Feature Extraction for Claim Check-Worthiness Prediction Tasks Using LLM

Yuka Teramoto[1(✉)], Takahiro Komamizu[2], Mitsunori Matsushita[3], and Kenji Hatano[1]

[1] Doshisha University, 1–3 Tatara-Miyakodani, Kyotanabe, Kyoto 610–0394, Japan
teramoto@mil.doshisha.ac.jp, khatano@mail.doshisha.ac.jp
[2] Nagoya University, Furo, Chikusa, Nagoya, Aichi 464–8603, Japan
taka-coma@acm.org
[3] Kansai University, 2–1–1 Ryozenji, Takatsuki, Osaka 569–1095, Japan
m_mat@kansai-u.ac.jp

Abstract. This study explores the use of Large Language Models (LLMs) for Claim Check-Worthiness Prediction (CCWP), a crucial pre-screening task in fact-checking. We predict the time between a claim's occurrence and verification by analyzing data from fact-checking organizations. The results show that validation time is the same between the top 25% and bottom 75% of total checklist condition fulfillment claims. That is, further optimization is needed for LLMs to perform effective CCWPs.

Keywords: Fact-checking · Claim Check-Worthiness Prediction task (CCWP) · Large Language Models (LLMs) · Misinformation

1 Introduction

The rapid spread of misinformation can lead to significant societal disruptions, highlighting the importance of effective fact-checking mechanisms. In the fact-checking process, claims are reported by citizens and categorized by organizations, which then prioritize verification. Claim Check-Worthiness Prediction (CCWP) is a crucial task for prioritizing claims for screening. Because, a large number of fact-checking candidates are gathered at fact-checking centers, each with varying societal impacts. Fact-checking is a complex and sensitive task that can take days to weeks and places a significant burden on individuals. Furthermore, the rapid completion of CCWP is performed, the more effectively the impact of misinformation can be mitigated [10].

Computational assistance can enhance efficiency by helping prioritize information for verification. This study investigates whether LLMs can assist in the pre-screening phase based on criteria set by human experts. Previous studies have suggested that LLM and human collaboration are practical for enhancing

P. Delir Haghighi et al. (Eds.): iiWAS 2024, LNCS 15342, pp. 53–58, 2025.
https://doi.org/10.1007/978-3-031-78090-5_5

the efficiency of fact-checking tasks [3,11]. These papers insist that humans must handle the critical and sensitive aspects of the validation process because LLMs can introduce biases, hallucinations, and amplification of inaccurate data [8].

In this study, we investigate and report on the ability of LLMs in the human-in-the-loop fact-checking approach by using a structured checklist, focusing on claim checks. Specifically, we investigate whether the judgments made by LLMs according to the checklist are useful in identifying which claims should be urgently judged true or false. The use of checklists can produce the following effects: The impact of incorrect outputs is subdivided by dividing the forecasting task into tasks for evaluation in a checklist. It also reduces the difficulty of individual tasks and facilitates human verification of outputs.

We use assessments made by LLMs based on the checklist for predicting time lag between a claim's occurrence and the completion of a fact-checking task. Current CCWP methods use a binary variable to represent the need for verification, but the verification value is continuous. Additionally, existing data are unsuitable for validation in LLMs due to data leakage risks. This study uses data from actual fact-checking organizations to design a new task that predicts the days between a claim's occurrence and its verification.

2 Related Works

Many previous studies have tackled the tasks of CCWP [4]. In recent years, the check-worthiness tasks have attracted significant attention. CLEF-2024 CheckThat! lab provides a dataset designed for check-worthiness tasks and a competition opportunity for state-of-the-art methods. According to the report from the organizers of CheckThat! lab [2], methods using transformer-based models have increased in prevalence over the past few years.

Recently, researchers have dealt with comprehensive fact-checking tasks, including CCWP, collecting evidence, and determining veracity using LLMs [9]. Comprehensive fact-checking is essential to social unrest and public health issues. Therefore, caution should be exercised in using LLMs, because LLMs contain biases and can amplify these biases [8]. Moreover, LLMs can produce hallucinations that deviate from facts, and the mechanisms of these hallucinations are not fully understood [5]. Furthermore, existing researches [7,9] using LLMs still need to examine data leakage rigorously. Because the LLMs' training data may already include publicly available datasets for detecting claim check-worthy tasks and related information. On the other hand, when solving comprehensive or partial fact-checking tasks using LLMs in the real world, LLMs have already been trained, and they must also handle claims that arise subsequently.

3 Proposed Method

As Fig. 1, this study explores practical methods for incorporating LLMs into the manual fact-checking process. Specifically, we examine whether LLMs can execute CCWP, a preliminary step of the fact-checking process. In numerous

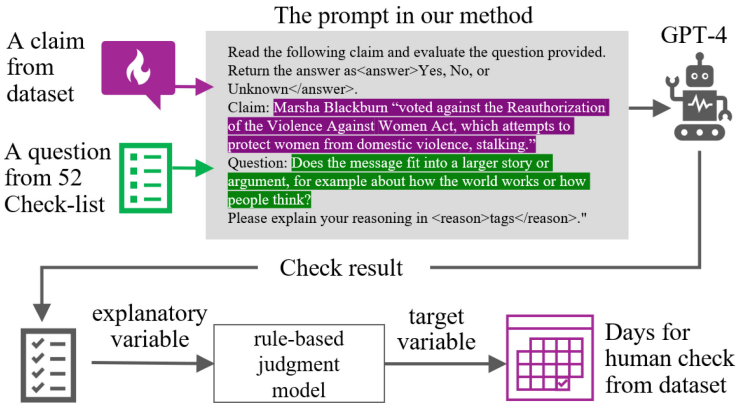

Fig. 1. Our proposed method

previous studies, the CCWP label has been represented as a Boolean value indicating whether a claim holds a checkable value. This approach was used because CCWP functioned as an initial screening process to determine whether fact-checkers should verify a reported claim. In contrast, our study envisages a new secondary screening exercise that assigns further priority to claims of assured importance that were deemed worthy of verification in the primary screening and survived the selection process. As a result, we have not made direct comparisons with existing studies. Existing check-value determination methods are likely to determine that all the data in this study are check-value. To use a simple metaphor, our task is to determine the rank of diamonds, not to separate stones from diamonds.

We use a structured checklist outlining human fact-checkers' steps to prioritize claims. Moreover, we considered the following two points to evaluate the practical application in real-world scenarios rigorously. First, we designed a new task to predict how quickly claims submitted to fact-checking organizations were addressed. Second, to demonstrate the effectiveness of LLMs on claims not included in their training data, we evaluated only claims that emerged after the cutoff date of the LLM's training data. We made the LLM execute the checklist-based claim evaluations, quantified the results, and assessed its ability to prioritize the fact-checking tasks. Sehat et al. created a 52-item checklist to determine the priority of fact-checking [12]. This checklist was developed based on a fact-checking survey of human experts. Examples of the questions are shown in Table 1. Fact-checkers respond to each item with "Yes," "No," or "Unknown." The higher the number of "Yes" responses, the more urgent it is to fact-check the corresponding claim.

This paper assumes that LLMs will likely perform comparably to humans on some of these questions in Sect. 2. Indeed, even in behavioral economics, it is known that LLMs can mimic human impressions and value judgments [6,14]. We observed the checklists and hypothesised that two types of these questions

Table 1. Examples of questions used to determine the priority of fact-checking

Question Category	Example
External information	Is there a lack of high quality information that is publicly accessible and refuting the message's claim?
Impressions from the text	Does the message directly call audience members to share the content further?

exist: those that ask for external information relevant to the claim and those that ask for characteristics or impressions derived from the text. The former includes items such as the characteristics of the claim's issuer and whether there have been any official announcements related to the claim. The latter includes whether the claim makes statements about global trends or contains aggressive bias against specific groups. The latter group of questions pertains to human impressions and value judgments derived from the text, which LLMs can likely replicate sufficiently. On the other hand, the former questions might be influenced by external factors. Examples of such external factors include the context of the documents accessible to the LLM and the structure of the websites where the claims are found. Therefore, there may be differences in the accuracy of LLM responses. There may be cases where the distinction between them is ambiguous. Therefore, we will treat these questions without distinguishing between them in our experiment.

This study uses the Fact-Check Insights dataset distributed by Duke University[1] to rigorously evaluate LLMs' responses in a zero-shot scenario. Standard benchmark datasets [2,13] for CCWP have limitations when using data beyond the cutoff date of the training data for GPT-4. The metadata of the Fact-Check Insights dataset includes the date when the claim was verified and the date when the fact-check article was published. By calculating the difference between these dates, we can determine how many days it took to address a claim in the real world. This study uses this number of days is used as the target variable. Figure 2 shows the prompt used when inputting to the LLM. Following the precedent set by prior research using LLMs as annotators [6], the response sections are structured with tags. The sentences following "Claim:" and "Question:" will be modified. For reviewing the claim, it is necessary to answer the 52 questions created in prior research [12]. Therefore, 52 different prompts will be created for each claim.

We use GPT-4 [1][2] for evaluation. There are three reasons for this choice:

1. The cutoff date of the model's training data is known.
2. Its outputs are relatively stable.
3. It is a model capable of obtaining sufficient claim data that emerged after the cutoff date from the Fact-Check Insights dataset.

[1] "Fact-Check Insights", https://www.factcheckinsights.org/, Last accessed on July 20, 2024.

[2] MODEL NAME: gpt-4-turbo, TRAINING DATA: Up to Dec. 2023.

4 Experimental Results

The Fig. 2 show simple aggregations of our analyses. A high number of "Yes" responses indicates that the complaint requires a prompt response. Therefore, we used the 25th percentile as the threshold to divide the total number of "Yes" into top and bottom groups because the 25th percentile represents the third quartile.

The response types are categorized into four groups: "Yes", "No", "Unknown", and "Error". Specifically, the number of instances classified as "Yes" is 3,328, "No" is 38,730, "Unknown" is 30,869, and "Error" is 113. This data is crucial for understanding how citizen reports on specific claims are categorized and which categories appear most frequently. Notably, the overwhelming number of instances classified as "No" indicates that many claims are evaluated negatively. As Fig. 2, the data for the bottom 75% of the respondents with a small number of "Yes" responses are often answered in a short time, but the high dispersion of the data shows a tendency for the respondents to take a long time to answer the questions. As Fig. 2, the fact-checkers are quicker to deal with claims with a high number of "Yes" responses from LLMs. However, it should be noted that a low number of "Yes" responses does not necessarily mean a low priority. Fact-checkers also responded to the bottom 75% of groups in a short time. We believe that adding other features to consider and increasing the complexity of the model is a potentially effective approach. Conversely, for the data with a large number of Yes responses, the number of days required is mostly lower than for the data with a small number of Yes responses, and the variation in the number of days is also smaller. Therefore, a large number of "Yes", is likely to be one of the conditions for a claim to be highly urgent.

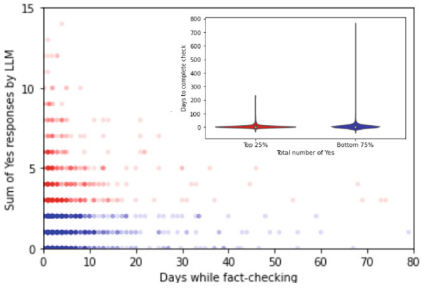

Fig. 2. The red section shows the top 25% of claims with the most "Yes" responses from the LLM checklist. The violin plot illustrates the distribution of checking days. (Color figure online)

5 Conclusion and Future Challenges

We developed scenarios for using LLMs in the fact-checking processes, considering the LLM's training data cutoff date and evaluating post-cutoff claims. Our analyses show that irrespective of the time to check, CLAIMs rarely meet the checklist conditions in determining LLM. To improve our method, we plan to develop a weighted model, analyze the impact of individual checklist items, and refine our classification system to identify critical claims better. By addressing these issues, we aim to create a more effective system for identifying and addressing urgent claims.

This work was partly supported by JST RISTEX #JPMJRS23L2 and Tateishi Science and Technology Foundation Research Grant C #237018 and the Grants-in-Aid for Academic Promotion, Graduate School of Culture and Information Science, Doshisha University.

References

1. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Barrón-Cedeño, A., et al.: The CLEF-2024 CheckThat! Lab: check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In: Goharian, N., et al. (eds.) Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, pp. 449–458. Springer Nature Switzerland, Cham (2024). https://doi.org/10.1007/978-3-031-56069-9_62
3. Das, A., Liu, H., Kovatchev, V., Lease, M.: The state of human-centered NLP technology for fact-checking. Inf. Process. Manage. **60**(2), 103219 (2023)
4. Hassan, N., Li, C., Tremayne, M.: Detecting check-worthy factual claims in presidential debates. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 1835–1838 (2015)
5. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12), 1–38 (2023)
6. Leng, Y.: Can LLMs mimic human-like mental accounting and behavioral biases? SSRN 4705130 (2024)
7. Li, X., Zhang, Y., Malthouse, E.C.: Large language model agent for fake news detection. arXiv preprint arXiv:2405.01593 (2024)
8. Neumann, T., Lee, S., De-Arteaga, M., Fazelpour, S., Lease, M.: Diverse, but Divisive: LLMs can exaggerate gender differences in opinion related to harms of misinformation. arXiv preprint arXiv:2401.16558 (2024)
9. Quelle, D., Bovet, A.: The perils and promises of fact-checking with large language models. Front. Artif. Intell. **7**, 1341697 (2024)
10. Rastogi, S., Bansal, D.: A review on fake news detection 3T's: typology, time of detection, taxonomies. Int. J. Inf. Secur. **22**(1), 177–212 (2023)
11. Schmitt, V., et al.: Evaluating Human-Centered AI Explanations: introduction of an XAI evaluation framework for fact-checking. In: Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation, pp. 91–100 (2024)
12. Sehat, C.M., Li, R., Nie, P., Prabhakar, T., Zhang, A.X.: Misinformation as a harm: structured approaches for fact-checking prioritization. Proc. ACM Hum.-Comput. Interact. **8**(CSCW1), 1–36 (2024)
13. Shaar, S., et al.: Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. arXiv preprint arXiv:2109.12986 (2021)
14. Wang, Y., Cai, Y., Chen, M., Liang, Y., Hooi, B.: Primacy effect of ChatGPT. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 108–115 (2023)