Outperforming Morningstar Analysts: Applying Enhanced LLMs to Financial Reports

1st Jonas Gottal
LMU Munich
Computer Science

2nd Mohamad Hagog
LMU Munich
Computer Science

***Abstract*—This study examines financial analyst reports from Morningstar, revealing two key findings: firstly, analysts appear to select stocks arbitrarily, and secondly, they provide comprehensive textual justifications that enable informed decisions surpassing mere chance. We employ enhanced large language models to efficiently analyze the textual data, facilitating a broad exploration of various pre-trained models to identify the subtle underlying sentiment and extract more value from the reports than the experts themselves.**

***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.

A. Data description

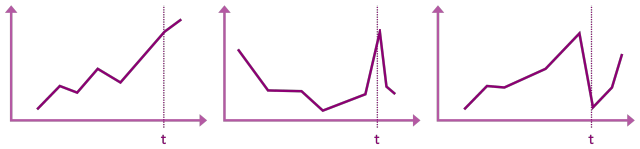
Where does the report come from? How does it look like and what is the most important information? Target group: investors not speculators.

- `ParseDate`: The date the information was retrieved.
- `Title`: The title of the analyst report.
- `CompanyName`: The name of the company.
- `TickerSymbol`: The ticker symbol (without exchange information) of the underlying stock.
- `Rating`: The analyst rating of the stock.
- `ReportDate`: The date of the release of the report.
- `AuthorName`: The name of the author of the report.
- `Price`: The price of the stock declared in the report.
- `Currency`: The given currency of the price.
- `PriceDate`: The date the price was retrieved from market data (Morningstar).
- `FairPrice`: The estimated fair price from the analyst.
- `Uncertainty`: The company's uncertainty quantified in 'Low', 'Medium', 'High' and 'Very High'.
- `EconomicMoat`: The ability to maintain competitive advantages quantified in 'Narrow' and 'Wide'.
- `CostAllocation`: Decisions on investments categorized as 'Poor', 'Standard', and 'Exemplary'.
- `FinancialStrength`: Rating of the ability to make timely payments and fulfill obligations – quantified in 'A', 'B', ... 'F'.
- `AnalystNoteDate`: The date of the analyst note.
- `AnalystNoteList`: The analyst note.
- `BullsList`: Arguments in favor of the company.
- `BearsList`: Arguments against the company.
- `ResearchThesisDate`: The date of the thesis.
- `ResearchThesisList`: Objective research thesis.
- `MoatAnalysis`: Added research on EconomicMoat.
- `RiskAnalysis`: Company's risk profile.
- `CapitalAllocation`: Text on the CostAllocation.
- `Profile`: Short text on the company profile.
- `FinancialStrengthText`: Short text on the financial strength of the company (conclusion).

How do we merge the market data? Where does it come from?

B. Data Preprocessing

How do we pre-process the data? What are the challenges? Why do we process it that way? Target group: investors not speculators.



(a) Archetype (b) Pure luck (c) Bad luck

Fig. 1: Different stock developments as charts: Comparing the archetype of a good investment to spurious spikes (positive and negative).

III. 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 efficient about it? How does it work?

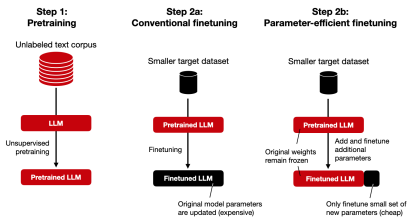


Fig. 2: process of PEFT [5]

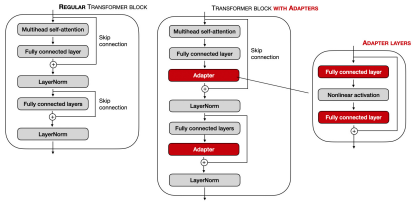


Fig. 3: process of PEFT (detailed) [5]

C. Hugging Face

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

IV. Literature

As described in [6], there are many different applications of LLMs, and specifically for analyst reports they can provide insights into subtle tone and sentiment to add value. But at the moment there is only research on publicly available data such as Reddit [7]. Furthermore, it has been shown that reports can have a small positive performance in the right market conditions [8], but it is unclear whether only recommendations themselves can influence markets [9] and the reports have no value at all [10]. Thus, the herding factor can have an influence on stock performance [11]. In [12] they use LLMs to better interpret and analyze Korean financial analysts' reports. However, this paper uses data from widely spread reports (from retail traders to professionals) on globally traded firms. We show that while the analysts can barely perform better than the current market, they provide comprehensive textual justifications that enable informed decisions. We employ enhanced large language models to efficiently analyze the textual data, facilitating a broad exploration of various pre-trained models to identify the subtle underlying sentiment and extract more value from the reports than the experts themselves.

LLMs to compare the analysts recommendations to an LLM sentiment from the analysts texts. Using

V. Approach

This is a classification problem based on sentiment scores Model building and training and Evaluation pipeline. Quickly describe own experiment and results as table for SST2.

A. Model selection

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

Sentiment

- `kwang123/bert-sentiment-analysis`
- `siebert/sentiment-roberta-large-english`
- `distilbert/distilbert-base-uncased-finetuned-sst-2-english`

Finance

- `ProsusAI/finbert`
- `yiyanghkust/finbert-tone`
- `bardsai/finance-sentiment-pl-fast`
- `RashidNLP/Finance-Sentiment-Classification`
- `ahmedrachid/FinancialBERT-Sentiment-Analysis`
- `soleimanian/financial-roberta-large-sentiment`
- `nickmuchi/sec-bert-finetuned-finance-classification`
- `nickmuchi/deberta-v3-base-finetuned-finance-text-classification`

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. First model against model, then with best models we compare adapter configurations, then we compare text inputs from the report and finally we compare it to directly fine-tuning the model itself. How do we exploit the best model? We try to optimize it further and let it run for all the different text columns.

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?

In order to validate learned models we compare predictions made from the model with previously separated test data. In this way, we can objectively evaluate them according to their predictive qualities. For this comparison, we utilize these fundamental principles:

Sensitivity or true positive rate (TPR) is derived from the true positives TP, i.e., the correctly identified positives P from the test set: $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}}$

Specificity or true negative rate (TNR) is derived from the true negatives TN, i.e., the correctly identified negatives N from the test set: $\text{TNR} = \frac{\text{TN}}{\text{N}}$
By plotting both the sensitivity and specificity in relation for various threshold values, we obtain the so called Receiver Operating Characteristic (ROC) curve. For the results of both measures 1.0 is the optimum, and if the curve is the diagonal, we observed a random process. In Figure 4, we can see some common examples of curves. In order to further summarise these evaluations, we can calculate the Area Under the Curve (AUC) to rank the models. Again, a value of 0.5 indicates a random process. [13] [4] [3]

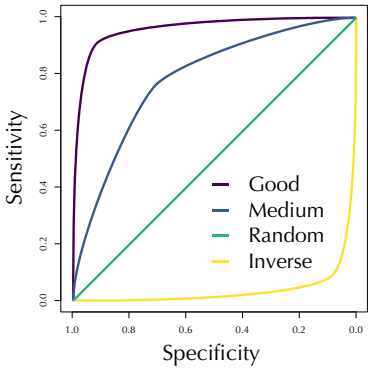


Fig. 4: Exemplary receiver operating curves for a *good, medium, random* and *inverse* model fit. The perfect score would be 1.0 on each dimension – specificity and sensitivity.

We furthermore compare the models with classical approaches based on the analysts forecasts themselves and other metrics from the report to create a baseline.

VI. Results

Show the results of the models. Show top 5 models and their ROC AUC after initial runs. First we compare the different models for the same adapter configuration and text input.

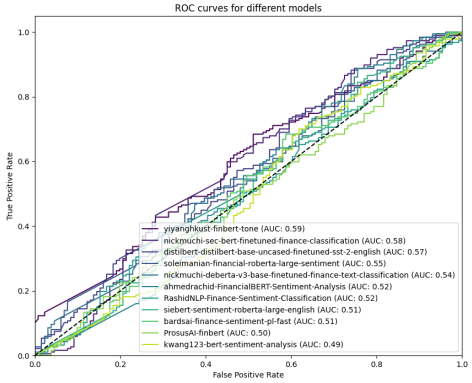


Fig. 5: First Experiment

For the best models we then compare the best adapter configurations found in literature:

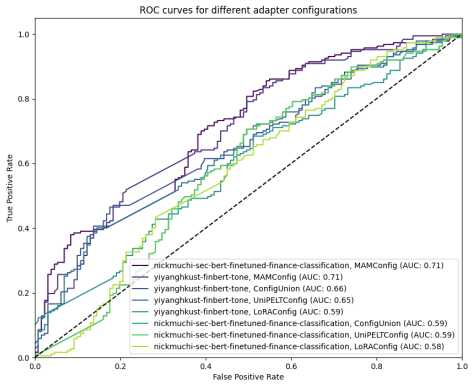


Fig. 6: Second Experiment

Next to `AnalystNoteList`, we also explore `BullsList`, `BearsList`, `ResearchThesisList`, `MoatAnalysis`, `RiskAnalysis`, `CapitalAllocation`, `Profile`, and `FinancialStrengthText`.

Finally we use the best found combination to finetune the underlying model directly. This only leads to a small improvement of 0.x AUC but the training time climbed from XX to YY.

A. Benchmarking

What are our benchmarks? What are the results? We use the analysts own predictions as input for a logistic regression

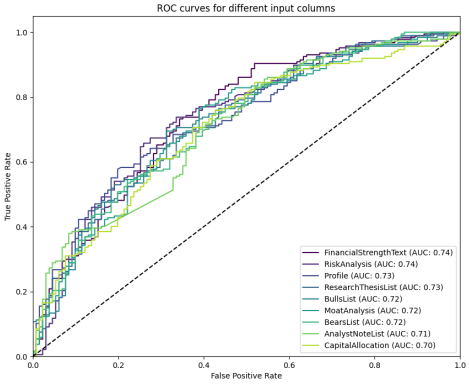


Fig. 7: Third Experiment

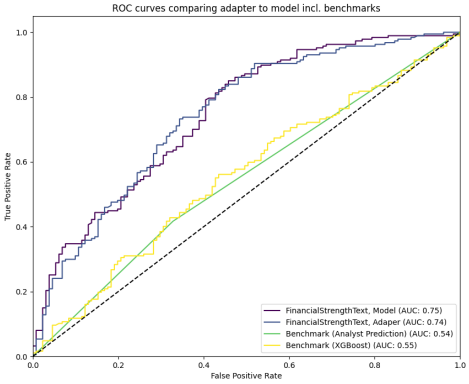


Fig. 8: Fourth Experiment

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

B. Models in comparison

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

C. Insights via CAPTUM

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

D. Training analysis

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

VII. 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.

[5] S. Raschka, "Finetuning llms efficiently with adapters," May 2023. [Online]. Available: https://magazine.sebastianraschka.com/p/finetuning-llms-with-adapters

[6] H. Zhao, Z. Liu, Z. Wu, Y. Li, T. Yang, P. Shu, S. Xu, H. Dai, L. Zhao, G. Mai, N. Liu, and T. Liu, "Revolutionizing finance with llms: An overview of applications and insights," 2024.

[7] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "What do llms know about financial markets? a case study on reddit market sentiment analysis," in *Companion Proceedings of the ACM Web Conference 2023*, ser. WWW '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, pp. 107 – 110. [Online]. Available: https://doi.org/10.1145/3543873.3587324

[8] C. Su, H. Zhang, and R. S. Hudson, "The time-varying performance of uk analyst recommendation revisions: Do market conditions matter?" *Financial Markets, Institutions & Instruments*, vol. 29, no. 2, pp. 65 – 89, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/fmii.12126

[9] M. Brauer and M. Wiersema, "Analyzing analyst research: A review of past coverage and recommendations for future research," *Journal of Management*, vol. 44, no. 1, pp. 218 – 248, 2018. [Online]. Available: https://doi.org/10.1177/0149206317734900

[10] V. Panchenko, "Impact of analysts' recommendations on stock performance," *The European Journal of Finance*, vol. 13, no. 2, pp. 165 – 179, 2007. [Online]. Available: https://doi.org/10.1080/13518470500459782

[11] M. Palmer, M. Eickhoff, and J. Muntermann, "Detecting herding behavior using topic mining: The case of financial analysts," in *Research Papers*, vol. 97, 06 2018.

[12] S. Kim, S. Kim, Y. Kim, J. Park, S. Kim, M. Kim, C. H. Sung, J. Hong, and Y. Lee, "Llms analyzing the analysts: Do bert and gpt extract more value from financial analyst reports?" in *Proceedings of the Fourth ACM International Conference on AI in Finance*, ser. ICAIF '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 383 – 391. [Online]. Available: https://doi.org/10.1145/3604237.3627721

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, US: Springer New York, 2009.