When Small Wins Big: Classification Tasks Where Compact Models Outperform Original GPT-4

Baptiste Lefort, Eric Benhamou, Jean-Jacques Ohana, Beatrice Guez, David Saltiel, Damien Challet

When Small Wins Big: Classification Tasks Where Compact Models Outperform Original GPT-4

B.Lefort[1,2] **E.Benhamou**[1,3] **JJ.Ohana**[1] **B.Guez**[1] **D.Saltiel**[1] **D.Challet**[2]

[1]Ai for Alpha
[2]Centrale Supelec
[3]Paris Dauphine PSL

Abstract

This paper evaluates Large Language Models (LLMs) on financial text classification, comparing GPT-4 (1.76 trillion parameters) against FinBERT (110 million parameters) and FinDROBERTA (82.1 million parameters). We achieved a classification task on short financial sentences involving multiple divergent insights with both textual and numerical data. We developed a market-based large dataset that enabled us to fine-tune the models on a real-world ground truth. Utilizing a market-based dataset for fine-tuning on extensive datasets, we achieved significant enhancements with FinBERT and FinDROBERTA over GPT-4. However, the use of a bagging majority classifier did not yield performance improvements, demonstrating that the principles of Condorcet's jury Theorem do not apply, suggesting a lack of independence among the models and similar behavior patterns across all evaluated models. Our results indicate that for complex sentiment classification, compact models match larger models, even with fine-tuning. The fine-tuned models are made available as open-source for additional research.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated significant success in the task of sentiment analysis [Araci, 2019, Yang et al., 2020, Sohangir et al., 2018, Day and Lee, 2016, Wu et al., 2023, Yang et al., 2023]. Their proficiency in interpreting intricate patterns and considering extensive contexts improved their effectiveness [Chen et al., 2023]. Previously, NLP models encountered learning constraints due to limited context windows and the small scale of training data. The limited quantity of parameters also hindered their ability to capture complex patterns [Day and Lee, 2016]. Large context windows are associated with prompt engineering in LLMs, which are highly responsive to the input prompts. Various prompting techniques have demonstrated significant enhancements, such as the Chain-of-Thought method cited in [Wei et al., 2023]. Adopting a human-like approach to problem-solving enhances the model's performance. Additionally, few-shot prompting contributes to improvements across various tasks in LLMs, as detailed in [Zhang et al., 2022]. The extensive number of parameters in the model allows for fine-tuning to perform in specific tasks.

Recent advancements in Natural Language Processing (NLP) have enhanced sentiment analysis tasks, which involve categorizing text into "positive", "negative", or "neutral" class. This task is extensively explored within the financial sector, encompassing intricate textual and numerical data [Mishev et al., 2020]. There are specific considerations [Dumiter et al., 2023, Brière et al., 2023] unique to this field, including:

- Brevity: Financial news articles are typically concise, packed with advanced insights.
- Mixed Data: They encompass both textual and numerical information, providing a comprehensive view of financial events.
- Financial news can rapidly lose its relevance, often discussing events that have occurred in the past, and the precise impact of such news is not always immediately known.
- Complexity: The categorization of financial texts can be challenging and contentious, even for humans, requiring detailed analysis.

The financial information is often a short text like headlines or tweets. This makes the interpretation difficult even for a human[Malo et al., 2013]. This challenging textual data is then a very favourable context for improving sentiment analysis understanding. In this study, we developed a methodology for improving understanding of LLM's sentiment classification facing a short text of interwoven issues.

Accepted for the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024).

We first created a comprehensive dataset of financial headlines which were automatically classified as positive, neutral, or negative. This classification was based on the subsequent day's market performance. Specifically, we identified any mentioned stocks or financial markets within each headline and used their next-day performance as a basis to determine the sentiment of the headline. If a headline did not reference any particular stock or financial asset, we considered the next day's overall equity market movement, averaging the performance across markets in the US, Europe, and Asia. A headline was considered positive (or negative) if the market movement exceeded (or fell below) a certain quantile threshold. Headlines that did not fit these criteria were labeled as indecisive. Further details are elaborated in the paper. We evaluated state-of-the-art financial classification models that were previously trained on expert-annotated data (by individuals with master's degrees in finance) against our dataset. Our analysis included a comparison of performance enhancements through a bagging approach and fine-tuning of Language Learning Models (LLMs) using our dataset as a baseline. We found that models, irrespective of their size, displayed equivalent performance levels when fine-tuned. Fine-tuning here refers to the process of training models on a segment of our novel dataset and assessing them on a separate, unseen portion of the dataset. We also present evidence, based on Condorcet's jury Theorem, illustrating that the performance of models in complex classification tasks is not determined by the quantity of their parameters. The key contributions of this paper include:

- Model performance is not solely dictated by its size; in fact, smaller models that have been fine-tuned, such as FinBERT and DistilROBERTA, significantly surpass the performance of larger, non-fine-tuned models like GPT-4. This underscores the critical role of fine-tuning in enhancing the ability of Language Learning Models (LLMs) to interpret financial news accurately.
- Moreover, a comparison between fine-tuned models reveals negligible differences in performance between smaller models and GPT-4, indicating that the benefits derived from fine-tuning are largely independent of the model's size. This observation particularly suggests that, following fine-tuning, GPT-4 does not provide any significant advantages over smaller models like FinBERT or DistilROBERTA.
- Additionally, employing a bagging majority classifier fails to produce significant improvements in performance, casting doubt on the relevance of Condorcet's jury Theorem in this context. This outcome is linked to the inability to satisfy the theorem's requirement for classifier independence, notably revealing that models tend to respond similarly when assessing the sentiment of financial news.

The rest of this paper is organized as follows. Section 2 briefly reviews the related works. Section 3 details our data collection process from Bloomberg Market Wraps and the creation of headlines to form a new comprehensive dataset, which serves as a baseline for the fine-tuning of Large Language Models (LLMs). In Section 4, we discuss the methodology employed to automatically assign labels to our generated headlines, utilizing the historical next-day return quantiles of the identified tickers for positioning. Section 5 provides a comparative analysis of various models tested against this market-based baseline dataset. Section 6 introduces the development of a majority vote classifier, also known as a bagging classifier, and examines its characteristics through the lens of the Condorcet jury theorem. The section also notes that, as our experiments indicate, the lack of notable enhancement with the bagging classifier suggests that the underlying assumptions—particularly the independence of the various models—are not necessarily met. Lastly, section 7 concludes and exposes future direction of research.

2 RELATED WORKS

In the field of sentiment classification for financial texts, LLMs have been widely adopted, demonstrating high accuracy in real applications [Hansen and Kazinnik, 2023, Cowen and Tabarrok, 2023, Korinek, 2023, Lopez-Lira and Tang, 2023, Noy and Zhang, 2023, Lefort et al., 2024, Zhao et al., 2024]. These studies collectively affirm the efficacy of LLMs in discerning sentiments in financial texts, marking a significant stride in the application of artificial intelligence in finance.

The advent of FinBERT, a model specifically honed for financial text analysis through the process of fine-tuning pre-existing LLMs on financial datasets, has markedly propelled the performance metrics forward [Araci, 2019]. This advancement underscores the potential of targeted model optimization to enhance accuracy and relevance in sector-specific applications.

Moreover, the introduction of niche datasets such as FinQA has played a pivotal role in refining the proficiency of LLMs in processing financial texts interspersed with numerical information [Chen et al., 2022]. These datasets train models to adeptly navigate the dual landscape of textual and numerical data, an essential capability for analyzing financial news that often melds narrative with figures.

The research into a more compact version of the BERT model, known as DistilBERT, has yielded optimistic results by demonstrating that reducing the number of parameters does not significantly compromise the model's ability to classify sentiments [Sanh et al., 2020]. This finding is crucial as it validates the efficiency of streamlined models in conducting sentiment analysis, particularly in the context of financial documents that typically present a complex blend

of text and data.

The process of fine-tuning models has been notably successful in dissecting intricate financial documents, which frequently combine various types of inputs [Li et al., 2023]. This technique enhances the model's understanding and interpretation of the multifaceted information contained within these documents.

Furthermore, the integration of Retrieval Augmented Large Language Models has elevated the capacity of LLMs to conduct sentiment analysis by weaving in a broader contextual backdrop [Zhang et al., 2023a]. This approach enriches the analysis by drawing upon external knowledge, which is particularly beneficial in the financial news sector where understanding the deeper, often unspoken implications of news is crucial.

The acquisition of extensive information, pertinent for grasping the sentiment in financial news, emerges as a significant hurdle, necessitating deep, contextual knowledge beyond the immediate content of the news articles [Kim and Nikolaev, 2023]. The challenge is compounded by the sheer volume of information that needs to be sifted through, as highlighted by [Loukas et al., 2023].

The task of classifying news headlines for sentiment analysis is further complicated by the inherent brevity of headlines, which often lack sufficient context. This scarcity of context poses a daunting challenge for LLMs, hindering their ability to accurately infer the underlying sentiments [Zhang et al., 2023b]. Despite these obstacles, certain LLMs have shown a remarkable ability to interpret headlines effectively, with the GPT model standing out for its adeptness in sentiment analysis within concise textual formats. A notable development by Lefort et al. [2024] introduced an indicator that correlates the sentiments extracted from headlines with movements in equity markets, thereby attesting to GPT's utility in sentiment analysis even in limited contexts.

In assessing the performance of LLMs in sentiment classification, comparisons are often drawn with annotations made by human experts [Brière et al., 2023, Araci, 2019]. However, this comparison is inherently flawed, as it may not accurately reflect the real-time dynamics of the market. Studies, such as [Lewis et al., 2021], have shown that LLMs exhibit a lower tendency to generate unfounded content ("hallucinate") when their analyses are grounded in factual data, emphasizing the importance of data veracity in enhancing the reliability of sentiment analysis outcomes.

3 DATA AND METHODOLOGY

3.1 DATA COLLECTION

Transitioning from collecting data to its utilization encompasses several stages and is not direct. The first step was to collect reliable financial news. For this purpose, we amassed a collection of Bloomberg Market Wraps spanning from 2010 to 2024. Bloomberg Market Wraps are a consolidated summary of daily financial news, done by human journalists specialized in finance, highly regarded and extensively followed by professionals within the financial sector. These summaries distill the day's most significant financial events and market movements into a digestible format, offering a rich, condensed source of relevant information. Their comprehensive nature renders them as particularly valuable textual data for analysis. In principle, Bloomberg Market Wraps should not miss any significant daily news and are spread to the financial community through multiple channels like the Bloomberg professional network but also various journalistic web channels like Yahoo finance, Investing.

Following the collection of these reliable sources, the subsequent phase entailed the extraction of headlines from the amassed financial news. This process yielded over 3,700 individual news items and more than 61,000 headlines. These headlines serve a dual purpose: they are instrumental in both the training and the evaluation phases of the models. This extensive compilation of headlines provides a robust dataset that mirrors the variety and complexity of financial news, facilitating the development of models capable of understanding and categorizing financial information with high accuracy.

3.2 HEADLINE GENERATION

After collecting the daily news, we extracted headlines highlighting the day's most important information, enabling us to summarize the information effectively. This approach helped us concentrate on the vital aspects by filtering out minor details and condensing the news into brief sentences. Additionally, it allowed us to gather more information by eliminating unnecessary noise and isolating the key facts in the headlines. The resulting headlines are both informative and useful for making investment decisions. Below is the prompt used to generate these headlines with GPT-4 model.

Headline Prompt:

> *You will be provided with a financial text, and your task is to extract a list of headlines from it. Each headline must be informative and provide relevant insights for a financial market analyst. Ensure that each headline contains a single piece of information. List these headlines in the specified format, with each headline separated by a line break and without additional commentary. Format your list as follows:*
>
> 1. *Headline for Theme 1*
> 2. *Headline for Theme 2*
> 3. ...

This two-step approach is valuable for improving model's classification performances, as it delivers meaningful and noise-free information following [Lefort et al., 2024].

4 A MARKET-BASED DATASET FOR EVALUATION

Numerous financial models that analyze sentiment in financial texts rely on benchmarks from multiple financial corpus. These corpora include news articles or extensive financial documents featuring both textual and numerical data [Chen et al., 2022, Araci, 2019]. The impact of the news is evaluated by a group of financial experts (or master's degree students) who assign a label for each text. However, this approach has its biases. Some headlines are difficult to classify even for financial expert and opinions may differ. This create a subjective label which can be misleading for evaluating the model's performances. Also, fine tuning the model based on this baseline is misleading and enable the model to learn on subjective data which cannot reflect the real market sentiment [Araci, 2019].

To address this bias, we established a new baseline that depends on the market return linked to the news headline. This approach allows us to assess the actual impact of the headline on the market and subsequently, the model's capability to identify the true effect.

4.1 TICKER IDENTIFICATION

Firstly, we utilize GPT-4 to identify and assign a list of tickers associated with the headline. The model has demonstrated efficiency in accurately determining the list of tickers. By doing so, we capture the genuine market response to the news, enabling us to grasp the actual impact of the news. After getting all the associated tickers for each headline, we evaluate the ticker return.

$$R_{T_k}(h_{i,t}) = \frac{P_{T_k}(t+1) + D_{T_k}(t+1)}{P_{T_k}(t)} - 1$$

The subtlety here in calculating the return of a particular stock lies in precisely incorporating any dividends paid. Here, the return of the ticker $R_{T_k}(h_{i,t})$ represents the return of ticker T_k given the headline $h_{i,t}$ at time t. This return is calculated by dividing the next day price $P_{T_k}(t+1)$ and any potential due dividend $D_{T_k}(t+1)$ at time $t+1$ by the previous price $P_{T_k}(t)$ at time t, and then subtracting 1. This formula neatly encapsulates the mechanism to measure the actual impact of a financial news headline on the market performance of a specific stock ticker.

A full summary of our process is given in figure 1

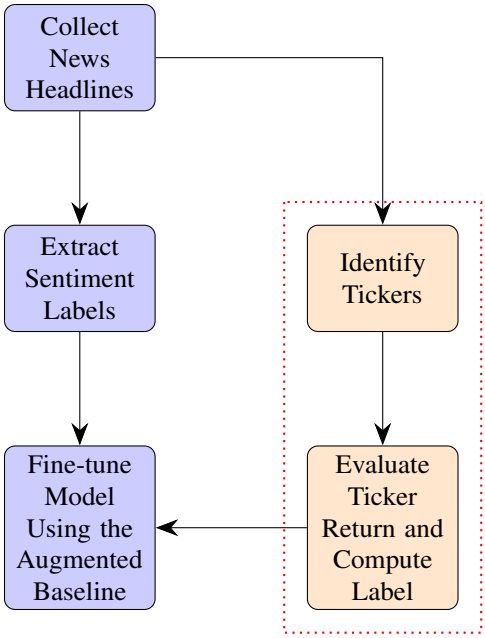

Figure 1: Workflow diagram to generate the dataset. Orange blocks represent the new steps introduced in this work, specifically the identification of relevant tickers and the evaluation of their market return to automatically annotate our database. The data augmentation is highlighted by the dotted red rectangle.

4.2 TICKER SENTIMENT CLASSIFICATION

The classification of financial headlines based on their impact on stock price movements is implemented through a quantile approach to ensure proper labelling. This approach utilizes historical stock performance data to categorize the impact of news-induced percentage changes in stock prices. We provide a formal definition of the classification algorithm:

Let pct_changes denote a dictionary where the keys represent stock tickers (T_k) and the values are the percentage changes (ΔP_{T_k}) in stock prices as a result of specific news headlines. Given a date (D), the function classify_headline aims to assess the impact of news on stock prices systematically.

For each ticker T_k, the algorithm executes the subsequent steps:

1. **Historical Data Retrieval:** For ticker T_k, historical closing prices over the preceding five years are obtained. This timeframe is determined by subtracting one day from D to establish the end date (D_{end}) and subtracting 1250 days (5×250) to pinpoint the start date (D_{start}).
2. **Percentage Change Calculation:** The daily percentage change in closing prices ($\Delta P_{\text{hist},T_k}$) is calculated

for the historical dataset.

3. **Quantile Determination:** Two pivotal quantiles, $Q_{0.3,T_k}$ and $Q_{0.6,T_k}$, are computed from $\Delta P_{\text{hist},T_k}$. These quantiles act as thresholds for classifying the impact of the current percentage change.
4. **Classification:**
   - If $\Delta P_{T_k} > Q_{0.6,T_k}$, the news impact is classified as positive $(+1)$, indicating a substantial positive market reaction.
   - If $\Delta P_{T_k} < Q_{0.3,T_k}$, the impact is classified as negative (-1), signifying a significant negative market reaction.
   - Otherwise, the news impact is deemed neutral (0), suggesting an insignificant or mixed market reaction.

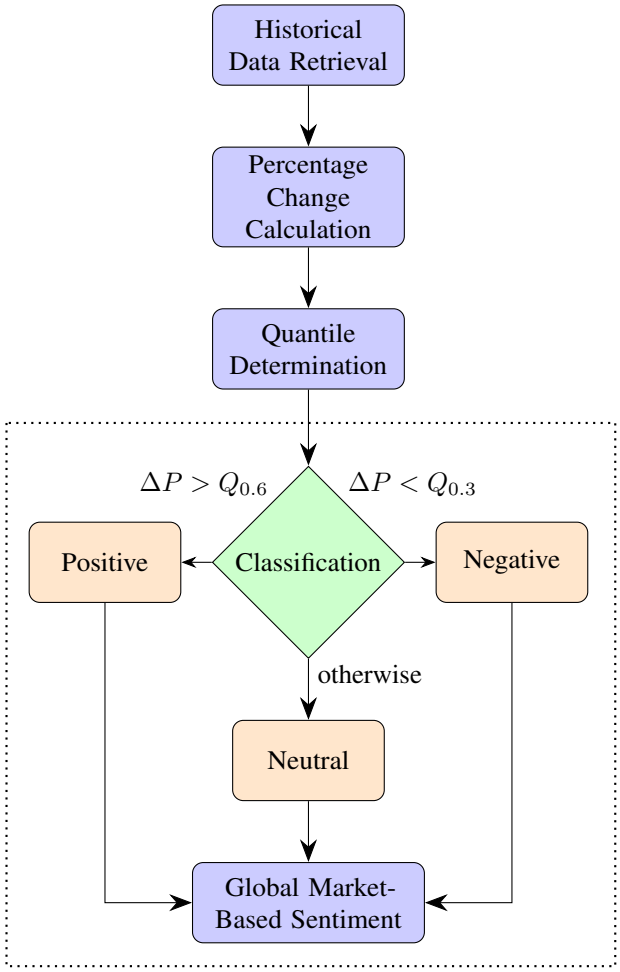

Figure 2: Automatic Classification of Financial Headlines. Blue blocks represent data processing steps, green block represents decision points for classification and orange blocks corresponding labels.

Formally, the classification function $C(T_k, \Delta P_{T_k})$ for a ticker T_k given its percentage change ΔP_{T_k} is defined as follows:

$$C(T_k, \Delta P_{T_k}) = \begin{cases} +1 & \text{if } \Delta P_{T_k} > Q_{0.6,T_k} \\ -1 & \text{if } \Delta P_{T_k} < Q_{0.3,T_k} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This classification framework facilitates a detailed understanding of the news impact, leveraging historical volatility and performance benchmarks to gauge the significance of current events. Finally we obtain the global market-based sentiment of a headline by taking the median of its ticker list sentiment and projecting it in one of these values{-1, 0, 1} respectively for negative, neutral, positive. The distribution of the sentiments is balanced as detailed in table 1.

Table 1: Number of instance in each class with percentage

Negative	Neutral	Positive
19254 (*31%*)	16202 (*27%*)	25795 (*42%*)

5 COMPARISON OF THE MODELS ON THE MARKET-BASED DATASET

Initially, we evaluate the performance of several leading models on this classification task, utilizing the GPT-4 model, frequently recognized for its efficiency in financial reasoning and analysis. We compare GPT-4 against two other open-source Large Language Models renowned for their accuracy in sentiment analysis within the financial domain: FinBERT and DistilROBERTA, both specifically fine-tuned for financial news sentiment analysis. These models are accessible on the HuggingFace.co platform: FinBERT and DistilROBERTA fine-tuned on financial news.

5.1 MODELS SIZE RECAP

The models vary in the number of parameters and have been trained on a smaller dataset compared to GPT-4. For reference, Table 2 provides the parameter count for each model.

Table 2: Model Parameters Number

Model	Param. Number
GPT-4	1.76×10^{11}
Distil ROBERTA	110×10^{6}
FinBERT	82.1×10^{6}

5.2 MODELS BEFORE FINE-TUNING

Despite its considerably larger parameter count, GPT-4 exhibits only marginally better performance than the other models. The substantial difference in the number of parameters between GPT-4 and the more compact models does not translate into a significantly enhanced decision-making capability in analyzing financial texts, as shown in table 3. This table highlights that, even though GPT-4 has advanced capabilities for financial text analysis, its improvement in decision-making is not markedly superior to that of BERT-based models. It is crucial to note that the F-score is weighted, prioritizing classes with a higher number of correct classifications.

Table 3: Model Performances Before Fine-Tuning

Model	Precision	Recall	F-score
GPT-4	0.48	0.49	0.47
Distil ROBERTA	0.45	0.45	0.44
FinBERT	0.46	0.46	0.44

Challenges arise particularly with the neutral class, which all models struggle to classify accurately. This difficulty is attributed to the complex financial insights often embedded in headlines with a neutral market impact. Accurate classification in these cases demands deep financial market knowledge and experience, which the models lack. To assess the models' ability to develop this intricate reasoning, they were fine-tuned on a baseline dataset.

5.3 FINE-TUNED MODELS

To enable Large Language Models (LLMs) to uncover hidden patterns for prediction, we fine-tuned them using a market-based dataset, allocating 70% of the dataset for training, which encompasses approximately 40,000 headlines. This represents a significant volume of data for model training. All models underwent fine-tuning with identical parameters and the same dataset. Furthermore, the evaluation set remained consistent across all models.

Table 4: Fine-Tuned Model Performances

Model	Precision	Recall	F-score
SFT GPT	0.54	0.53	0.51
SFT Distil ROBERTA	0.53	0.51	0.49
SFT FinBERT	0.54	0.52	0.50

SFT refers to "Supervised Fine-Tuning", a process where models learn to predict specific labels from given inputs. Following this process, Large Language Models (LLMs) maintain similar levels of performance, with Table 4 illustrating that no model outperforms the others significantly. This observation indicates that the total parameter count of these models does not majorly influence their effectiveness in this specific classification task.

Additionally, figures 3, 4, 5, 6, 7, and 8 illustrate the distinctive capabilities of each model in class prediction. GPT-4 excels in detecting negative headlines, as shown in Figure 5, whereas Distil ROBERTA is superior in recognizing headlines with a neutral impact, according to Figure 3. The models exhibit comparable performance in identifying headlines with a positive impact. Since all the LLMs have equal overall good performances, employing a majority voting strategy (refereed equivalently as bagging) could enhance performance, as suggested by [Abburi et al., 2023].

Additionally, the market-adapted fine-tuned models have been made publicly accessible and open source on HuggingFace, available at FinBERT fine-tuned for market data and DistilROBERTA fine-tuned for market data.

5.4 MAJORITY VOTE CLASSIFICATION

To exploit the strengths of each model we used the ensemble method for providing a final classification, as described in section 6. We selected the majority class given by an ensemble of LLMs that presents similar individual performances. For each of the headline, we assign the most given sentiment by the LLMs. We provide in the table 5 the performances of the several model ensemble that we did.

Below is the list of the three different bagging configurations, including SFT models.

- Bagging 1: SFT GPT + SFT Distil ROBERTA + SFT FinBERT
- Bagging 2: SFT Distil ROBERTA + SFT FinBERT
- Bagging 3: All the models SFT and Not

Table 5: Ensemble Method Performances

Model	Precision	Recall	F-score
Bagging 1	0.55	0.53	0.51
Bagging 2	0.53	0.52	0.52
Bagging 3	0.53	0.53	0.52

The ensemble approach failed to enhance overall efficacy markedly. Additionally, the SFT GPT model marginally outperforms the ensemble techniques, indicating noteworthy insights into each Large Language Model's (LLM) unique classification capabilities. The inability to achieve a substantial improvement in overall performance and the non-confirmation of Condorcet's jury theorem underscore the dependency among models, which exhibit comparable responses to the classification challenge. This clarifies that variations in a model's parameter count do not influence its behavior in addressing this specific classification task.

6 MAJORITY VOTE CLASSIFIER AND NON INDEPENDENCE

The bagging (Bootstrap Aggregating) method aims to improve the stability and accuracy of machine learning algorithms by combining the predictions of several base estimators.

6.1 MATHEMATICAL FORMALISM

Let X be the input space, and $Y = \{y_1, y_2, \ldots, y_k\}$ the output space of labels. Consider a set of base classifiers $\{C_1, C_2, \ldots, C_n\}$, where each classifier $C_i : X \to Y$ maps an input $x \in X$ to a label in Y. The bagging model also referred to as the majority classifier model aggregates the individual classifiers to chooses the label with the majority count.

Definition 6.1. Majority classifier: The predicted label $\hat{y}$ for an input $x \in X$ by the bagging classifier C_{bag} is determined by the majority vote among the base classifiers as follows:

$$\hat{y} = \underset{y \in Y}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbf{1}_{C_i(x)=y}, \tag{2}$$

If multiple labels y receive the same highest number of votes from classifiers, we take the average label, provided it does exist (for instance 0 if -1 and 1 are the highest classes). Otherwise, we take randomly one of the best labels.

6.2 CONDORCET'S JURY THEOREM

The Condorcet theorem [de Condorcet, 1785], states that if there is a majority preference for one option over all others in pairwise comparisons, then that option should be chosen as the overall winner in a collective decision-making process, provided that individual classifiers are better than a random guess, there are independent and follow the same law. More formally:

Definition 6.2. IWT Ensemble: We say that $\{C_1, C_2, \ldots, C_n\}$ is an IWT Ensemble (Independent and Well-Trained) if the classifiers satisfy the following conditions:

1. Independence: The base classifiers make their predictions independently of each other.
2. Identical Distribution: The underlying labels Y are drawn from the same distribution.
3. Better Than Random: Each classifier C_i has an accuracy better than random guessing, i.e., for each C_i, the probability $P(C_i(x) = y | x \in X, y \in Y) > \frac{1}{|Y|}$, where $|Y|$ denotes the total count of distinct labels. For binary classification scenarios, this threshold traditionally stands at 0.5. In our specific context, where there are three possible classes (positive, neutral, or negative), the threshold adjusts to one-third.

If only the first two conditions hold, we call the classifiers set a independent classifiers ensemble or ICE.

Theorem 6.1. Condorcet Jury Theorem for Classifiers: For an IWT Ensemble $C_1, C_2, \ldots, C_n$, the majority vote classifier C_{bag} exhibits higher accuracy than any individual classifier within the ensemble. Conversely, if the classifiers perform worse than random guessing, then the majority vote classifier will exhibit lower accuracy than the individual classifiers.

Proof. See for instance [Sancho, 2022]. □

Corollary 6.1. Additionally if the independence assumption does not hold, the majority classifier C_{bag} should not perform better than the best classifier.

Proof. Immediate as a consequence of the Condorcet's jury theorem. □

6.3 EXPERIMENTAL REEVALUATION

In our experiment, employing a bagging majority vote classifier on financial text classification did not result in an significant improvement, challenging the validity of the IWT assumtions. As model precision are all above one third (the random guess baseline on a three classes classification), this suggests that the various LLMs models are not independent and share similarity in their behaviors.

7 CONCLUSION

This paper evaluates the effectiveness of Large Language Models (LLMs) in financial text classification, specifically comparing the performance of GPT-4-1106-preview (1.76 trillion parameters) with FinBERT (110 million parameters) and FinDROBERTA (82.1 million parameters). Our findings indicate that the standard GPT-4-1106-preview model performs less effectively than smaller models like FinBERT and DistilROBERTA, which demonstrate significant performance improvements when fine-tuned. Moreover, when both smaller models and GPT-4 undergo fine-tuning, the enhancements observed are minimal, suggesting that the benefits of fine-tuning are largely consistent across different model sizes. Furthermore, our experiments with a bagging majority classifier fail to produce notable performance enhancements, revealing a failure of Condorcet's jury Theorem assumptions and indicating a lack of model independence. This suggests that the fine-tuned models exhibit similar behavior patterns in analyzing sentiment scores of financial news, regardless of their size.

References

Harika Abburi, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. Generative ai text classification using ensemble llm approaches, 2023.

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

Marie Brière, Karen Huynh, Olav Laudy, and Sébastien Pouget. Stock market reaction to news: Do tense and horizon matter? *Finance Research Letters*, 58, Part D:104630, 2023. ISSN 1544-6123. doi: 10.1016/j.frl.2023.104630. URL https://www.sciencedirect.com/science/article/pii/S1544612323010024.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation, 2023.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data, 2022.

Tyler Cowen and Alexander T. Tabarrok. How to Learn and Teach Economics with Large Language Models, Including GPT. *SSRN Electronic Journal*, XXX(XXX):0–0, 3 2023. ISSN 1556-5068. doi: 10.2139/SSRN.4391863. URL https://papers.ssrn.com/abstract=4391863.

Min-Yuh Day and Chia-Chou Lee. Deep learning for financial sentiment analysis on finance news providers. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1127–1134. IEEE, 2016.

Marquis de Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix, 1785.

F.C. Dumiter, F. Turcaș, Ș.A. Nicoară, C. Bențe, and M. Boiță. The impact of sentiment indices on the stock exchange—the connections between quantitative sentiment indicators, technical analysis, and stock market. *Mathematics*, 11:3128, 2023. doi: 10.3390/math11143128. URL https://doi.org/10.3390/math11143128.

Anne Lundgaard Hansen and Sophia Kazinnik. Can ChatGPT Decipher Fedspeak? *SSRN Electronic Journal*, XX(XX):XX, 3 2023. ISSN 1556-5068. doi: 10.2139/SSRN.4399406. URL https://papers.ssrn.com/abstract=4399406.

Alex G Kim and Valeri V Nikolaev. Context-based interpretation of financial information. *Chicago Booth Research Paper*, 23(08), June 2023. doi: 10.2139/ssrn.4317208. URL https://ssrn.com/abstract=4317208.

Anton Korinek. Language Models and Cognitive Automation for Economic Research. *Cambridge, MA*, XX(XX): XX, 2 2023. doi: 10.3386/W30957. URL https://www.nber.org/papers/w30957.

Baptiste Lefort, Eric Benhamou, Jean-Jacques Ohana, David Saltiel, Beatrice Guez, and Damien Challet. Can chatgpt compute trustworthy sentiment scores from bloomberg market wraps?, 2024. URL https://arxiv.org/abs/2401.05447.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

Lezhi Li, Ting-Yu Chang, and Hai Wang. Multimodal gen-ai for fundamental investment research, 2023.

Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *SSRN Electronic Journal*, XXX(XX-XX):XX, 4 2023. ISSN 1556-5068. doi: 10.2139/SSRN.4412788. URL https://papers.ssrn.com/abstract=4412788.

Lefteris Loukas, Ilias Stogiannidis, Odysseas Diamantopoulos, Prodromos Malakasiotis, and Stavros Vassos. Making llms worth every penny: Resource-limited text classification in banking. In *4th ACM International Conference on AI in Finance*, ICAIF '23. ACM, November 2023. doi: 10.1145/3604237.3626891. URL http://dx.doi.org/10.1145/3604237.3626891.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8:131662–131682, 2020. doi: 10.1109/ACCESS.2020.3009626.

Shakked Noy and Whitney Zhang. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *SSRN Electronic Journal*, XX(XX):XX, 3 2023. doi: 10.2139/SSRN.4375283. URL https://papers.ssrn.com/abstract=4375283.

Álvaro Romaniega Sancho. On the probability of the condorcet jury theorem or the miracle of aggregation. *Mathematical Social Sciences*, 119:41–55, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25, 2018.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*, 2023.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models, 2023a.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. Prompt-based meta-learning for few-shot text classification. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1357, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.87. URL https://aclanthology.org/2022.emnlp-main.87.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check, 2023b.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, Ninghao Liu, and Tianming Liu. Revolutionizing finance with llms: An overview of applications and insights, 2024.

Additional Experimental Results

B.Lefort[1,2] **E.Benhamou**[1,3] **JJ.Ohana**[1] **B.Guez**[1] **D.Saltiel**[1] **D.Challet**[2]

[1]Ai for Alpha
[2]Centrale Supelec
[3]Paris Dauphine PSL

A CLASSIFICATION CONFUSION MATRIX

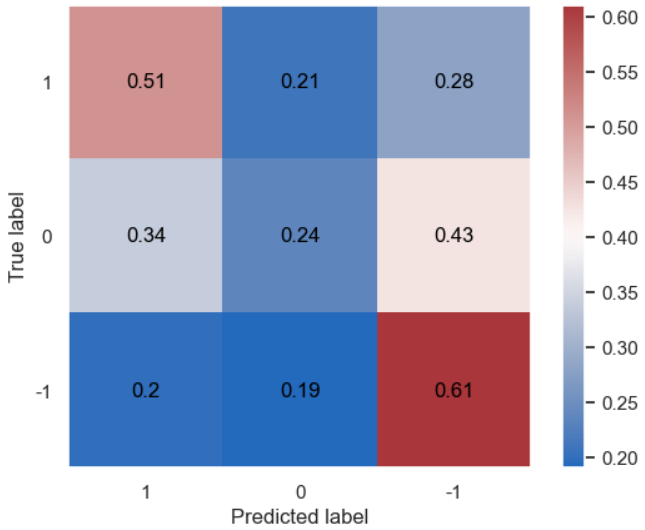

Figure 3: Confusion matrix with proportion of correct classification by class against the market-based baseline using Distil ROBERTA

Accepted for the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024).

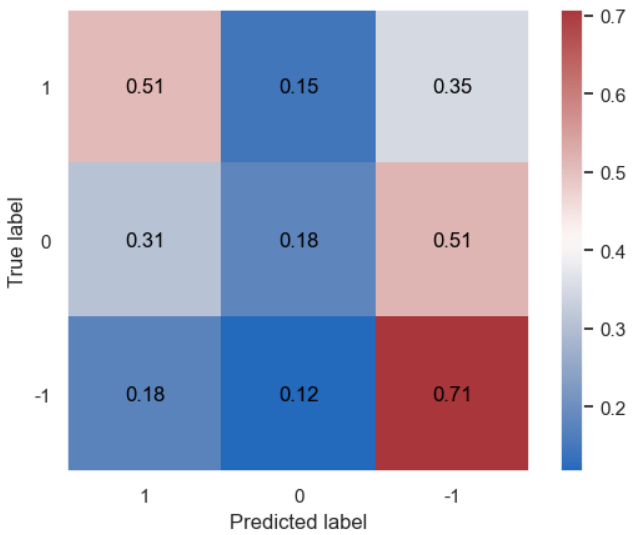

Figure 4: Confusion matrix with proportion of correct classification by class against the market-based baseline using FinBERT

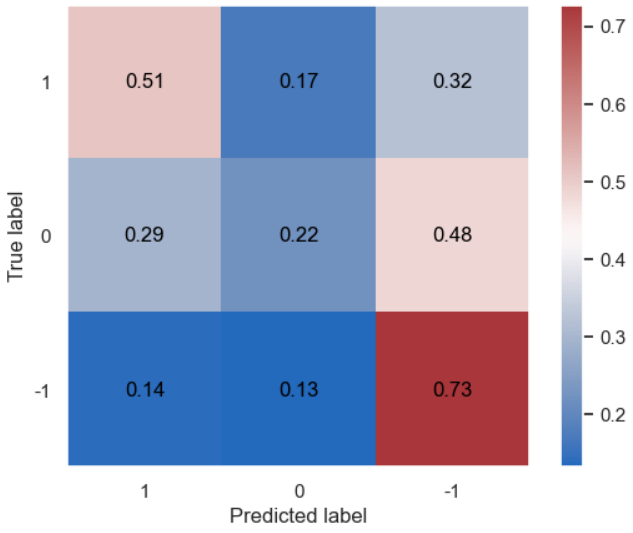

Figure 5: Confusion matrix with proportion of correct classification by class against the market-based baseline using GPT-4

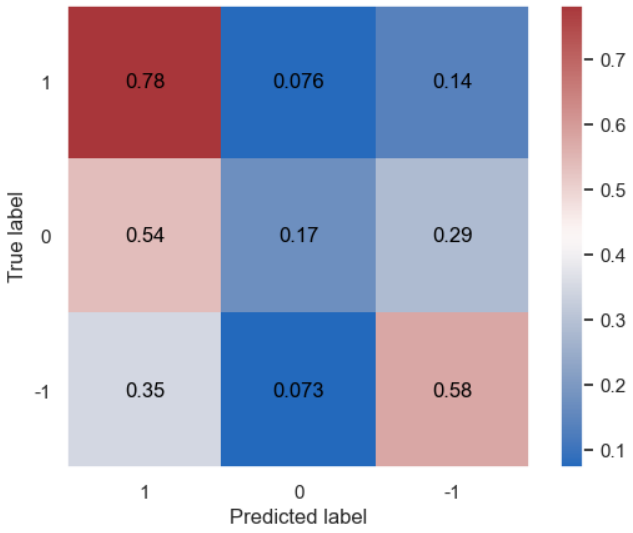

Figure 6: Confusion matrix with proportion of correct classification by class against the market-based baseline using Fine-tuned Distil ROBERTA

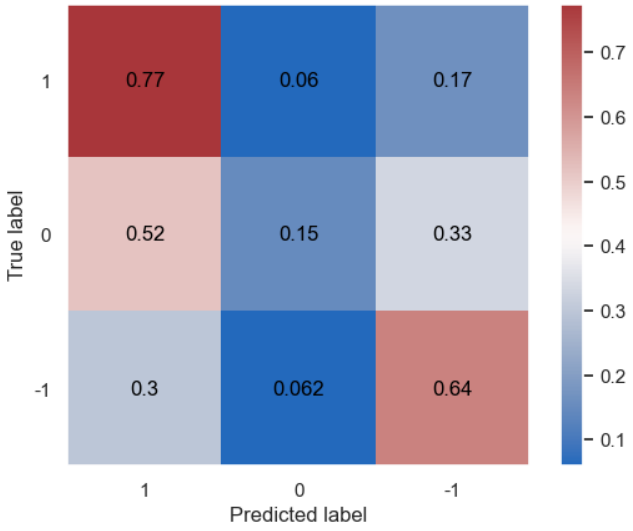

Figure 7: Confusion matrix with proportion of correct classification by class against the market-based baseline using Fine-tuned FinBERT

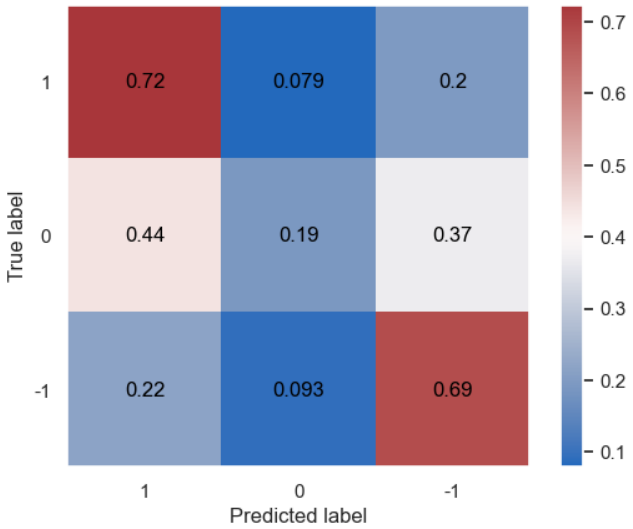

Figure 8: Confusion matrix with proportion of correct classification by class against the market-based baseline using Fine-tuned GPT