Outperforming Morningstar Analysts: Applying Enhanced LLMs to Financial Reports

1st Jonas Gottal
LMU Munich
Computer Science

2nd Mohamad Hagog
LMU Munich
Computer Science

Abstract—Describe Goal, data, method and results.
Index Terms—LLMs, PEFT, Adapters, Hugging Face, Finance

I. INTRODUCTION

Motivation and hypothesis [1]–[4]

Why are we doing this and what doe we hope will be the results

II. DATA

Report and financial market data. Where does the report come from? How does it look like (JSON preview)? What is the most important information? How do we merge the market data? Where does it come from?

A. Data Preprocessing

How do we preprocess the data? What are the challenges? Why do we process it that way?

III. METHODOLOGY AND FOUNDATIONS

What are the foundations of our approach? What is the goal? Sentiment Classification.

A. Transformer models

What are transformer models? Build an intuition on self-attention and multi-head attention. Use the presentation pictures (youtube video) and reproduce the images quickly in a sketch style on iPad.

B. Parameter efficient fine tuning (PEFT)

We want to explore many models and fine-tune them on our data. So we need a more efficient approach: PEFT. What are adapters? Why do we use them? How do we use them? What is so effficient about it? How does it work? Show image from IntSys lecture (incl sources) Quickly describe own experiment and results as table for SST2.

C. Hugging Face

What is Hugging Face? Why do we use it? What are the benefits?

D. Evaluation

How do we evaluate the models? What are the metrics? What is the baseline? What is ROC (build intuition). Why ROC AUC and not F1 score?

IV. APPROACH

Model building and training and Evaluation pipeline.

A. Model selection

How did we select the models? What are the models? (all relevant to sentiment and finance)

B. Adapter configuration

How can adapters be configured? What are the options? What is the industry standard for our problem and what did we use?

C. Exploration to Exploitation

How do we explore the models? Just run them all on the same configuration and compare them. How do we exploit the best model? We try to optimize it further and let it run for all the different text columns.

D. Benchmarking

What are our benchmarks? What are the results? We use the analysts own predictions as input for a logistic regression model as benchmark. And we use all the categorial information from the results and yahoo finance as input for a XGBoost random forest model as benchmark. Results: show `build_roc` for n=0 and baseline=True

V. RESULTS

Show the results of the models. Show top 5 models and their ROC AUC after initial runs. Show the results of the top 5 ROC after optimization.

A. Models in comparison

Discuss results and wheather the performance is enough to outperform the analysts.

B. Insights via CAPTUM

What are the models using as input? Use CAPTUM and quickly describe the gradient approach to obtain feature attribution.

C. Training analysis

Show some plots about runtime and training time. Compare with Loss, Accuracy etc

VI. Conclusion

Wrap it up and discuss the results. What are the implications? What are the limitations?

References

[1] C. Poth, H. Sterz, I. Paul, S. Purkayastha, L. Engländer, T. Imhof, I. Vulić, S. Ruder, I. Gurevych, and J. Pfeiffer, "Adapters: A unified library for parameter-efficient and modular transfer learning," 2023.

[2] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.

[3] S. J. Russell and P. Norvig, *Artificial Intelligence – A Modern Approach, Global Edition*, 4th ed., ser. Pearson Education. Harlow, United Kingdom: Prentice Hall, 2021.

[4] G. Kauermann, H. Küchenhoff, and C. Heumann, *Statistical Foundations, Reasoning and Inference*. Cham, Switzerland: Springer Nature Switzerland, 2021.