Financial Sentiment Analysis from News Headlines and Tweets

Aysun Can Türetken

October 29, 2023

Abstract

In today's fast-paced financial markets, understanding sentiment is pivotal. News headlines and tweets convey sentiments about financial assets, requiring the ability to discern positive, negative, and neutral sentiments for informed decisions. Traditional deep learning models in financial sentiment analysis are giving way to fine-tuning large language models (LLMs) pretrained on small textual corpora with sentiment labels, which offer superior performance and streamlined training. Challenges linked to working with limited datasets are currently tackled by using multi-task learning or multi-instruction fine-tuning. This study demonstrates that leveraging a more extensive training dataset with recent 7-billion-parameter LLMs achieves state-of-the-art sentiment classification performance without requiring any of these techniques. The practicality of this approach offers a promising path for similar NLP tasks in the field.

1 Introduction

Financial sentiment analysis plays a pivotal role in today's dynamic and data-driven financial markets. Understanding the sentiment surrounding stocks, commodities, and currencies, as expressed in news headlines and tweets, is essential for informed decision-making. Positive sentiment can boost market confidence and drive up asset prices, while negative sentiment can trigger panic and precipitate sell-offs. With the rapid dissemination of information on social media and news platforms, real-time sentiment analysis becomes indispensable for investors and financial institutions.

In the context of this project, my main aim is to perform sentiment analysis on financial texts, involving news headlines and tweets, specifically categorizing sentiments into three groups: neutral, positive, and negative. This classification not only provides a solid foundation for making investment decisions but also enables us to leverage publicly available datasets that already come with sentiment these labels.

Modern techniques in financial sentiment analysis have witnessed a paradigm shift, with an increasing reliance on deep learning models for sentiment classification tasks. In the earlier stages of development, these models were exclusively trained from scratch using financial data. However, a recent trend has emerged, focusing on fine-tuning large language models (LLMs) originally pre-trained on extensive text corpora to adapt them for specialized financial tasks. This approach has demonstrated not only superior performance but also greater ease of training, as only a small subset of LLM parameters needs to be fine-tuned.

Many of these LLM-based fine-tuning methods have leaned towards using relatively small datasets, often comprising just a single dataset of few thousand text samples. This, unfortunately, often leads to issues such as dataset bias and over-fitting, as models become overly attuned to the idiosyncrasies and the distribution of the limited training data.

To address these limitations, recent methods usually opt for a multi-task learning objective, encompassing sentiment analysis classification alongside several other tasks such as named entity recognition, question answering and news headline classification. This

broader training scope allows models to be exposed to a more diverse range of financial datasets, hence improve their overall performance. Alternatively, other strategies aim to enrich the meager sentiment data available by providing a wide array of instructions to guide label output during fine-tuning.

In contrast, in this work, I will demonstrate that leveraging a more extensive training dataset yields state-of-the-art (SOTA) performance in sentiment classification, all without the complexities associated with multi-task learning frameworks or requiring multi-instruction-based fine-tuning. Through the fine-tuning of two recent 7-billion-parameter LLMs on four publicly-accessible financial datasets, each annotated with sentiment labels, this approach consistently attains superior or comparable accuracy while retaining a high level of practicality.

2 Related Work

Recent advances in financial sentiment analysis have predominantly revolved around training Large Language Models (LLMs) using various transformer architectures. While these approaches have outperformed traditional methods like Support Vector Machines (SVMs) and decision trees, they demand extensive textual data to capture the nuances of financial language. Domain-specific datasets, exemplified by the proprietary 50-billion parameter BloombergGPT [1], have proven effective for various financial tasks but are not readily accessible to the broader research community due to their proprietary nature and substantial computational requirements. In contrast, publicly available task-specific labeled financial datasets, while accessible, often suffer from limited scale, presenting a significant challenge for training LLMs.

One strategy to address this challenge involves fine-tuning pre-trained LLMs, such as the foundational LLaMA models [2], obtained through training on large public and general-purpose English text. This approach may also entail a further pre-training phase without ground truth annotations on domain-specific financial corpora before task-specific fine-tuning with class labels as done by FinBERT [3]. Nevertheless, pre-training poses inherent risks, including the potential for knowledge loss relevant to the final task and target task-specific domain adaptation issues.

Another approach is to employ a multi-task learning framework, in which the model is trained on multiple small financial datasets for various related objectives including named entity recognition, question answering, news headline classification, stock movement prediction in addition to sentiment classification [4], [5]. This allows the model to tap into a larger amount financial labelled data and hence shown to perform better than training the same model on each task with its associated small dataset.

Recent efforts for building task-specific models also include using multiple instructions with the same or similar meanings during model fine-tuning [6] or Chain-of-Thought (CoT) prompting [7]. Although providing improvements over their respective baselines, however, all these past models fall short of tackling the original problem of having limited amount of task-specific training data with sentiment labels. Solving this limitation is arguably what brought many supervised foundational models, such as ChatGPT [8], their exceptional capabilities thanks to the high quality of human-labelled datasets they are trained on.

This study illustrates that by harnessing a larger training dataset and utilizing recent 7-billion-parameter LLMs, state-of-the-art sentiment classification performance can be attained without resorting to the aforementioned methodologies. The viability of this approach underscores the significance of dataset size in supervised fine-tuning of LLMs for domain-specific financial tasks.

3 Datasets

An exhaustive exploration of publicly available web datasets uncovered four English financial sentiment datasets, as detailed in the subsequent subsections. These datasets vary significantly in size, posing a challenge: a straightforward concatenation for training would disproportionately favor the largest dataset, potentially diminishing performance on the smaller ones. To address this, I balanced out the datasets by replicating samples within each dataset to align its sample count closely with the largest one (i.e., the NGI dataset). This resulted in a combined training dataset of 68659 samples which I used used for fine-tuning the LLMs. The resulting samples were randomly shuffled to avoid any dataset-specific biases that may arise during fine-tuning.

3.1 Twitter Financial News Sentiment (TFNS)

This dataset is comprised in news tweets that are obtained in the context of Twitter discussions and hence contain financial jargon such as stock tickers, indices and financial abbreviations. It includes 9540 text samples for training, each annotated with one of three investment-related labels: bearish, bullish, and neutral. I mapped bearish and bullish respectively to negative and positive. The dataset also contains a validation split of 2390 samples, which I used for tests and comparisons of the pretrained models.

3.2 Financial Phrasebank (FPB)

This is one of the most frequently used and cited datasets in the financial sentiment analysis field. It consists of financial news downloaded from the LexisNexis database using an automated web scraper. They are sampled randomly from a subset of 10000 articles to obtain good coverage across small and large companies. To ensure high-quality annotations, each text sample is categorised into the three sentiment classes that I use in this work by 5-8 annotators with backgrounds in finance and business. As there are multiple annotators, the dataset comes in multiple sizes depending on the percentage of agreeing annotators: 50agree, 66agree, 75agree, and 100agree. In this work, since most of the financial analysis techniques use the 50agree subset, I used this subset, which contains 4840 sentences. Furthermore, as there is no split provided in the original dataset, I used a randomly sampled 90% for training and 10% for tests.

3.3 Financial Opinion Mining and Question Answering (FiQA)

This dataset is introduced as a financial opinion mining challange that took place in Lyon France in 2018. The annotations are continuous label scores ranging from +1 signifying positive sentiment to -1 carrying negative sentiment. I converted these continuous scores to the three sentiment classes using two thresholds values of -0.33 and +0.33. The dataset includes train, validation and test splits of 961, 102 and 150 samples respectively. I used the train and validation splits for training and the test split for evaluation of the trained models.

3.4 News with GPT Instructions (NGI)

This dataset is comprised in news headlines each assotiated with five labels: strong positive, moderately positive, mildly positive, neutral, mildly negative, moderately negative, andstrong negative. For training with thee labels, I mapped 'strong' and 'moderately' positive/negative labels into their respective positive/negative classes. Mildly positive/negative and neutral labels are all mapped to neutral. The dataset comes with 16200 train samples and 405 test samples, which I respectively used for training and tests.

4 Preprocessing

The text samples in the final combined dataset are structured to incorporate an instruction (also called, a task prompt) that clearly states the sentiment classification task. As shown by earlier fine-tuning approaches on LLMs, this allows for a better zero-shot performance and hence serves as a good initialization for fine-tuning, and improves convergence. I use the following template for instructions:

```
### INSTRUCTION: [task prompt] ### TEXT: [text input] ### ANSWER: [output]
```

In contrast to [6], which uses multiple similar task prompts during fine-tuning, in this work, I used a single one given below, which proved to be sufficient for good performance:

```
Classify the financial sentiment in the following text into one of the
three classes: negative, neutral, or positive. Give a single word answer
(either negative, neutral or positive).
```

During training and inference, the models are fed with statements following the above template except for the [output] segment, which serve as our labels and hence can only be one of the three sentiment words {negative, neutral, negative}. For the specialized tokenizers that come with the baseline pre-trained LLMs, each of these three words translate into a single token, which speeds up the inference as there is only a single token to be generated by the fine-tuned LLM.

5 Model Training

In this work, two models were adapted for the financial sentiment analysis task: LLaMA-2-7B-Chat and Mistral-7B–Instruct. The LLaMA-2-7B-Chat has been widely used by the research community and practitioners in many downstream tasks including the ones in the financial domain and serves as a good baseline to compare against. On the other hand, the Mistral-7B–Instruct model is much more recent (Oct 10) and has been shown to outperform all existing 7-billion parameter models in addition to several larger ones, including LLaMA-2-13B, on various NLP benckmarks [9].

Both models were fine-tuned on the final combined dataset presented in the dataset section using the Hugging Face's SFTTrainer. For fast training and inference, the sequence length is fixed to 768 tokens, which is larger than the largest input token size in the combined dataset including the instruction phrase introduced earlier.

To minimize the memory usage, both models were loaded into the memory in 4-bit precision and fine-tuned with the efficient QLoRA approach [10], which performs memory-efficient 4-bit precision training of the fine-tuned (modified) parameters. For LLaMA-2-7B-Chat and Mistral-7B–Instruct, the fraction of the fine-tuned parameters were around 0.5% and 1.5% of the total number of trainable parameters respectively. Both models were fine-tuned for one epoch and with a batch size of 2 samples. For LLaMA-2-7B-Chat, the learning rate and weight decay were set to 1e-4 and 1e-2 respectively, while Mistral-7B–Instruct, a lower learning rate (2e-5) and larger weight decay (1e-2) were used to counteract over-fitting.

The training was performed on a single A100 GPU with 40GB memory space using Google Colab. Both models took around 31.3GB during training and less than 15GB during inference. For both models, the training took approximately 6 hours and the evaluation on the combined test dataset less than 10 minutes. The total cost of the development cycle including model fine-tuning was less than 55 CHF. The fine-tuned models are hosted in the Hugging Face website.

6 Evaluation

I evaluated the two fine-tuned models as well as their original baselines (zero-shot) on my combined test dataset of 7070 samples. For a fair evaluation, I generated 10 tokens for the baselines, instead of only 1 used for the fine-tuned models, and the then checked if generated tokens contain any of the allowable three sentiment class labels. If no sentiment token could be found, the predicted label fell into a fourth 'undefined' category.

The results presented in Table 1 show that both methods give similar performance while Mistral-7B-Instruct performs slightly better on three datasets out of four. However, both fine-tuned models show significant improvements over their baselines, which shows the effectiveness of task-pecific fine-tuning. For more detailed results including per-class precision, recall, F1 scores and confusion matrices, please refer to the results directory inside the source code folder.

My model achieve more than 90% accuracy and weighted F1 score on the TFNS dataset By contrast, the Instruct-FinGPT model [6] trained also on the TFNS and FiQA datasets remain below 90% on both metrics. The multi-task FinMA-30B model [4] achieves slightly lower results on FPB but much better results on FiQA. What is very surprising is that the GPT-4 model achieves close to SOTA performance while being fine-tuned only on 5 samples on these datasets. This impressive result demonstrates the importance of deep language understanding, which can be achieved by training LLMs on large and high-quality datasets, in adapting these models for domain-specific tasks through fine-tuning.

Method	FPB		TFNS		FiQA		NGI	
	F1	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**
LLaMA-2-7B-Chat ZS	0.53	0.55	0.26	0.34	0.33	0.40	0.52	0.58
MISTRAL-7B-Instruct ZS	0.58	0.61	0.48	0.52	0.47	0.53	0.58	0.63
LLaMA-2-7B-Chat FT	0.89	0.89	0.91	0.91	0.71	0.71	0.69	0.69
MISTRAL-7B-Instruct FT	0.89	0.89	0.90	0.90	0.75	0.75	0.70	0.70
FinBERT[3]	0.84	0.86	0.67	0.73	—	—	—	—
Instruct-FinGPT[6]	—	—	0.84	0.88	—	—	—	—
FinMA-30B[4]	0.88	0.87	—	—	0.87	—	—	—
BloombergGPT[1]	0.51	—	—	—	0.75	—	—	—
GPT-3 5-Shot[7]	0.78	0.78	—	—	0.78	—	—	—
GPT-4 0-Shot[7]	0.83	0.83	—	—	0.87	—	—	—
GPT-4 5-Shot[7]	0.86	0.86	—	—	0.88	—	—	—

Table 1: Weighted average F1-Score and accuracy values for the evaluated approaches and datasets. The terms ZS and FT signify zero-shot and fine-tuned versions of the models.

7 Conclusion

This work showcases the remarkable potential of leveraging recent 7-billion-parameter LLMs for achieving state-of-the-art sentiment classification in financial markets, obviating the need for complex multi-task learning or multi-instruction fine-tuning. This practical approach not only enhances performance but also paves the way for more efficient and effective solutions in the realm of natural language processing for similar financial tasks.

Another potential axis of research would be to better define the sentiment classification problem by adding the entity name to be predicted in addition to the sentiment class. This could lead to more fine-grained and context-aware sentiment analysis, further enhancing the utility of these models in financial decision-making and beyond. Future work in this area may also involve exploring the adaptability of these models to other NLP tasks and further enhancing their robustness in dynamic financial environments.

References

[1] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[4] Q. Xie, W. Han, X. Zhang, Y. Lai, M. Peng, A. Lopez-Lira, and J. Huang, "Pixiu: A large language model, instruction data and evaluation benchmark for finance," *arXiv preprint arXiv:2306.05443*, 2023.

[5] R. S. Shah, K. Chawla, D. Eidnani, A. Shah, W. Du, S. Chava, N. Raman, C. Smiley, J. Chen, and D. Yang, "When flue meets flang: Benchmarks and large pre-trained language model for financial domain," *arXiv preprint arXiv:2211.00083*, 2022.

[6] B. Zhang, H. Yang, and X.-Y. Liu, "Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models," *arXiv preprint arXiv:2306.12659*, 2023.

[7] X. Li, X. Zhu, Z. Ma, X. Liu, and S. Shah, "Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks," *arXiv preprint arXiv:2305.05862*, 2023.

[8] OpenAI, "Gpt-4." https://platform.openai.com/docs/models/gpt-4.

[9] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[10] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[11] HuggingFace, "Twitter financial news sentiment (tfns)." https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment.

[12] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 782–796, 2014.

[13] HuggingFace, "Financial phrasebank (fpb)." https://huggingface.co/datasets/financial_phrasebank.

[14] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur, "Www'18 open challenge: financial opinion mining and question answering," in *Companion proceedings of the the web conference 2018*, pp. 1941–1942, 2018.

[15] HuggingFace, "News with gpt instructions (ngi)." https://huggingface.co/datasets/oliverwang15/news_with_gpt_instructions.

[16] OpenAI, "Gpt-3.5." https://platform.openai.com/docs/models/gpt-3-5.