Outperforming Morningstar Analysts: Applying Enhanced LLMs to Financial Reports

1st Jonas Gottal
LMU Munich
Department of Computer Science

2nd Mohamad Hgog
LMU Munich
Department of 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. The code is available at github.com/trashpanda-ai.**

***Index Terms*—LLMs, PEFT, Adapters, Hugging Face, Finance**

I. Introduction

Many search for guidance in financial analyst reports for investment purpose. Especially during the COVID-19 pandemic, the number of retail traders has increased significantly. And Morningstar is one of the most popular platforms for financial reports that are widely used by both retail traders and professionals.

However, the question arises whether these reports are actually valuable or just an illusion of expertise. In this paper, we investigate the value of analyst reports and whether they can outperform the market. We hypothesize that the reports contain valuable information that can be extracted using enhanced large language models (LLMs) to outperform the market.

II. Data

Financial analysts provide detailed reports on single companies (equities), detailing their current situation and future prospects. These reports are based on a variety of factors, including their financial strength, economic moat and cost allocation. The reports also contain a rating and a fair price estimate.

A. Data description

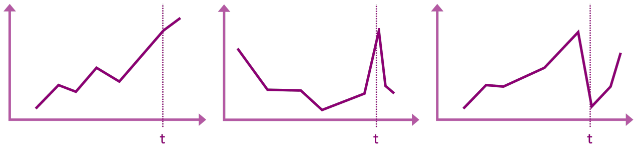
The data is retrieved from Morningstar and contains the following columns:

- `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).

The most important information for the application of LLMs is the textual data – starting with `AnalystNoteList`, which we expect to have the most effective content due to its subjective tone. We also analyze more objective texts like `BullsList` and `BearsList` as well as `ResearchThesisList`, `MoatAnalysis`, `RiskAnalysis`, `CapitalAllocation`, `Profile` and `FinancialStrengthText`. Since the target group of financial analyst reports is investors, we focus on the long term perspective and not on short-term speculation. Therefore, we look at the stock price development over a longer period. The data is retrieved from Yahoo Finance and contains the following columns: `Date, Open, High, Low, Close, Adj Close` and `Volume`. We base our target variable for the sentiment analysis on the `Adjusted Close` price.

B. Data Preprocessing

Since we care about investment and not speculation, we pre-process the data to remove spurious spikes and other noise. We focus on the archetype of a good investment as shown in Figure 1a and ignore spurious spikes as shown in Figure 1b and Figure 1c. This means we consider the average performance within a certain time period – here the maximum amount since the last report was released: 120 days. Inspired by the commonly used moving averages in mathematical finance [1], the preprocessing is done by calculating the average stock price over 120 days since publication and comparing it to the stock price at the time of the report. If the average stock price is higher[1] than the price at time of release, we label it as a good investment, otherwise as a bad investment. This is our binary target variable for the following analysis.



(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).

The final data (1586 reports) is split into a training, development and test set with a ratio of 70:10:20 (1103, 165, 318). The training set is used to train the models, the development set to optimize the hyperparameters and the test set to evaluate the models.

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? [2]

C. Hugging Face

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

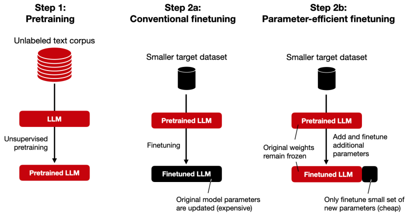


Fig. 2: process of PEFT [3]

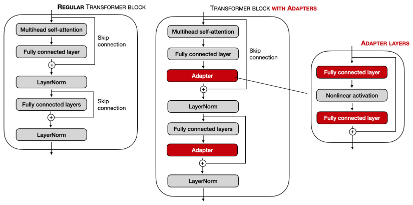


Fig. 3: process of PEFT (detailed) [3]

IV. Literature

As described in [4], there are many different applications of LLMs, and specifically for analyst reports they can provide insights into subtle tone and sentiment to add value. However, at the moment there is only research on publicly available data such as Reddit [5]. Furthermore, it has been shown that reports may have a small positive performance in the right market conditions [6], but it is unclear whether only recommendations themselves can influence markets [7] and reports have no value at all [8]. Thus, the herding factor may have an impact on stock performance [9]. In [10] they use LLMs to better interpret and analyze Korean financial analysts' reports. In this paper, however, we use data from widely spread reports (from retail traders to professionals) on globally traded firms. We show that while analysts can barely outperform the current market, they provide comprehensive textual justifications that enable informed decisions. We use enhanced large language models to efficiently analyze the textual data, allowing a broad exploration of different pre-trained models to identify the subtle underlying sentiment and extract more value from the reports than the experts themselves.

V. Approach

Our ultimate objective can be interpreted as a binary classification problem based on sentiment analysis. We use the textual data from the reports to predict whether the stock is a good investment or not. In order to do so we compare the performance of different models and configurations to find the best one. We also compare the performance of the models to the analysts' own predictions and other metrics from the report to create a baseline. In order to explore many different models and configurations, we use the parameter efficient fine-tuning approach. We use adapters to fine-tune the models and Hugging Face to access the pre-trained models. Even though literature has shown adapters

[1] A sideway movement within a 2% margin is not classified as "higher".

to be more efficient while retaining the performance of the underlying model, we setup a preliminary experiment with the Stanford Sentiment Treebank v2 (SST2) data set for sentiment classification to validate our approach and compared the efficiency and performance of the models:

	Fine-tuned LLM	**Fine-tuned Adapter**
Training Runtime	1h 58m	**24m**
Evaluating Accuracy	**0.908**	0.902
Evaluation Runtime	6.83s	**1.84s**
Evaluation Loss	0.42	**0.31**

TABLE I: Comparison of Fine-tuned LLM and Fine-tuned Adapter

Therefore, we used PEFT to find the best performing setup and built a automated training and evaluation pipeline to efficiently explore the model landscape of Hugging Face.

A. Model selection

Since our approach is centered around sentiment classification of financial reports, we allocated the most fitting pre-trained models from both worlds – sentiment analysis 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`

Among the finance subset, there are also models that are fine-tuned on the task of sentiment classification with finance related text data like the public available SEC filing reports.

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 and Exploitation

Since running all possible combinations of configurations is intractable, we assume independence of our various parameters in our experimental setup and base our approach on exploration and exploitation. We explore all possible valid models and configurations and exploit the best ones in the next iteration. First we compare the different models, then we optimize for the best adapter configuration, and after that we search for the best text input. Finally, we compare our PEFT approach to fine-tuning the LLM itself. All experiments are performed on the same data, using the same training and test split to ensure comparability, and we use the same evaluation metrics, training parameters and hyperparameters for all models. Everything is run on the same hardware.

D. Evaluation

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: $\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}}$

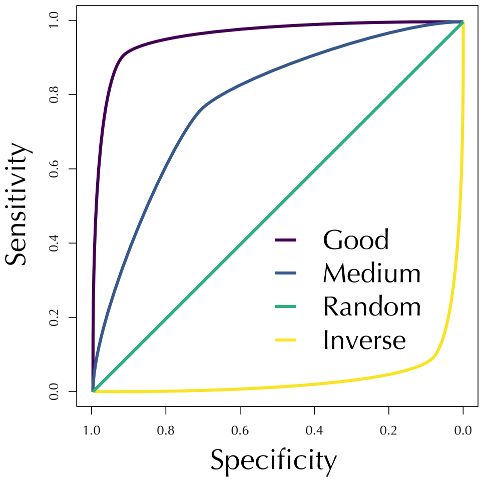


Fig. 4: Exemplary receiver operating curves for a *good, medium, random* and *inverse* model fit. The perfect score would be 1.0 on each dimension – specificity and sensitivity.

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 summarize 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. The reason we use ROC rather than accuracy, precision, recall and f1 score is that ROC is a more comprehensive metric that provides a more detailed view of model performance. In addition, ROC curves make it easier to compare the performance of different models across all classification thresholds. This also makes it more robust to class imbalance, which is very important for inherently unbalanced data such as financial markets, which often follow trends. It also helps to understand how the performance of the model changes with different classification thresholds, which can be crucial for decision making in different conditions where the cost of false positives and false negatives can vary. [11] [12] [13]

VI. Results

In this section we present the results of our experiments. We compare the models with classical approaches based on the analysts' forecasts themselves and other metrics from the report to establish a baseline. We present the results of our model training in three steps: First, we compare the different models, then we optimize the best adapter configuration, and finally we search for the text input that provides the highest predictive power. We also directly fine-tune the underlying model to see if PEFT loses significant predictive power. We also use feature attribution via CAPTUM to gain insight into the models and provide an overview of training time, runtime and number of parameters.

A. Benchmarking

In order to evaluate our models accurately, we first create a benchmark based on the analysts' own predictions. We use the analysts' `FairPrice` to create binary predictions and use them as input for a logistic regression model as a first benchmark. We also use all the categorical (and ordinal) information from the reports as input for an gradient boosted tree model (XGBoost) as a benchmark: `Rating`, `AuthorName`, `Uncertainty`, `EconomicMoat`, `CostAllocation`, `FinancialStrength`, `Sector`, `Exchange`, `Prediction`. We assume that certain authors might be better than others, and that the uncertainty, economic moat, cost allocation and financial strength might be valid indicators of a firms success. We also assume that the sector and exchange might have an impact on the stock performance. We use the same training and test split as for the LLMs and the same evaluation metrics. The results are shown in Figure 5.

The results show that neither the analysts nor the XGBoost model can outperform randomness – it only reproduces the current distribution of the market (trend following). This is a strong indication that the reports themselves may not contain any valuable information.

B. Models in comparison

Our approach assumes that all changes in our experiments are independent and that the results are comparable. As first step we compare the different models on the same adapter configurations and text inputs to find the best one.

We continue for the two best models (both trained on financial sentiment data) and compare the best adapter configurations found in literature:

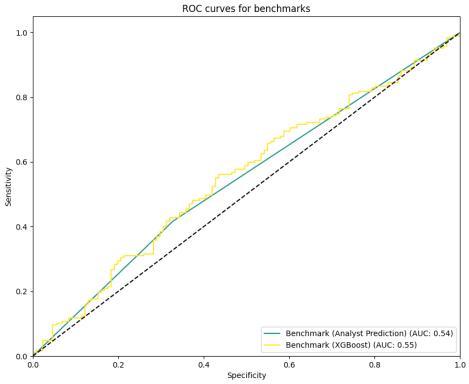


Fig. 5: Benchmark Experiment

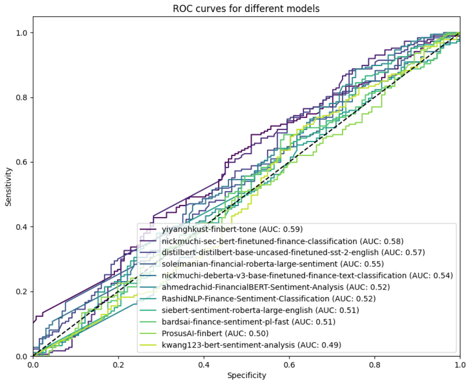


Fig. 6: First Experiment

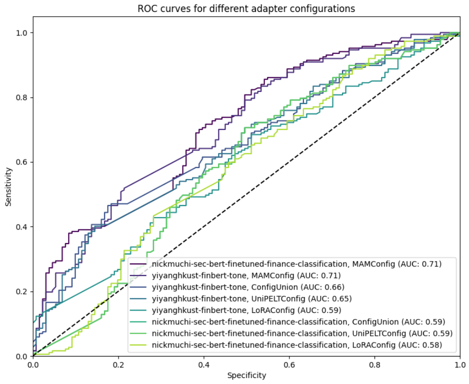


Fig. 7: Second Experiment

Next to `AnalystNoteList` we also explore `BullsList`, `BearsList`, `ResearchThesisList`, `MoatAnalysis`, `RiskAnalysis`, `CapitalAllocation`, `Profile`, and `FinancialStrengthText` as input data. We find that `FinancialStrengthText` as a concise summary provides the best predictive power.

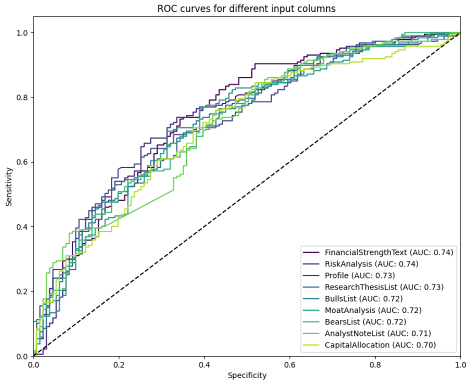


Fig. 8: Third Experiment

Finally, we use the best combination found to directly fine tune the underlying model. This leads to only a small improvement of 0.0031 AUC and the training time was roughly the same, but the number of parameters could be drastically decreased from 110 million to 23 million.

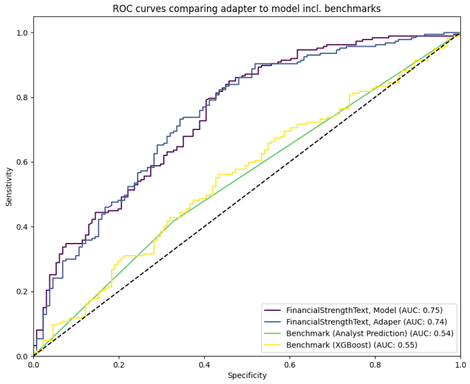


Fig. 9: Fourth Experiment

While there is no overall cutoff value and it highly depends on the specific application, with values above 0.7 [14], we reached a robust predictive quality.

C. Insights via CAPTUM

What are the models using as input? Use CAPTUM and quickly describe the gradient approach to obtain feature attribution. [15]

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] J. C. Hull, *Options, Futures, and other Derivatives*, 11th ed. Boston: Pearson, 2021.

[2] 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.

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

[4] 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.

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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.

[10] 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

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

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

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

[14] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. John Wiley & Sons, 2013.

[15] 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.