Correcting misinformation on social media with a large language model

Authors

- Jakub Kubiak, 156049
- Maciej Janicki, 156073
- Julia Lorenz, 156066

Introduction

Misinformation is an ongoing issue on social media platforms posing significant challenges to public discourse and individual decision-making. For a long time now, misinformation detection and correction have been tasks that required informed human intervetion, which is both time-consuming and not scalable. Recent advancements in large language models (LLMs) have opened new avenues for addressing this problem, enabling automated systems to generate corrective responses to misinformation. This project aims to reproduce the method introduced in the paper *Correcting misinformation on social media with a large language model*(Zhou et al. 2024). The paper proposes a novel approach to misinformation correction called MUSE, which combines search-engine querying with LLM-based summarization. The goal of this project is to implement, test, and deploy the model on a different domain.

Related Work

As a significant issue in the digital space, misinformation has been the focus of various studies. Prior research suggests that the main cause as well as the main problem of misinformation spread is the sheer volume and speed at which information is shared on social media (Adams et al. 2023). As a result, traditional methods of misinformation correction that often rely on manual fact-checking, however accurate, are not scalable, thus, inherently limited given the vast amount of content generated daily. Early work in computer science and natural language processing primarily focused on automated misinformation and fake news detection (Shu, Sliva, et al. 2017; Khan et al. 2021). These approaches typically frame the problem as a supervised classification task. While such methods have demonstrated promising results, they are mainly designed to identify misinformation rather than to correct it. More recently, the emergence of LLMs has enabled a shift from detection oriented systems toward generative approaches (Zhou et al. 2024; Guo et al. 2022) capable of producing natural language explanations and corrective responses.

Datasets Description

Dataset 1: Tweets Correction

The dataset is derived from the MUSE model's GitHub repository and is used to evaluate the generated responses. It includes the text of the original tweet, the generated corrective response, and the source of the response, as well as other metadata. The response source can be one of four categories: produced by highly helpful humans, produced by humans with average helpfulness, generated by the GPT-4 model, or generated by the MUSE model.

Dataset 2: Fake News Detection

The second dataset is derived from the FakeNewsNet repository, which compiles news articles from multiple sources, including both genuine and fake content. For this project, we focus on the subset of articles identified as fake news from the BuzzFeed website. FakeNewsNet was introduced as part of a series of studies on fake news detection on social media (Shu, Sliva, et al. 2017; Shu, Wang, et al. 2017; Shu et al. 2018). Each record includes the title of the article and the corresponding main text.

Dataset 3: Fact-checked Online Content

Final dataset also derivered from Kaggle website contains fact-checked posts from various online platforms, covering a wide range of topics between 2008 and 2022. Each entry includes the post's title, content of the post, a truthfulness rating status, and a link to the original content.

Method Description

The method consists of four main steps, firstly, the input content, which is a social media post is preprocessed and input into a LLM with an appropirete prompt to generate three adquete search queries. Next, there queries are used to retrieve relevant websites and sources of information from the web. Then, relevant content is extraced from the retrieved sources. Finally, the post's content along with the extracted information is fed into the LLM to generate a corrective response.

During our project we introduced an additonal step before the query generation, where we prompt the LLM to rewrite the input content to make it more suitable for query generation. This step is intended to investigate whether improving the clarity and structure of the input content can lead to more effective search queries and ultimately better corrective responses.

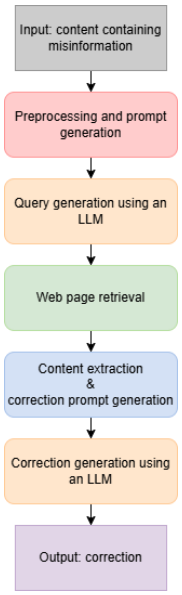

Figure 1: Method Pipeline

Evaluation Startegy

We implemented the method step by step, starting with preprocessing the content and creating the initial prompts for query generation. The preprocessing step involved cleaning the text, removing unnecessary characters, changing temporal expressions to applicable dates from content's metadata, and formatting it appropriately for input into the LLM.

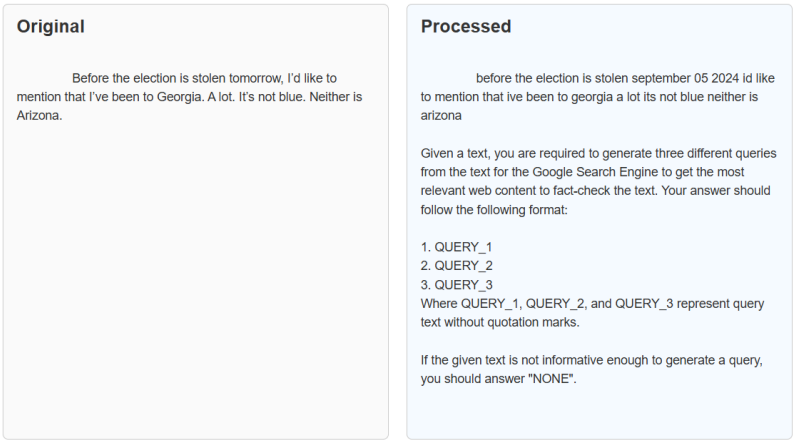

Figure 2: Preprocessing Step

Next, we focused on the query generation step. We designed prompts to instruct the LLM to generate three relevant search queries based on the preprocessed content. We evaluated the quality of the generated queries based on their relevance to the original content and their potential to retrieve useful information manually.

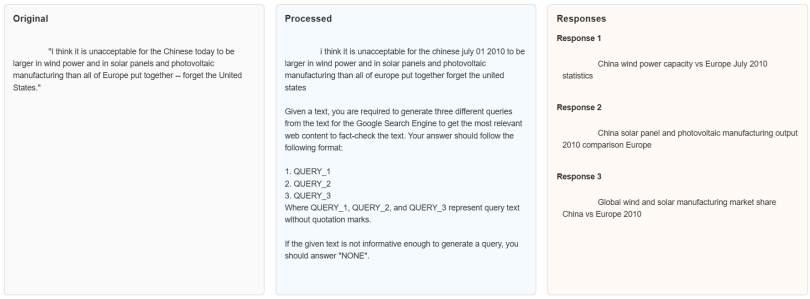

Figure 3: Query Generation Step

Outcomes and Conclusions

References

MUSE GitHub repository: https://github.com/Social-Futures-Lab/MUSE/tree/main
FakeNewsNet Kaggle repository: https://www.kaggle.com/datasets/mdepak/fakenewsnet
Verified Posts Kaggle repository: https://www.kaggle.com/datasets/mattimansha/verified-posts-fact-checking-online-content

Adams, Zoë, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. "(Why) Is Misinformation a Problem?" *Perspectives on Psychological Science* 18 (6): 1436–63. https://doi.org/10.1177/17456916221141344.

Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos. 2022. "A Survey on Automated Fact-Checking." *Transactions of the Association for Computational Linguistics* 10 (February): 178–206. https://doi.org/10.1162/tacl_a_00454.

Khan, Junaed Younus, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. 2021. "A Benchmark Study of Machine Learning Models for Online Fake News Detection." *Machine Learning with Applications* 4: 100032. https://doi.org/https://doi.org/10.1016/j.mlwa.2021.100032.

Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media." *arXiv Preprint arXiv:1809.01286*.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. "Fake News Detection on Social Media: A Data Mining Perspective." *ACM SIGKDD Explorations Newsletter* 19 (1): 22–36.

Shu, Kai, Suhang Wang, and Huan Liu. 2017. "Exploiting Tri-Relationship for Fake News Detection." *arXiv Preprint arXiv:1712.07709*.

Zhou, Xinyi, Ashish Sharma, Amy X. Zhang, and Tim Althoff. 2024. *Correcting Misinformation on Social Media with a Large Language Model*. https://arxiv.org/abs/2403.11169.