Zero is Not Hero Yet: Benchmarking Zero-Shot Performance of LLMs for Financial Tasks

Aditya D Kulkarni (016904537)

Introduction

- Delve into the performance of LLMs in zero shot scenarios
- Understand how these models perform without labeled data which is a common problem in finance
- Address related questions on data annotation, performance gaps and feasibility of employing generative models in finance.

Chat GPT's Impact in Finance

- Since its introduction, it has shown promise in interpreting complex financial communication
- Ability to analyze and predict trends in AI related financial assets
- There is work done to understand the capabilities of ChatGPT in NLP, not much work done for financial NLP
- According to a survey, ChatGPT outperforms fine tuned models only on 22.5% tasks.

Contd...

Key Insights

- Even though zero-shot ChatGPT fails to outperform fine-tuned PLMs, it provides impressive performance across all the tasks without having access to any labeled data
- The performance gap between fine-tuned model PLMs and ChatGPT is larger where the dataset is not publicly available yet

Contd...

Key Insights

- The performance of fully open-source LLMs for all financial tasks is significantly lower as compared to ChatGPT
- In certain scenarios, even if the user is willing to accept the performance difference between zero-shot LLMs and fine-tuned PLMs, the amount of time required to assign labels to data is 1000 times greater when using generative LLMs.

Dataset and Tasks

Sentiment Analysis

- Correlates with market sentiment which influences price movements
- Soon after BERT, FinBERT was developed for financial sentiment analysis

Contd...

Numerical Claim Detection

- Extraction of numerical claims from the financial text like analysts' reports, earnings calls, news, etc
- Dataset contains binary labels "in-claim", and "out-of-claim"
- "in-claim" refers to sentences that contain specific and measurable financial claims, which are not factual statements
- "out-claim" refers to sentences that contain factual information from the past

Contd...

Named Entity Recognition

- NER plays a crucial role in financial NLP
- Enables the extraction and categorization of key financial entities such as company names, person names, locations, etc.
- Facilitates the identification of important entities involved in financial news and reports
- Crucial for extracting meaningful information and enhancing decision-making processes in financial NLP

Experiments

4 Experiments in Total

- Fine Tuning PLM
- Zero Shot with Generative LLMs
  - FOMC Communication
  - Sentiment Analysis
  - Numerical Claim Detection
  - Named Entity Recognition

Results

- Fine Tuning
  - RoBERTa-base and RoBERTa-large have similar performance and they outperform zero-shot LLMs
  - ChatGPT follows the instruction 100% of the time, while Dolly fails to follow instructions for 6.05% of the time and H2O for 42.14% of the time
- Sentiment Analysis
  - The H2O model doesn't follow the instruction at all
  - ChatGPT achieves impressive performance close to 0.9 F1 score for sentiment analysis which is not far from the performance of fine-tuned RoBERTa

Contd...

- Numerical Claim Detection
  - The gap between fine-tuned model PLMs and ChatGPT is larger while the gap between Dolly and ChatGPT is smaller
  - One possibility could be contamination, considering that the sentiment analysis dataset is publicly available while the claim dataset is not
- Named Entity Recognition
  - The NER prompt proves to be more challenging compared to other tasks, leading both open-source models to struggle in adhering to the given instruction
  - Decoding labels from the prompt output is relatively more difficult in token classification compared to sequence classification tasks resulting in a higher percentage of samples with missing labels

Conclusion

- While fine-tuned PLMs generally outperformed zero-shot ChatGPT, it still demonstrated impressive performance across various tasks without the need for labeled data
- A notable performance gap was identified between fine tuned PLMs and ChatGPT, particularly in cases where the dataset was not publicly available. This discrepancy could be attributed to potential contamination issues within the data
- Despite potential performance gaps between zero-shot LLMs and fine-tuned PLMs, the time required to label a single sample using generative LLMs was significantly higher, potentially by a factor of 1000