A Blessing in Disguise: Can Large Language Models be Useful for Misinformation Detection?

No Author Given

No Institute Given

Abstract. In the times of advanced generative artificial intelligence, distinguishing truth from fallacy and deception has become a critical societal challenge. This research attempts to analyze the capabilities of large language models for the detection of misinformation. Our study employs a versatile approach, covering multiple Large Language Models (LLMs) with both few-shot and zero-shot prompting. These models are rigorously evaluated across a diverse range of fake news and rumour detection datasets. In addition to the LLM-based analyses, we introduce a novel dimension to the study by incorporating sentiment and emotion annotations into the data. Leading sentiment analysis and emotion detection models are employed to annotate the emotional and sentiment-laden aspects of the data. This augmentation seeks to understand the potential influence of emotions on the detection of misinformation using LLMs.

Keywords: Misinformation Detection · Rumour and Fake news · LLMs

1 Preface

In the current digital era abundant with technologies of generative artificial intelligence, there is an unprecedented surge in misinformation which can be attributed to several key factors. Firstly, the widespread accessibility of the internet and social media platforms has democratized information sharing, allowing anyone to disseminate content without fact-checking. This has created an environment where both *fake news*, deliberately fabricated to deceive, and *rumour news*, which stems from unverified or loosely sourced information, can spread rapidly. Additionally, the algorithms used by social media platforms often prioritize sensational or controversial content, boosting the creation and circulation of misleading information for increased engagement and visibility. As a result, distinguishing between fact and fiction has become increasingly challenging. In addition, Large Language Models (LLMs), while powerful in generating human-like text are known to hallucinate and produce mistakes. Furthermore, malicious actors can exploit these models to generate deceptive content that appears legitimate. As LLMs become increasingly integrated into online platforms, there is a heightened risk of amplifying the spread of misinformation[1], underscoring the need for reliable detection mechanisms. Motivated by these, **we reverse the**

[1] We consider 'fake news' and 'rumour news' as equally harmful misinformation.

problem and analyze whether it is actually possible to **harness LLMs for misinformation detection**. Our study makes the following contributions:

- *Evaluating the efficacy of multiple LLMs in discerning rumour content across six distinct rumour datasets.*
- *Investigating the impact of various prompting techniques, such as few-shot and zero-shot, on the LLMs' ability to detect rumour content, while also examining their sensitivity to hyperparameters like 'temperature'*[2]*. Additionally, we explore the effects of modifying prompting instructions.*
- *Assessing whether incorporating sentiment and emotions in the input data enhances or rather hampers the LLMs' capacity to identify rumour content.*

2 Related Studies

***Exploring Misinformation**: These works encompass various aspects of fake news and rumour detection.* Zhou et al. [2] emphasized on the impact of fake news on democracy, reviewing detection methods. Shu et al. [1] provided a comprehensive survey covering characteristics, algorithms, metrics, and datasets. Zhang et al. [3] offered an in-depth overview and future directions. Work in [9] focused on rumour classification, including detection, tracking, stance, and veracity, evaluating existing approaches. Bondielli et al. [5] analyzed automatic detection, addressing definitions, data challenges, and techniques. Authors in [4] introduced the TriFN framework for precise fake news classification. Zhou et al. [6] focused on early rumour detection using reinforcement learning. Sicilia et al. [7] presented a novel system for detecting social media rumours, while Zubiaga et al. in [8], utilized context, rather than tweet querying, for classification.

Exploring Large Language Models (LLMs) for Rumour Detection: Hu et al. [11] highlighted the proficiency of LLMs like GPT 3.5, but also emphasized BERT's superior performance. Pavlyshenko et al. in [13] focused on fine-tuning the Llama 2 LLM for tasks spanning disinformation analysis to fake news detection. Additionally, authors of [12] addressed the detection of AI-generated text and attribution to specific language models. Furthermore, Sun et al. [14] have introduced Med-MMHL, a comprehensive dataset for medical misinformation detection, encompassing diverse diseases and incorporating both human-generated and LLM-generated content.

***Misinformation Detection with Sentiment and Emotions**:* Research by Alonso et al. in [10] provided a comprehensive overview of sentiment analysis in fake news detection, highlighting crucial factors and future requirements. Iwendi et al. [15] addressed COVID-19-related fake news using Information Fusion techniques. Kula et al. [16] employed deep learning for fake news detection, while authors in [19] highlighted the role of sentiments in social network-based detection. Bakir et al. [17] scrutinized the 2016 US presidential election campaign, proposing solutions and introducing the concept of "empathic media".

[2] The "temperature" parameter in a language model like GPT controls the level of randomness in generated text

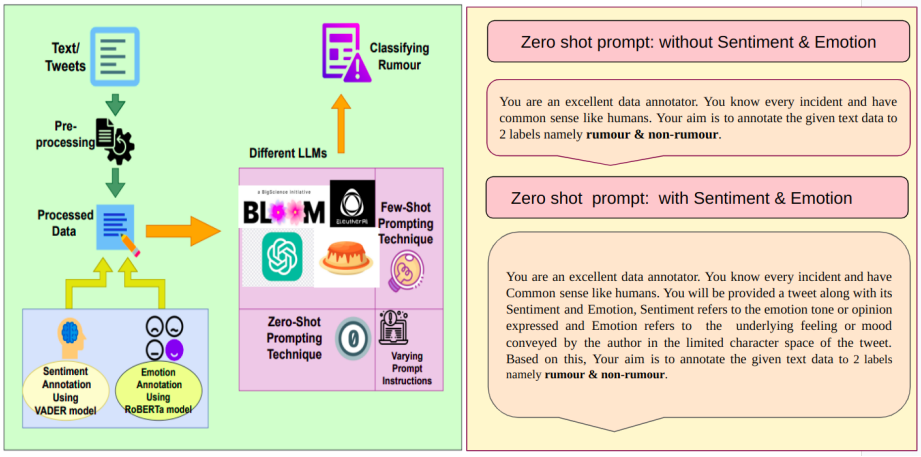

Fig. 1. Our Rumour Detection Framework and examples of Zero-Shot Prompting with and without Consideration of Sentiment and Emotions

Lastly, Mackey et al. [18] introduced a BERT-based model enhancing fake news classification with emotional cues.

3 Methodology

Our approach follows a systematic pipeline as shown in Figure 1. We commence by preprocessing a set of 500 rumour texts/tweets randomly selected from the datasets, which involve tasks such as removing URLs, emoticons, and hashtags. Subsequently, the processed texts/tweets undergo sentiment and emotion annotation using the VADER [21] and DistilRoBERTa-base [20] models, respectively. Additionally, we manually verified 50 annotated texts/tweets from each dataset to ensure accuracy. Next, a series of experiments are conducted, where the texts (with sentiment and emotions in one setting, and without in another) are fed to multiple LLMs namely, GPT-3.5 [24], BLOOM [22], FLAN-T5 [23], and GPT-Neo [25] for classification into rumour or non-rumour. Different prompt methods, including zero-shot and few-shot, are explored. Additionally, various settings are tested to gauge the performance of these LLMs. The rationale behind utilizing VADER for sentiment analysis in the experiments was due to its specialized design for microblog-like content, which often features unique linguistic characteristics. Furthermore, its incorporation of standard grammatical and syntactical rules, allows VADER to capture sentiments effectively. Likewise, the choice of using the DistilRoBERTa-base model for emotion analysis was driven by its extensive training on a diverse range of English text sources, including Twitter, Reddit, student self-reports, and TV dialogues. This aligns with the datasets utilized in our experiments. Additionally, considering our computing resource constraints, VADER and DistilRoBERTa-base were practical and justified choices for the experiments.

Datasets: We have utilized six datasets for misinformation detection. The *PHEME* dataset [26], released in 2016, captures Twitter conversations during five distinct breaking news events, meticulously labelled to distinguish rumours from non-rumors. The FakeNewsNet dataset [27], a combination of *GossipCop*[3] and *PolitiFact*[4] datasets, released in 2018, encompass labelled news articles categorized by experts as real or fake, along with relevant social context details. The *Snopes* dataset [28] created in 2020 from the Snopes website, a widely recognized fact-checking platform, includes diverse textual claims, each paired with veracity labels (true, false, or mixed) and contextual information such as origin date and source. The *IFND (Indian Fake News Dataset)* [29], introduced in 2021, is a comprehensive resource primarily focused on events spanning from 2013 to 2021. Additionally, an augmentation algorithm was employed to enhance the fake news section of the IFND dataset, generating authentic-seeming fake news statements to improve the dataset's reliability. The 2020 *ESOC COVID-19 Misinformation* Dataset [30] documents instances of misinformation across social media and news platforms, providing detailed information including sources, keywords, and direct links.

Employed Large Language Models (LLMs) Information: We provide now details on the four LLM models that we have used. First, *GPT-3.5*: introduced by OpenAI in June 2020 is a paid service model with a decoder-only transformer architecture trained on over 500 billion tokens. Next, *BLOOM*: a part of the BigScience project, is a freely accessible large language model published in 2021. It features a decoder-only transformer model and is trained on around 366 billion tokens. Then, *FLAN-T5*: an open-source instructional model, FLAN-T5 is an adapted variant of the T5 model. It has an encoder-decoder architecture and has been trained on a corpus of 1.5 trillion tokens. Lastly, *GPT-Neo*: released by EleutherAI, GPT-Neo is a large-scale language model that was meant to replicate the performance of models like GPT-3 but without the associated cost. It was introduced in 2021 and is available for free. GPT-Neo utilizes a similar architecture to GPT-3, employing a decoder-only transformer.

Prompting Methods and Settings Used: We have used zero-shot prompting and few-shot prompting. Additionally, we have employed diverse configurations to evaluate the performance of the different LLMs. This includes testing with varying temperature parameters, as well as randomly annotating sentiments and emotions to observe their impact.

Experimental Settings: Given resource constraints, we opted for the following specific models: GPT-3.5 (gpt-3.5-turbo), FLAN-T5 (google/flan-t5-small), BLOOM (bigscience/bloomz-560m), and GPT-Neo (EleutherAI/gpt-neo-1.3B). Our code implementation leveraged fundamental Python libraries including Pandas, Numpy, and CSV handling. The models were executed on two types of GPUs: the NVIDIA Tesla V100-PCIE-32GB and the GeForce GTX 1080-Ti-11GB, both operating on Unix/Linux systems. Additionally, the default value of the 'temperature' hyperparameter used by us across all LLMs was 0.7.

[3] `https://www.gossipcop.com/`
[4] `https://www.politifact.com/`

Few-Shot								
LLM Used →	*GPT-3.5*		*BLOOM*		*FLAN-T5*		*GPT-Neo*	
DATASET ↓	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*
PHEME	0.640	**0.680**	0	0	0.592	0.002	0.556	0
POLITIFACT	0.326	0.286	**1**	0	0.748	0.024	0	0
GOSSIPCOP	**0.430**	0.318	0	0	0.236	0.034	0	0
SNOPES	0.274	0.176	0	0	**0.368**	0.264	0	0
IFND	**0.356**	0.344	0	0	0.068	0.022	0	0
ESOC COVID-19	**0.262**	0.086	0	0	0.176	0.060	0	0
Zero-Shot								
LLM Used →	*GPT-3.5*		*BLOOM*		*FLAN-T5*		*GPT-Neo*	
DATASET ↓	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*
PHEME	0.638	0.630	0	0	0.412	0.168	**0.862**	0.828
POLITIFACT	0.626	0.494	**1**	0	0.526	0.136	0.804	0.768
GOSSIPCOP	**0.506**	0.296	0	0	0.454	0.086	0	0
SNOPES	0.098	0.274	0	0	**0.612**	0.236	0	0
IFND	0.562	0.374	**1**	0	**1**	0.090	0	0
ESOC COVID-19	**0.630**	0.130	0	0	0.140	0.186	0	0

Table 1. The accuracy values with different datasets and LLMs are presented for both Few-Shot and Zero-Shot settings, With "W-SE" and WithOut "WO-SE" indicating the inclusion or lack of Sentiments and Emotions.

4 Results and Discussions

Table 1 highlights several key observations. Generally, Zero-Shot produces more accurate results (more instances of complete accuracy), suggesting that the models excel without specific examples. In instances where the model demonstrates perfect accuracy, it's possible that this is due to the fact that the LLM model encountered this data with clearly defined labels during its training phase. However, when attempting to incorporate sentiment and emotions, it may interpret them as noise, resulting in a drop from 100% accuracy to 0%. Incorporating sentiments and emotions (W-SE) does not yield a performance enhancement compared to the scenario where they are excluded (WO-SE), particularly in the Zero-Shot setting. Also, the content of the dataset is particularly crucial, with POLITIFACT posing a challenge due to its lower accuracy scores. In Few-Shot settings, we observe that GPT-3.5 shows the most promising results in comparison to other LLMs and GPT-Neo performs the least, while in the Zero-Shot case, GPT-Neo is the best performing LLM for the PHEME dataset. FLAN-T5 achieves the highest accuracy for SNOPES under the Few-Shot settings. BLOOM shows complete accuracy for three combinations. A surprising observation is instances of '0' accuracy in certain Language Models (LLMs) which most likely result from inadequate fine-tuning and limited training on dedicated datasets. LLMs without tailored preparation lack specialized knowledge for accurate classification, especially in the case of refined and complex datasets. Highly intricate or ambiguous data, like the POLITIFACT dataset in Few-Shot scenarios, can

confound LLMs in verifying claims. Additionally, out-of-distribution data poses a challenge, particularly in Zero-Shot contexts, as models may struggle with unfamiliar data. The '0' accuracy cases underscore the role of proper fine-tuning and exposure to diverse datasets in enhancing LLM performance across tasks and domains.

LLM Used → GPT-3.5	*temperature = 0.2*		*temperature = 0.7*		*temperature = 1.0*	
DATASET ↓	*Zero-Shot*	*Few-Shot*	*Zero-Shot*	*Few-Shot*	*Zero-Shot*	*Few-Shot*
PHEME (WO-SE)	0.476	0.082	0.638	**0.640**	0.430	0.126
PHEME (W-SE)	0.376	0.048	0.630	**0.680**	0.404	0.094
POLITIFACT (WO-SE)	0.194	0.422	**0.626**	0.326	0.212	0.444
POLITIFACT (W-SE)	0.170	0.264	**0.494**	0.286	0.200	0.308
GOSSIPCOP (WO-SE)	0.472	0.363	**0.506**	0.430	0.468	0.414
GOSSIPCOP (W-SE)	0.362	0.188	0.296	0.318	**0.378**	0.252

Table 2. Accuracy results for varying temperature hyperparameter.

In Table 2, we experiment with different temperature settings. We can observe that often, higher temperature leads to higher accuracy scores. For instance, in the PHEME dataset without sentiment and emotions (WO-SE), the Zero-Shot accuracy increases from 0.476 at temperature 0.2 to 0.638 at temperature 0.7, and then slightly decreases to 0.430 at temperature 1.0. Likewise, in the Few-Shot setting, the accuracy increases from 0.082 to 0.640 as the temperature value rises from 0.2 to 0.7, before dropping to 0.126 at temperature 1.0. Similar trends can be observed across different datasets, highlighting the sensitivity of the model's performance to the temperature hyperparameter.

LLM Used → GPT-3.5	*Zero-Shot*		*Few-Shot*	
DATASET ↓	*C-SE*	*R-SE*	*C-SE*	*R-SE*
PHEME	0.630	0.540	**0.680**	0.306
POLITIFACT	**0.494**	0.336	0.286	0.120
GOSSIPCOP	0.296	0.282	0.318	**0.360**

Table 3. Accuracy results for correctly and randomly labelled sentiments and emotions (C-SE and R-SE, respectively).

Finally, Table 3 highlights the fact that in both the Zero-Shot and Few-Shot scenarios, correctly labelled sentiments and emotions (C-SE) improve the accuracy when compared to the cases with their values being set randomly (R-SE) suggesting they are still relevant. This is somewhat in contrast to the observations made earlier. Furthermore, GPT-3.5 performs notably better in Few-Shot scenarios across all datasets, indicating the importance of providing at least some task-specific training data for optimal performance.

Conclusion and Limitations: The findings presented in this paper suggest the utility of LLMs for misinformation detection. Additionally, they shed light on the influence of sentiment and emotions, demonstrating how they can either aid or hinder this task. It is important to remember that the specific datasets and scenarios limit our study's scope analyzed, potentially restricting its generalizability.

References

1. Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. "Fake news detection on social media: A data mining perspective." ACM SIGKDD explorations newsletter 19, no. 1 (2017): 22-36.
2. Zhou, Xinyi, and Reza Zafarani. "A survey of fake news: Fundamental theories, detection methods, and opportunities." ACM Computing Surveys (CSUR) 53, no. 5 (2020): 1-40.
3. Zhang, Xichen, and Ali A. Ghorbani. "An overview of online fake news: Characterization, detection, and discussion." Information Processing & Management 57, no. 2 (2020): 102025.
4. Shu, Kai, Suhang Wang, and Huan Liu. "Beyond news contents: The role of social context for fake news detection." In Proceedings of the twelfth ACM international conference on web search and data mining, pp. 312-320. 2019.
5. Bondielli, Alessandro, and Francesco Marcelloni. "A survey on fake news and rumour detection techniques." Information Sciences 497 (2019): 38-55.
6. Zhou, Kaimin, Chang Shu, Binyang Li, and Jey Han Lau. "Early rumour detection." 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), pp. 1614-1623. 2019.
7. Sicilia, Rosa, Stella Lo Giudice, Yulong Pei, Mykola Pechenizkiy, and Paolo Soda. "Twitter rumour detection in the health domain." Expert Systems with Applications 110 (2018): 33-40.
8. Zubiaga, Arkaitz, Maria Liakata, and Rob Procter. "Exploiting context for rumour detection in social media." In Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9, pp. 109-123. Springer International Publishing, 2017.
9. Zubiaga, Arkaitz, Maria Liakata, and Rob Procter. "Learning reporting dynamics during breaking news for rumour detection in social media." arXiv preprint arXiv:1610.07363 (2016).
10. Alonso, Miguel A., David Vilares, Carlos Gómez-Rodríguez, and Jesús Vilares. "Sentiment analysis for fake news detection." Electronics 10, no. 11 (2021): 1348.
11. Hu, Beizhe, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. "Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection." arXiv preprint arXiv:2309.12247 (2023).
12. Abburi, Harika, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. "Generative ai text classification using ensemble llm approaches." arXiv preprint arXiv:2309.07755 (2023).
13. Pavlyshenko, Bohdan M. "Analysis of Disinformation and Fake News Detection Using Fine-Tuned Large Language Model." arXiv preprint arXiv:2309.04704 (2023).
14. Sun, Yanshen, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. "Med-MMHL: A Multi-Modal Dataset for Detecting Human-and LLM-Generated Misinformation in the Medical Domain." arXiv preprint arXiv:2306.08871 (2023).
15. Iwendi, Celestine, Senthilkumar Mohan, Ebuka Ibeke, Ali Ahmadian, and Tiziana Ciano. "Covid-19 fake news sentiment analysis." Computers and electrical engineering 101 (2022): 107967.
16. Kula, Sebastian, Michał Choraś, Rafał Kozik, Paweł Ksieniewicz, and Michał Woźniak. "Sentiment analysis for fake news detection by means of neural networks." In Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20, pp. 653-666. Springer International Publishing, 2020.

17. Bakir, Vian, and Andrew McStay. "Fake news and the economy of emotions: Problems, causes, solutions." Digital journalism 6, no. 2 (2018): 154-175.
18. Mackey, A., Susan Gauch, and Kevin Labille. "Detecting fake news through emotion analysis." In Proceedings of the 13th International Conference on Information, Process, and Knowledge Management, pp. 65-71. 2021.
19. Ajao, Oluwaseun, Deepayan Bhowmik, and Shahrzad Zargari. "Sentiment aware fake news detection on online social networks." In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2507-2511. IEEE, 2019.
20. Jochen Hartmann, "Emotion English DistilRoBERTa-base". https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.
21. Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In Proceedings of the international AAAI conference on web and social media, vol. 8, no. 1, pp. 216-225. 2014.
22. Workshop, BigScience, Teven L. Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow et al. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model." ArXiv, (2022). Accessed October 2, 2023. /abs/2211.05100.
23. Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li et al. "Scaling instruction-finetuned language models." arXiv preprint arXiv:2210.11416 (2022).
24. Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.
25. Black, Sid, Leo Gao, Phil Wang, Connor Leahy and Stella Rose Biderman. "GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow." (2021).
26. Zubiaga, Arkaitz, Geraldine Wong Sak Hoi, Maria Liakata, and Rob Procter. "PHEME dataset of rumours and non-rumours." (2016).
27. Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media." Big data 8, no. 3 (2020): 171-188.
28. Vo, Nguyen, and Kyumin Lee. "Where are the facts? searching for fact-checked information to alleviate the spread of fake news." arXiv preprint arXiv:2010.03159 (2020).
29. Sharma, Dilip Kumar, and Sonal Garg. "IFND: a benchmark dataset for fake news detection." Complex & Intelligent Systems (2021): 1-21.
30. Shapiro, Jacob N., Jan Oledan, and Samikshya Siwakoti. "ESOC COVID-19 Misinformation Dataset." Empirical Studies of Conflict. 2020. https://esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset.