UNIVERSITY POLITEHNICA OF BUCHAREST
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

Scientific Report #2

Network-aware Fake News mitigation on social media
Artificial Intelligence

Octavian Mitrică

Scientific advisors:

S.L. Dr. Ing. Ciprian-Octavian Truică
Conf. Dr. Ing. Elena-Simona Apostol

BUCHAREST
2024

CONTENTS

1 INTRODUCTION

In recent times, online social networks have been rapidly expanding and serve as a major source of global and local news for people all over the world. This comes with both benefits and drawbacks, offering users seamless communication and instant access to information but also facilitating the spread of fake news. The task of detecting fake news or unreliable sources is still a complex problem [1] and with social media platforms like Twitter spreading false information six times faster than the truth [2] there comes the urgent need for AI methods that can fact-check this kind of issues. This problems are further aggravated by visual information since news stories that include images spread faster than those containing only text [3].

Fact-checking websites like PolitiFact and Snopes rely on professionals to manually verify the information. However, the fast spread of misinformation in the digital age makes these manual efforts slow and difficult to scale effectively [4]. Therefore, automated solutions have become a necessity and there have been many discoveries in this field. The most effective researches revolve around NLP techniques [5], LSTMs [6], BERT [7], and their derivatives. Recently, the research area expanded to multimodal solutions, taking into account text and visual information. Hybrid models with BERT and image transformers provide the best results [8] [9]. Supervised learning methods for fake news detection have proven effective at identifying misinformation, but they rely on human-annotated data for training and could become costly in practice.

Large Language Models (LLM) received praise and attention by effectively solving various problems, providing accurate responses even in NLP tasks [10], [11], [12], [13] thus motivating researchers to explore their potential in fact-checking and fake news detection. In this research report, we will present experiments on proven prompt engineering ideas with a personal twist and showcase our best-performing combination of techniques. The focus will be to build a multimodal fake news detector LLM that outperforms supervised learning methods.

The main objectives for this report are the following:

- *Experiment:* Adjust different tactics of prompt engineering to work on our multimodal dataset and analyze the results.
- *Build:* Combine the prompting techniques that produce the best results and fit them for multimodal fake news detection.
- *Analyze:* Inspect the answers provided by the LLM and formulate the next steps.

This Scientific Report is structured on 5 **chapters**. Chapter 2 presents a survey and introduction to relevant work and state-of-the-art. Chapter 3 showcases the proposed solution and tactics used for prompting an LLM for predictions, as well as the datasets used for experiments.

Chapter 4 ranks the obtained results and explains the architecture and setup details. Finally, chapter 5 sums up all the gathered information, limitations, and further work.

2 RELATED WORK

In this chapter, we present the previous approaches and research work on prompt engineering tactics and their results on the task of fake news detection, split into two categories of input: textual information, and textual and visual information - multimodal.

2.1 Textual Information

This section showcases experiments based on determining the veracity of a news article or social media post by prompting an LLM with a constructed prompt based on textual information found in the article or post such as title, author, or the location (domain) of said item. Some of these tactics prove to be more efficient than supervised learning methods that require training and computing resources and can be more cost-effective when implemented at large scale.

Article [14] serves as the prime candidate for introduction to the capabilities of large language models. The authors prove that only using an LLM such as GPT 3.5 for fake news detection does not outperform the top, fine-tuned BERT. The gap stems from the LLM's inability to properly integrate rationales to conclude the final answer. However, they have also shown that this reasoning provided by the LLMs can be a good advisor for small language models yielding results that outperform the other baseline methods. Their experiments with the baseline tactics are a great way to understand the processing behind the LLM's rationale. These experiments include techniques like zero-shot prompting, zero-shot chain-of-thought prompting, few-shot prompting, and few-shot chain-of-thought prompting that we will discuss later, in chapter 4.

Zhang et al. [10] explored a step-by-step type of prompting in order to separate a claim into several subclaims and then verify them via progressive question answering. One of the fundamental problems of LLMs is hallucination, and the authors tackle this issue by providing real-time contextual information. The LLM is prompted to construct questions for the subclaims that will later be searched on the web. The findings are then fed back to the LLM. After every subclaim has been verified or fact-checked with up-to-date information, the LLM is ready to output the final prediction. Their method proved state-of-the-art results when it was published.

In this article [15], the authors introduced a method to detect rumors on social media by leveraging the comments that each post received. They designed strict prompts to make LLMs focus on the right clues in both the news and the comments and achieved better than state-of-the-art results. This came with limitations as well, since large structural data is hard to be well described in text, leaving space for future exploration.

This paper [16], presents a unique way of leveraging NLP techniques in order to make a final prediction. It splits the claim similar to the method used in [10] and aligns each resulted subclaim with its response (given by another LLM). The final answer is decided using logical operations to invalidate or support the given claim.

Li et al. [17] published another collection of experiments similar to the ones documented in [14]. Their contribution lies in a new way of gathering contextual information by using specific tools and leveraging the step-by-step method to extract the final answer. Authors call it an agentic approach since each tool has its own very specific task. One example is the phrase tool which looks at the title of the claim and decides whether it uses sensational teasers or provocative language. This method is very powerful because it looks at other textual data, not only the title, such as the domain, writing style, commonsense, or even political standing. After all of the tools give their respective observations, the final prompt is formed by joining the claim along with all of the observations and is fed to another LLM that uses a step-by-step workflow to give the final answer. The results outperform all of the other baselines tested in the paper.

2.2 Textual and Visual Information (Multimodal)

This section presents research and experiments done by prompting the LLM with both text and image. These tactics are very useful in our context of detecting fake news in social media posts and give the LLM more room for reasoning. Limitations also arise when the prompt gets increasingly longer and the large language model seems to lose the initial context sometimes.

Wu et al. [18] introduced a model that explores the multimodal domain in the context of LLMs. The authors came up with a method that asks the LLM to describe an image using text. The textual description of the image and the caption used by the original author are then fed to another LLM for analysis. If the two captions relate to the same event then it is considered real, otherwise, it is labeled as fake. Furthermore, they used this image description as a feature extractor and passed it to an ensemble classifier, AdaBoost. By using LLMs as feature extractors and combining them with machine learning techniques, they proved an accurate and stable representation of coherence between image-caption triplets and improved baseline models by a large margin.

This article [19], presents a unique way of leveraging the image context and up-to-date web searching for an accurate prediction of fake news in the domain of politics. With the use of GPT-4V, they input a claim and an associated image into the LLM and ask it to formulate one or more questions about the context. These questions will be piped into another LLM that searches the web and gathers the answers for them. Finally, the claim, along with the image and answered questions are fed to the deciding LLM, which makes an informed evaluation. Limitations here include different question generation, different search results, based on the search engine, and the cost of using such a refined LLM model.

3 PROPOSED SOLUTION

In this chapter, we discuss and summarize the objectives mentioned before and how our experiments will be shaped around them. We give a brief explanation of the thought process, datasets used, and the final solution.

The task is to classify a social media post containing some text data and one image into fake news (labeled as 0) or real news (labeled as 1).

To get a good understanding of how well the prompt engineering tactics work on our proposed dataset and task we propose a set of experiments, utilizing both the typical approaches and the ones presented by researchers mentioned above. These will be the topics of discussion:

Zero-Shot Prompting. The prompt is constructed using only the task description and the provided news (or social media post in our case). To enhance the response quality and reduce the likelihood of refusals, we can incorporate a role-playing approach when explaining the task.

Zero-Shot CoT Prompting [11]. In addition to the standard zero-shot prompt, we encourage the model to use chain-of-thought reasoning by simply adding the sentence "Let's think step by step." at the end of it. We forward the post and reasoning to another LLM to get the final prediction.

Few-Shot Prompting [13]. This method provides the LLM with a few examples in the form of post-label as a demonstration for rationale. We do tests with 2, 4, and 8 examples. The experiments show that as the number of examples increases, so does the accuracy.

Few-Shot CoT Prompting [12]. Similar to the zero-shot CoT, we provide the LLM with the respective examples and ask it to go step-by-step providing a rationale. We forward the post, examples, and rationale to another LLM for final assessment.

Zero-Shot Agent Prompting [17]. Leverages our specially designed tools (defined in section 4.1) to get individual information about pieces of the post. The post and the reviews are passed to another model for the final answer.

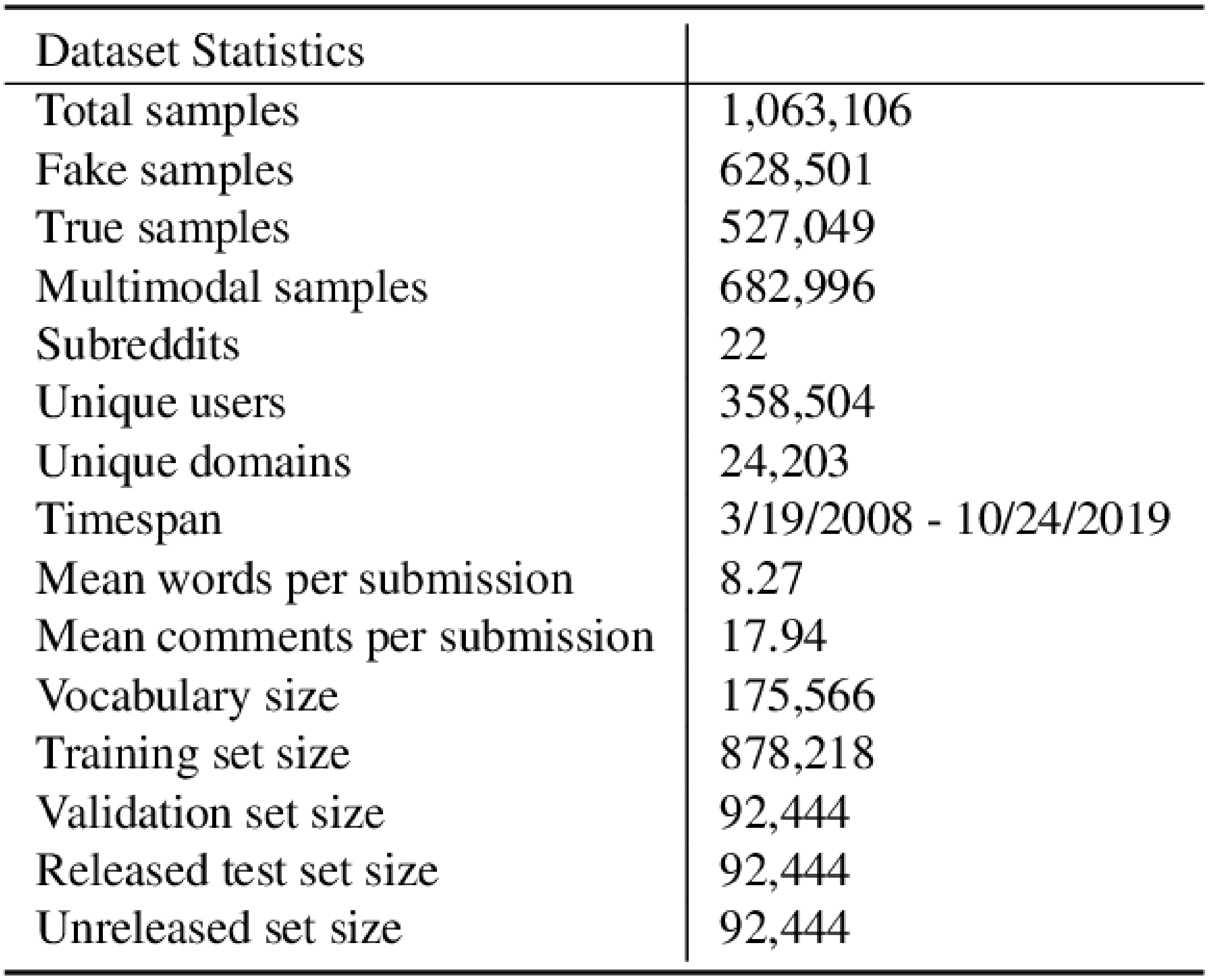
Zero-Shot Tools with RAG Prompting. This is a method we created by combining the RAG method [19] and the tools defined in [17]. This tactic asks a model for an analysis of the image and then another model is given the title of the post along with the review of the image. This second LLM will formulate a question about the given context that will help with categorizing the post. The question is then answered by the third LLM, which uses a web search engine to find relevant information. The fourth and final model assesses the post, image, and question-answer pair and gives its final prediction. This method proved to be less efficient than expected when used against social media posts. Further experiments with actual news headlines may provide more interesting results.

We will formulate our own prompts for each experiment and compile the baseline results in chapter 4. Based on the final findings, we construct a new approach that yields better results than the supervised learning baselines presented in the last research report, namely DistilBERT + ViT.

Our final solution utilizes a combination of the few-shot method [13], the chain-of-thought method [12], and the FactAgent method [17].

3.1 Proposed architecture

We propose a **Few-Shot Agen Prompting** method that leverages tools similar to the ones proposed by Li et al. [17] and the power of example that comes with few-shot prompting. Figure 1 shows a representation of the process.



Dataset Statistics	
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 - 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

Figure 1: Our proposed pipeline

3.2 Current implementation

Our implementation will make a total of 4 calls to an LLM for output:

- Title tool: a tool designed to determine if a social media post's title contains sensational teasers, provocative or emotionally charged language, or exaggerated claims to grab attention from readers or to suggest a compilation of rumors.
- Author tool: a tool that provides insight into the author's username and the type of content posted on the respective subreddit that contains the post (can it be trustful?).
- Image tool: a tool that analyzes a given image, trying to determine if it presents signs of alteration or meme-like characteristics or if it describes real events.
- Final assessment: takes all of the information given in the prompt and predicts the veracity of the post.

The final prompt is built after all of the tools have responded, as follows:

- 2, 4, or 8 examples in the form post-reasoning-label
- Output of Title tool, Author tool, and Image tool

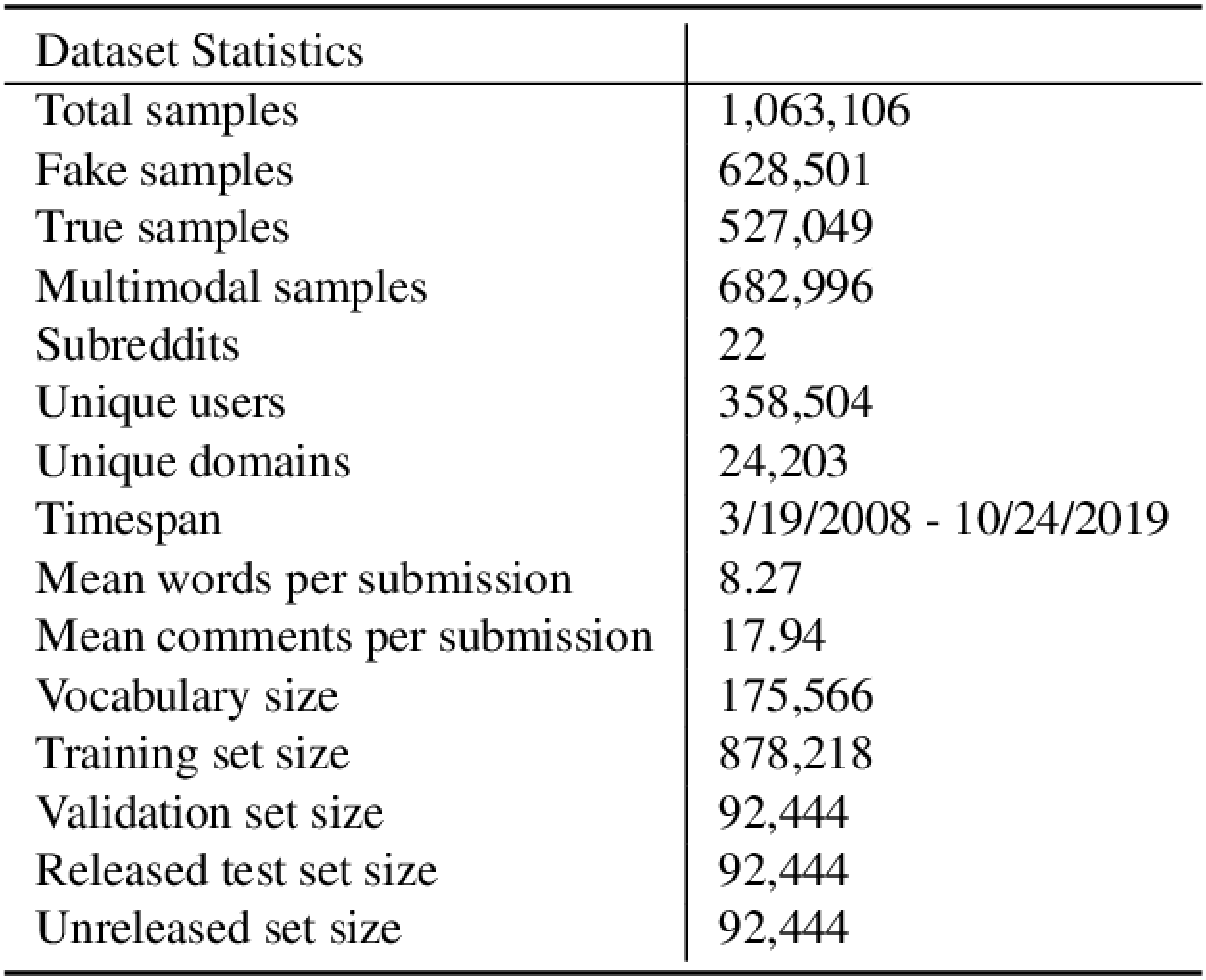
3.3 Dataset

The dataset we mainly focused on with our experiments is Fakeddit [20]. As mentioned before, it is comprised of over 1 million samples of social media posts taken from Reddit. Figure 2 shows the stats for this dataset. The annotation was done automatically, judging by the subreddit containing the post and the number of positive and negative votes. The labels were also manually verified to ensure the reliability of the data.

Twitter15 and Twitter16 [21] datasets are the other ones we used in our research and they contain news articles from Twitter, along with their labels and metadata such as users, followers, and followers.

In our experiments, we tested the multimodal solutions on Fakeddit, for the unimodal solutions we used both Fakeddit and the Twitter datasets.

We split the data into 500 rows each for validation and testing. Only rows without missing data were kept, along with the columns for title, image ID, and the 2-way label. To get a better understanding of the dataset, Figure 3 showcases a couple of examples.



Dataset Statistics	
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 - 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

Figure 2: Dataset statistics [20]

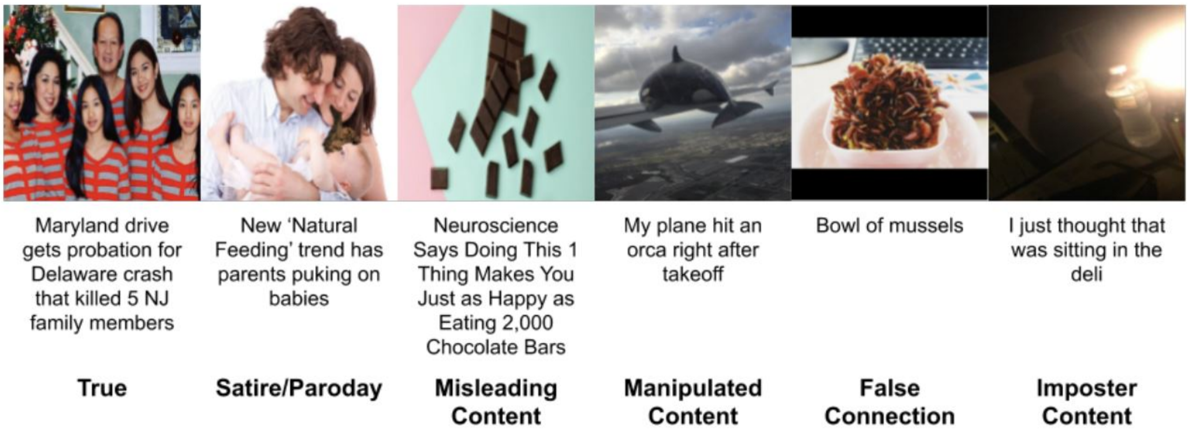

Maryland drive gets probation for Delaware crash that killed 5 NJ family members
New 'Natural Feeding' trend has parents puking on babies
Neuroscience Says Doing This 1 Thing Makes You Just as Happy as Eating 2,000 Chocolate Bars
My plane hit an orca right after takeoff
Bowl of mussels
I just thought that was sitting in the deli
True
Satire/Paroday
Misleading Content
Manipulated Content
False Connection
Imposter Content

Figure 3: Dataset sample [20]

4 PRELIMINARY RESULTS

In this section, we present all of the results obtained from our experiments. We will compare the new findings against our baseline models from the last research report (Tables 1 and 2).

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Twitter15	DistilBERT + BiGRU-CNN	61.23	60.84	61.93	61.38
Twitter16	DistilBERT + BiGRU-CNN	53	52	54	53
Twitter15+16	DistilBERT + BiGRU-CNN	66.36	65.78	65.8	65.79
Fakeddit	ResNet18	60.4	30.2	50	37.7
Fakeddit	BERT + ViT	73.37	73.73	73.15	73.26
Fakeddit	DistilBERT + ViT	74.32	73.77	73.15	73.46

Table 1: Preliminary results

Model	Accuracy	Precision	Recall	F1-Score
DistilBERT + BiGRU-CNN	76.73	76.66	75.11	75.88
ViT	74.32	73.77	73.15	73.46
Multimodal	**77.70**	**77.27**	**76.72**	**76.99**

Table 2: Supervised learning results

4.1 Experimental setup

Our setup is quite self-explanatory. We used GPT3.5-turbo and GPT4-mini LLM models through the OpenAi API and constructed our own prompts, following the mentioned articles and methods.

4.2 Results

The experiments packed with information for the final LLM seem to lose context somewhere on the way and that causes a loss of balance as well in the final assessment. The visual information proves to be vital in getting a high accuracy and LLM's ability to detect fake news is substantial. Finding the right balance between information quantity and quality is a tricky part of the prompt engineering task and more experiments are needed to further improve the model we have.

Model	Accuracy	Precision	Recall	F1-Score
Zero-Shot Prompting (GPT3.5)	76.52	0	0	0
Zero-Shot CoT Prompting (GPT3.5)	78.05	0	0	0
Few-Shot Prompting (GPT3.5)	67.17	0	0	0
Few-Shot CoT Prompting 4ex (GPT3.5)	78.96	0	0	0
Few-Shot CoT Prompting 8ex (GPT3.5)	83.74	0	0	0
Zero-Shot Tools RAG (GPT4mini)	70.71	0	0	0
Few-Shot Tools Prompting 8ex (GPT4mini)	**85.68**	0	0	0

Table 3: Few-Shot Agent Prompting results

Model	Accuracy	Precision	Recall	F1-Score
MCWDST [22]	76.90	77.27	76.89	76.82
Fakeddit [20]	89.09	-	-	-
DEAP-FAKED [23]	89.55	-	-	-
(BERT+Dense)+Xception [24]	91.87	93.39	93.29	93.25
Multimodal transformers [25]	**92.51**	93.83	93.74	93.79
FactAgent with Expert Workflow [17]	88.00	88.00	89.00	88.00
LLM+SLM [14]	78.60	78.40	81.40	80.40
ToRAG + CoTVP + CoVe [19]	84.00	85.00	86.00	85.50
Ours - Zero-Shot Tool	85.68	0	0	0

Table 4: Comparison with state-of-the-art-models

4.3 Discussions

Results from Table 3 show that our method outperforms the other baselines and is the best result in terms of accuracy that we have so far. Looking at Table 4, our model hangs up there with the state-of-the-art and looks promising for future research. Moving forward with the task of fake news detection, we will look more into up-to-date information gathering and how to best engineer a prompt that encourages the LLM to keep context and not drift off.

5 CONCLUSIONS

In conclusion, we confirmed that large language models have a substantial capability to detect fake news. The multimodal aspect of the task seemed to help the model achieve better reasoning and comparing the context of the image with the news title can be a future area to improve on.

The main bonus of working with LLMs is that they don't need training and can run at a relatively low cost. As a solution for a social media company that can either have its own LLM or a subscription to one already, it proves to be essential to take into consideration when mentioning the automatic fake news detection task.

Of course, all of this does not come without its flaws. The somewhat randomness of the model's answers and the web searches are the main points of concern. Also, we need to take into consideration the hallucination problem.

Further experiments can improve this kind of detection and we will continue with it.

BIBLIOGRAPHY

[1] S. A. Esma Aïmeur and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, 2023.

[2] D. R. Soroush Vosoughi and S. Aral, "The spread of true and false news online," *Science*, 2018.

[3] J. B. E. D. C. M. S. G. S. Savvas Zannettou, Tristan Caulfield and G. Suarez-Tangil, "On the origins of memes by means of fringe web communities.," *Proceedings of the Internet Measurement Conference*, 2018.

[4] L. Graves and M. A. Amazeen, "Fact-checking as idea and practice in journalism," *Oxford Research Encyclopedia of Communication*, 2019.

[5] M. Mayank, S. Sharma, and R. Sharma, "Deap-faked: Knowledge graph based approach for fake news detection," 07 2021.

[6] C.-O. Truica, E. S. Apostol, R.-C. Nicolescu, and P. Karras, "Mcwdst: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media.," *IEEE Access*, vol. 11, pp. 125861–125873, 2023.

[7] V. Slovikovskaya and G. Attardi, "Transfer learning from transformers to fake news challenge stance detection (FNC-1) task," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Marseille, France), pp. 1211–1218, European Language Resources Association, May 2020.

[8] S. K. Uppada, P. Patel, and S. B., "An image and text-based multimodal model for detecting fake news in osn's," *Journal of Intelligent Information Systems*, pp. 1–27, 2022.

[9] P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, "Multi-modal fake news detection via bridging the gap between modals," *Entropy*, vol. 25, p. 614, 04 2023.

[10] X. Zhang and W. Gao, "Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method," 2023.

[11] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," 2023.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[14] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi, "Bad actor, good advisor: Exploring the role of large language models in fake news detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, p. 22105–22113, Mar. 2024.

[15] Q. Liu, X. Tao, J. Wu, S. Wu, and L. Wang, "Can large language models detect rumors on social media?," 2024.

[16] Anonymous, "Zero-shot fact verification via natural logic and large language models," in *Submitted to ACL Rolling Review - June 2024*, 2024. under review.

[17] X. Li, Y. Zhang, and E. C. Malthouse, "Large language model agent for fake news detection," 2024.

[18] G. Wu, W. Wu, X. Liu, K. Xu, T. Wan, and W. Wang, "Cheap-fake detection with llm using prompt engineering," 2023.

[19] M. A. Khaliq, P. Chang, M. Ma, B. Pflugfelder, and F. Miletić, "Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models," 2024.

[20] K. Nakamura, S. Levy, and W. Y. Wang, "Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Marseille, France), pp. 6149–6157, European Language Resources Association, May 2020.

[21] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, (New York, NY, USA), p. 1867–1870, Association for Computing Machinery, 2015.

[22] C.-O. Truică, E.-S. Apostol, R.-C. Nicolescu, and P. Karras, "Mcwdst: A minimum-cost weighted directed spanning tree algorithm for real-time fake news mitigation in social media," *IEEE Access*, vol. 11, p. 125861–125873, 2023.

[23] M. Mayank, S. Sharma, and R. Sharma, "DEAP-FAKED: knowledge graph based approach for fake news detection," *CoRR*, vol. abs/2107.10648, 2021.

[24] S. K. Uppada, P. Patel, and S. B., "An image and text-based multimodal model for detecting fake news in osn's," *J. Intell. Inf. Syst.*, vol. 61, p. 367–393, nov 2022.

[25] P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, "Multi-modal fake news detection via bridging the gap between modals," *Entropy*, vol. 25, no. 4, 2023.