Fine-tuning

We follow the approach outlined in (Ho et al., 2022) to introduce CoT fine-tuning based on FMD-Mllama, resulting in a system referred to as CoT-FMD-Mllama. The configuration for the CoT fine-tuning experiment is the same as that used in the ablation study. CoT-FMD-Mllama is trained with batch sizes of 4 and 32 to evaluate the impact of batch size on performance, while all other hyperparameters remain consistent with FMD-Mllama. The results are shown in Table 4. We design specialized prompts for GPT-4o to generate the CoT based on the processed input and output. The generated CoT is then added to the output for fine-tuning the model. The prompt provided to GPT-4o to generate the CoT as shown in Figure 4. The input and output of CoT-FMD-Mllama are different from FMD-Mllama, the difference shown in Figure 5 and Figure 6.

```
msgs = [
    {"role": "system", "content": "You are a reasoning expert,
please provide a detailed reasoning process."},
    {"role": "user", "content": f"Task: Based on the provided
contextual information and image information, generate a
reasoning process (Chain of Thought) that leads to the known
answer. The Claim is the question and the Claim summaries
include more information about the claim.
Claim: {claim}
Claim summaries: {claim_summaries}\n\nContextual Information:
{contextual_info}
Image Information: {image_info}
Known Answer: {answer}
Please provide the reasoning process step-by-step to arrive at the
above answer."}
```

Figure 4: The prompt provided to GPT-4o to generate the CoT. The prompt requests the model using the content of "user" to generate the reason process.

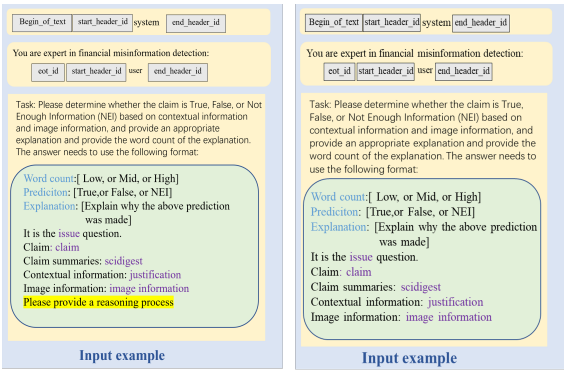

Figure 5: The input of CoT-FMD-Mllama and FMD-Mllama. The left is the input of CoT-FMD-Mllama, the right is the input of FMD-Mllama. The highlight is the difference between the two models.

The results present a performance comparison between FMD-Mllama and CoT-FMD-Mllama. CoT-FMD-Mllama (batch size 4) achieves a micro-F1 score of 70.37, which is 9.18 points lower than FMD-Mllama, and an overall score of 69.33, which is 9.91 points lower than FMD-Mllama. CoT-FMD-Mllama (batch size 32) achieves a micro-F1 score of 75.25, which is 4.3 points lower than FMD-Mllama, and an overall score of 62.83, which is 16.41 points lower than FMD-Mllama.

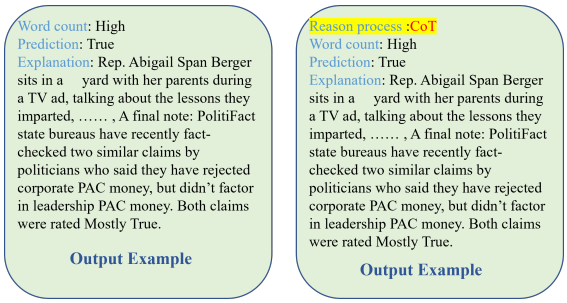

Figure 6: The output of CoT-FMD-Mllama and FMD-Mllama. The left is the input of FMD-Mllama, the right is the input of CoT-FMDMllama. The highlight is the difference between the two models.

According to the experiments on CoT, we conclude that the CoT fine-tuning decreases the overall effectiveness of the model. One possible reason is that the CoT fine-tuning increases the model's complexity, as it must not only generate judgments and explanations but also generate the CoT, which raises the difficulty of the generation task. The Financial Misinformation Detection Challenge includes both judgment and explanation tasks, and the CoT fine-tuning further complexity to these tasks. Additionally, batch size impacts the performance of CoT-FMD-Mllama. As the batch size increases, the micro-F1 score improves, but the ROUGE-1 score decreases. This suggests that with larger batch sizes, the model may shift its focus towards generating the CoT, which could negatively impact judgment and explanation generation.

More strategies are needed to refine CoT fine-tuning, enabling the model to enhance its reasoning ability while staying focused on the task at hand, without being adversely affected by the need to generate the CoT.

5 Conclusion

We introduce multimodal approaches that significantly enhance the performance of the model for the Financial Misinformation Detection Task. Our final system achieves an overall score of 79.24, significantly outperforming the two baselines provided by the shared task, respectively. Additionally, the simple adoption of CoT fine-tuning can actually harm the model's performance.

Acknowledgment

We thank reviewers for their suggestions and the data provided by the organizers of the shared task. This work is supported by Yunnan Fundamental Research Projects (grant NO. 202401CF070189).

References

David Boissonneault and Emily Hensen. 2024. Fake news detection with large language models on the liar dataset.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

Wang Jian, Jian Ping Li, Muhammad Atif Akbar, Amin Ul Haq, Shakir Khan, Reemiah Muneer Alotaibi, and Saad Abdullah Alajlan. 2024. Sa-bi-lstm: Self attention with bi-directional lstm-based intelligent model for accurate fake news detection to ensured information integrity on social media platforms. *IEEE Access*, 12:48436–48452.

Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Fmdllama: Financial misinformation detection based on large language models. *Preprint*, arXiv:2409.16452.

Junwen Lu, Xintao Zhan, Guanfeng Liu, Xinrong Zhan, and Xiaolong Deng. 2023. Bstc: A fake review detection model based on a pre-trained language model and convolutional neural network. *Electronics*, 12(10):2165.

Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521, Miami, Florida, USA. Association for Computational Linguistics.

Zhiguo Qu, Yunyi Meng, Ghulam Muhammad, and Prayag Tiwari. 2024. Qmfnd: A quantum multimodal fusion-based fake news detection model for social media. *Information Fusion*, 104:102172.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2022. Fake news detection in dravidian languages using transformer models. In *International Conference on Computer Vision, High-Performance Computing, Smart Devices, and Networks*, pages 515–523. Springer.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.