Abstract

Fact-checking is the process by which journalists verify whether a claim is true. With the rapid spread of misinformation and the rapid growth of social media, it is crucial to automate fact-checking to help network moderators reduce its impact on social news platforms. Moreover, the output of the fact-checking task requires not only the stance for the claim but also the explanation for the stance of the claim. In addition, current trends combine data from multiple modalities, such as text and images; thus, the fact-checking system must be able to process multimodal data.

This thesis aims to construct an end-to-end Fact-checking system on a multimodal dataset, comprising three main stages: retrieving evidence, verifying the claim based on the evidence, and justifying the claim's truthfulness. There are several appropriate approaches for an automatic end-to-end Fact-checking system, such as deep learning with multimodal fusion techniques and a sequence-to-sequence model for generating results in natural language. However, the main challenge for the deep learning approach is the lack of large-scale human-annotated data and potential bias in the dataset. Large language models (LLMs) trained on vast amounts of data can exploit valuable latent patterns even with limited or no training data, enabling efficient instruction prompting for fact-checking tasks. The flexibility of LLMs can then be leveraged to construct an end-to-end Fact-checking system.

We proposed MCVE - a transformer model with fusion techniques that efficiently combine text and image feature vectors within a classification network to determine the truthfulness label of the claim. Besides, we introduce a simple idea that leverages the image caption as a feature to combine with textual evidence, forming the input sequence to a sequence-to-sequence model for generating a justification of the claim's veracity. The experimental results show the efficiency of our deep neural network-based fusion method for the Multimodal fact-checking task, as evaluated on two multimodal fact-checking datasets. Next, we propose three instruction prompting frameworks — ZeFAV,

TabV4FC, and M-RAV — to leverage the robustness of LLMs for verifying claims across diverse evidence modalities: text, images, and tables. Then, we evaluate our frameworks on different fact-checking datasets with multimodal evidence, including text-only, text-image, and text-table, and the empirical results show the efficiency of our proposed frameworks in leveraging the LLMs for the Fact-checking task. Finally, we introduce FALCON - an end-to-end system that implements LLMs with a UI to demonstrate the fact-checking task in practice. The system can be hosted entirely locally with a single GPU, without any external API dependencies, demonstrating the potential for scalability and maintenance in a real system.

In the future, we would like to improve the performance of the end-to-end fact-checking system by optimizing inference runtime and integrating temporal information into LLMs to better understand time-based events in the data and verify the correctness of information, especially time-sensitive data such as financial, legal, and political.

Keywords: fact-checking, deep neural network, large language models, multimodal, end-to-end systems