Generative Large Language Models for the Fake Detection task

Project title:

Generative Large Language Models for the Fake Detection task -A comparative performance analysis.

Supervisor(s): Uku Kangur (University of Tartu, Estonia), Paolo Zicari (University of Calabria, Italy), Rajesh Sharma (CSAI, Plaksha University, Mohali, India).

Project objectives

In today's digital landscape, the rapid spread of misinformation and fake news poses a significant threat to public trust, democratic stability, and individual decision-making. With the widespread use of social media and online platforms, false information can quickly reach a lot of people, influencing opinions on critical issues such as health, politics, and climate change.

Fake news detection is a complex task that involves analyzing the linguistic features of text, understanding context, detecting bias, and verifying facts using external sources.

To address these challenges, researchers are exploring advanced AI techniques based on Natural Language Processing (NLP) and large language models (LLMs) to build more accurate, explainable, and adaptable systems for detecting misinformation.

Generative LLMs (ChatGPT, DeepSeek, LLAMA, etc.), trained on billions of words (books, websites, code, etc.), can have a fundamental role in the fake news domain thanks to their capability of better understanding natural language, reasoning, and finding coherence.

The main objective of the proposed research is to analyze how the generative LLMs perform in the specific task of fake detection.

The challenging questions we want to answer are:

1) Is the huge knowledge in the pre-trained generative LLMs useful for detecting fakes?
2) Is there a direct correlation between fake detection performances and the size of the model (in terms of number of neurons/weights)?
3) Are small non-generative Language Models opportunely fine-tuned on the fake detection task better than big Large Language Models pre-trained on billions of documents? That is, is it better to know less but be skilled on the specific task through fine-tuning training or to have encyclopedic knowledge without a specific skill?
4) Do models perform better on specific topics than on other topics? That is, which is the best model for detecting fakes in political news, which for gossip news, and so on?
5) Are there performance improvements with different prompting (Zero-shot prompting, CoT prompting)?
6) Are there performance improvements if we give examples of real and fake news in the prompt of the model (Few-shot prompting with examples)
7) Are there performance improvements if the model can use external data (RAG)?

Experiments

1) Selecting datasets:
    a. Politifact,
    b. GossipCop,
    c. …
2) Selecting LLMs:
    a. llama3.2,
    b. mistral,
    c. phi3,
    d. phi4,
    e. Gemma3,
    f. deepseek-v2,
    g. qwen2.5,
    h. Vicuna,
    i. Grok 4
    j. GPT 5
    k. BERT (as non-generative model)

    l. …
3) Selecting prompting:
    a. Zero-shot prompting
    b. Few-shot prompting with examples
    c. CoT prompting
    d. RAG
    e. …
4) Set the training, validation and test sets (training and validation to be used only for Bert-like models, test set to be used for examples in prompting by examples) for n runnings
5) Launch the experiments
6) Fill the results table
7) Analyze the results

The final objective is to get the following comparative table

Columns: LLM, Size, Dataset, prompting technique, performances (acc, F1 micro and macro, precision for fake and real, roc-auc, …)

Dataset	**LLM**	**Size**	**prompt**	**Acc**	**F1 micro**	**F1 macro**	**Prec_fake**	**Prec_real**	**Roc_auc**	**…**
Politifact	DeepSeek		Zero-shot prompting							
			Few-shot prompting with examples							
			…							
	LLAMA		Zero-shot prompting							
			Few-shot prompting with examples							
			…							
	…									
	Bert									
GossipCop	DeepSeek		Zero-shot prompting							
			Few-shot prompting with examples							
			…							
	LLAMA		Zero-shot prompting							
			Few-shot prompting with examples							
			…							
	…									
	Bert									

…										

An analysis of the performances is followed depending on the results