GMU-MU at the Financial Misinformation Detection Challenge Task: Exploring LLMs for Financial Claim Verification

Alphaeus Dmonte[1], Roland Oruche[2], Marcos Zampieri[1], Eunmi Ko[3], Prasad Calyam[2]
[1]George Mason University, USA
[2]University of Missouri-Columbia, USA
[3]Rochester Institute of Technology, USA
admonte@gmu.edu

Abstract

This paper describes the team GMU-MU submission to the Financial Misinformation Detection challenge. The goal of this challenge is to identify financial misinformation and generate explanations justifying the predictions by developing or adapting LLMs. The participants were provided with a dataset of financial claims that were categorized into six financial domain categories. We experiment with the Llama model using two approaches; instruction-tuning the model with the training dataset, and a prompting approach that directly evaluates the off-the-shelf model. Our best system was placed 5th among the 12 systems, achieving an overall evaluation score of 0.6682.

1 Introduction

With the widespread use of social media, the spread of false and misleading information has been on the rise. This includes information in domains such as politics, healthcare, finance among others. In the financial domain, data shared through social media channels is made widely available through the web impacting important business decisions, financial policies, etc. This data can ultimately also affect financial markets. Hence, it is essential to check the accuracy of such information. Given the sheer volume of information on the web, it is not feasible to manually check and evaluate potentially inaccurate information and claims. Hence, automated approaches for misinformation detection and claim verification are required to identify and mitigate the spread of false and inaccurate information.

Several approaches have been proposed over the years for automatic claim verification including traditional machine-learning models, as well as more recent deep-learning models (Wang, 2017). Models such as BERT (Devlin et al., 2019) have shown state-of-the-art performance in accurately identifying fake news and misinformation (Kaliyar et al., 2021). The recent emergence of Large Language Models (LLMs) has shown exceptional abilities in several NLP tasks. In the financial domain, these models have been used for several applications including sentiment analysis, entity recognition, and summarization among others (Nie et al., 2024). LLMs have been employed for misinformation detection and automated claim verification with several techniques like in-context learning, fine-tuning, retrieval augmented generation, etc (Dmonte et al., 2024a; Chen and Shu, 2024). However, most of these approaches have been evaluated on general-domain datasets and the financial misinformation detection and claim verification using LLMs is underexplored.

A typical claim verification pipeline consists of identifying the claim, retrieving evidence, rationale selection, veracity label prediction, and explanation generation. This challenge focuses on the last two components of the claim verification pipeline. Given the claim and the associated evidence, the objective is to use LLMs to verify if a claim is *True, False* or there is *Not Enough Evidence*, and provide explanations for the predicted label considering the associated evidence. We employ two approaches for this task; instruction-tuning and prompting an LLM.

2 Related Work

Several approaches for automatic claim verification have been proposed over the years. These include traditional machine-learning approaches like Logistic Regression, SVM, etc, and deep learning models like LSTMs (Wang, 2017). However, these approaches do not consider contextual dependencies within the text. Models like BERT (Devlin et al., 2019) that consider contextual dependencies within the text have been shown to outperform the previous approaches (Soleimani et al., 2019). With the exceptional capabilities of LLMs in several NLP tasks, these models have recently been explored

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 308–312
January 19–20, 2025.

claim	justification	label	evidence
When John Kasich became governor of Ohio, there...	Hoping to add some political muscle to Republic...	True	In his endorsement speech, Schwarzenegger called...
Did a Twitter Ad Show Rebel Wilson During Her C...	On Dec. 20, 2020, the person who controlled the...	False	On Dec. 20, 2020, the person who controlled the...
'Unidentified Flying Object' Seen as SpaceX Roc...	On the morning of 1 September 2016 a SpaceX Fal...	NEI	On the morning of 1 September 2016 a SpaceX Fal...
We have the most productive workers in the world.	On the third night of the Democratic convention...	True	When the OECDcompares the GDP per hour workedac...

Table 1: Example instances from the dataset. Only the fields used in the experiments are shown here.

for claim understanding and verification (Dmonte et al., 2024b,a). Several approaches like in-context learning, fine-tuning, retrieval augmented generation (RAG), etc. have been explored for the task. For example, Zhang and Gao (2023) evaluate the LLMs in a few-shot setting and introduce a hierarchical prompting approach, showing improved performance over supervised training approaches. While Chiang et al. (2024) fine-tuned LLMs for multi-stage claim verification.

In the financial domain, several approaches for fake news, misinformation, and disinformation detection have been proposed. These include traditional machine learning and deep learning models like SVM, LSTM, CNN, etc (Zhi et al., 2021), and transformer-based models like BERT (Zhang et al., 2022; Mohankumar et al., 2023) and RoBERTa (Kamal et al., 2023; Rangapur et al., 2023). However, the task of financial claim verification is underexplored. More recently, Rangapur et al. (2023) introduced a dataset for multimodal financial claim verification. They experimented with several approaches including models like RoBERTa (Liu, 2019) and LLMs like GPT-4 (Achiam et al., 2023), Claude 3 (Anthropic, 2024), etc. Liu et al. (2024) fine-tune LLMs for the task. Our work aims to advance financial claim verification efforts by investigating approaches to evaluate open-source LLMs for this task.

3 Experiments

3.1 Datasets

We utilize the Fin-Fact (Rangapur et al., 2023) dataset provided by the COLING-2025-FMD. The dataset consists of financial claims related to income, tax, economy, budget, finance, and debt. The instances were extracted from PolitiFact, Snopes, and FactCheck, which are online platforms for fact-checking. The training data consists of 1,953 instances, while the test dataset consists of 1,303 instances. We further split the training set into a train-validation set with an 80:20 split. The following fields are included in the dataset.

- **Claim:** the core assertion.
- **Posted Date:** temporal context.
- **Sci-Digest:** claim summaries.
- **Justification:** contextual information offering insights into the claim's accuracy.
- **Issues:** the domain of the claim.
- **Image Data:** visual information.
- **Label:** the veracity label of the claim which can be *True*, *False*, or *Not Enough Information*.
- **Evidence:** the ground truth explanations.

The training dataset includes all the fields, while the *label* and *evidence* fields are not included in the test dataset. For our experiments, we use only the *claim, justification, label,* and *evidence* fields. Table 1 shows the example instances from the dataset.

Instruction:
Given the input claim and the corresponding evidence, determine if the claim is True, False, or Not Enough Information (NEI). Please provide an explanation justifying the prediction.
Input:
Claim: {claim}
Evidence: {context}
Response:

Figure 1: The prompt used to instruction-tune the model.

3.2 Implementation Details

We experiment with the following two approaches.

Instruction Tuning We fine-tune Llama-3.1-8B (Dubey et al., 2024) model with the training dataset. The *claim* and *justification* columns were used as input to the model. Figure 1 shows the instruction prompt used to fine-tune the model consisting of the task-specific instruction as well as the input claim and associated evidence.

Table 2 shows the hyperparameter values used to fine-tune the model.

Parameter	Value
epochs	10
batch size	8
learning rate	1e-4
max grad norm	1.0
gradient accumulation steps	2

Table 2: The hyperparameter values used to fine-tune the LLM.

Prompting We use a few-shot prompt to evaluate the performance of the off-the-shelf model. The prompt instruction includes the steps to be executed to verify the claim against the associated context and generate an explanation. We first ask the model to identify the main assertion or claim spans from both the claim and the associated context. The model should then compare these identified text spans and generate a veracity label. Finally, the model should provide a justification for the predicted label while considering the claim and the associated context. The claim and the evidence, which serve as additional context to the model are given as input.

We provide three examples from the training dataset to further enhance the model's ability to perform this task. Figure 2 shows the detailed prompt used in our experiments.

4 Results and Discussion

Table 3 and 4 show the performance of our approaches on the test dataset compared to the top-3 teams and the baseline models. The test dataset was divided into a public and private split, where the performance of the approaches on the private split served as an official leaderboard for the challenge. On the public split, our instruction-tuning approach was ranked eighth and achieved an overall score of 0.7026, with an F1 score of 0.8299 and a ROUGE-1 score of 0.5752, outperforming both the baselines. In comparison, our prompting approach underperformed baseline 1 but outperformed baseline 2. An overall score of 0.5831 was achieved with this approach, with an F1 score and ROUGE-1 scores of 0.7468 and 0.4194, respectively. Similar to the performance on the public split, our instruction-tuned model outperformed both the baselines and was ranked fifth, with overall, F1, and ROUGE-1 scores of 0.6682, 0.7575, and 0.5789, respectively. The prompting approach achieved an overall score of 0.5495, while the F1 and ROUGE-1 scores were 0.6802 and 0.4187, respectively, outperforming baseline 2 while having a score closer to baseline 1.

We analyze the predictions of our approaches to understand the lower performance of our approaches compared to the top three teams. We observe that for both approaches, the Llama 3 model tends to generate inconsistent labels, especially if there is not enough information to make a prediction. In this case, the model either assigns a random True or False label, or it outputs *mixture* indicating neither true nor false. We also observe that in some instances, the model generates incomplete explanations. This can be attributed to the maximum new tokens hyperparameter, which decides the maximum number of new tokens generated. We also observe that, in some instances the explanations generated contain repetitions, suggesting the lower ROUGE scores of our approaches compared to the top three teams. To assess if the few-shot prompt followed the instruction steps for prediction, we randomly select a few instances and output the model's reasoning steps. We observe that the model considers the intermediate instruction steps when making the prediction. Furthermore, the lower performance of the model can also be attributed to the model generality. Since the Llama 3 model was trained on general domain data, it may be unable to understand domain-specific jargon resulting in inconsistencies while analyzing the claim and evidence. Our approaches use only the textual data to verify the claims. Incorporating image data as well as other meta-data can further enhance the model performance, as such data provides valuable information that can aid claim verification.

5 Conclusion

This paper presents our submission to the financial misinformation detection challenge. We use two different approaches to evaluate the LLMs. Results indicate that the models are able to predict

The task is to analyze the claim and the associated evidence and predict if the claim is False, True, or there is Not Enough Information, and provide a justification. Please follow these steps:

1. Identify the main claim span or assertion span from the input claim:
- For the given input claim, extract the exact text span mentioning the main claim or assertion.
- This can be a sub-text or the entire input text.

2. Identify the main claim span or assertion span from the input evidence:
- From the associated input evidence, extract the main assertion or claim span if any.
- There can be multiple claims or assertions in the evidence.

3. Make a prediction based on the claim/assertion spans:
- Consider the claim/assertion span extracted in step 1 and the claim/assertion spans extracted in step 2.
- Based on these spans, verify if the claim is True False, or there is Not Enough Information to verify.
- Label should be only one of the following: False, True, Not Enough Evidence.

4. Provide a justification explaining your prediction. Consider the claim and associated evidence when providing the justification.

Examples:
{examples}

Output Format:
Predicted Label: [your-label-prediction-here]
Justification: [your-justification-here]

Input:
Claim: {claim}
Evidence: {evidence}

Response:

Figure 2: The few-shot prompt used in our experiments. The prompt instruction include the steps to be performed for verifying the claim.

Rank	Team Name	Overall Score	Micro-F1	ROUGE-1	ROUGE-2	ROUGE-L
1	Dunamu ML	0.8492	0.8946	0.8038	0.7773	0.7879
2	TFinAI	0.8338	0.8688	0.7988	0.7682	0.7805
3	GGbond	0.8102	0.8503	0.7701	0.7302	0.7448
8	**GMU-MU**	**0.7026**	**0.8299**	**0.5752**	**0.4956**	**0.5137**
Baseline-1	FMDLlama	0.6089	0.7616	0.4563	0.3536	0.3817
15	**GMU-MU***	**0.5831**	**0.7468**	**0.4195**	**0.2726**	**0.3122**
Baseline-2	ChatGPT	0.5152	0.7634	0.267	0.102	0.1662

Table 3: Model performance on the public split. Our system performances are in bold. GMU-MU* represents our prompting approach, while the other is the instruction-tuned model performance.

Rank	Team Name	Overall Score	Micro-F1	ROUGE-1	ROUGE-2	ROUGE-L
1	Dunamu ML	0.8294	0.8467	0.8121	0.7873	0.7969
2	GGbond	0.7924	0.7955	0.7892	0.7517	0.7663
3	1-800-SHARED-TASKS	0.7768	0.8283	0.7253	0.6763	0.6911
5	**GMU-MU**	**0.6682**	**0.7575**	**0.5789**	**0.4956**	**0.5145**
Baseline-1	FMDLlama	0.5842	0.7182	0.4502	0.3464	0.3743
15	**GMU-MU***	**0.5495**	**0.6802**	**0.4187**	**0.2773**	**0.3122**
Baseline-2	ChatGPT	0.4813	0.7012	0.2614	0.0994	0.1632

Table 4: Model performance on the private split. The scores in bold represent the scores for our instruction-tuned model.

the veracity of the claims more precisely compared to generating the explanations. Furthermore, fine-tuning LLMs on the task outperforms the prompting approach. The generality of these models may also affect their performance. For future work, we would like to analyze the impact of few-shot examples. We further plan to use domain-specific LLMs. We also plan to explore multimodal models with the additional data fields, as the inclusion of the images along with the textual data can help improve the performance of the task.

Acknowledgments

We would like to thank the competition organizers for providing participants with this interesting dataset.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.

Shang-Hsuan Chiang, Ming-Chih Lo, Lin-Wei Chao, and Wen-Chih Peng. 2024. Team trifecta at factify5wqa: Setting the standard in fact verification with fine-tuning. *arXiv preprint arXiv:2403.10281*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024a. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Alphaeus Dmonte, Marcos Zampieri, Kevin Lybarger, Massimiliano Albanese, and Genya Coulter. 2024b. Classifying human-generated and ai-generated election claims in social media. *arXiv preprint arXiv:2404.16116*.

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*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Ashraf Kamal, Padmapriya Mohankumar, and Vishal Kumar Singh. 2023. Financial misinformation detection via roberta and multi-channel networks. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 646–653. Springer.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

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

Padmapriya Mohankumar, Ashraf Kamal, Vishal Kumar Singh, and Amrish Satish. 2023. Financial fake news detection via context-aware embedding and sequential representation using cross-joint networks. In *2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pages 780–784. IEEE.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2023. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *arXiv preprint arXiv:2309.08793*.

Amir Soleimani, Christof Monz, and Marcel Worring. 2019. Bert for evidence retrieval and claim verification. *arXiv preprint arXiv:1910.02655*.

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

Xiaohui Zhang, Qianzhou Du, and Zhongju Zhang. 2022. A theory-driven machine learning system for financial disinformation detection. *Production and Operations Management*, 31(8):3160–3179.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Xiaofan Zhi, Li Xue, Wengang Zhi, Ziye Li, Bo Zhao, Yanzhen Wang, and Zhen Shen. 2021. Financial fake news detection with multi fact cnn-lstm model. In *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, pages 1338–1341. IEEE.