Title: Financial Sentiment Analysis in General vs Domain Specific Language Models

Course: CSPB 4380, Spring 2025
Natural Language Processing
Author: Sulman Haque
Date: 05/05/2025

ABSTRACT:

Financial Sentiment Analysis (FSA) on documents can serve as a useful heuristic for investing professionals when evaluating and monitoring investments. FSA poses unique challenges including ambiguity on the rating of sentiment, reduced tonality and expressiveness of written financial information, and numerical relationships in the text to determine positive or negative outlooks, among others. Recently Large Language Models have been applied to sentiment analysis in the financial domain yielding many different approaches with varying results. Building on previous work, we establish baseline performance on sentiment classification for a general language model: Bert-based-uncased and a financial domain model: FinBERT-tone. Further we evaluate the improvement of fine-tuning a Bert-based-uncased model and discuss the size of the tuning dataset required to improve performance. The dataset used for training, validation and testing is the Financial PhraseBank. We observe across baseline FinBERT-tone performed much better than BERT-based-uncased. After fine-tuning the BERT-base, it eventually matched and even exceeded the performance of FinBERT-tone. This indicates a fine-tuned general model can improve and surpass a domain specific model in certain tasks.

INTRODUCTION:

Industry Need

Decision-making in the financial industry is heavily reliant on numerical and textual data. Financial sentiment is often a useful heuristic that can influence investor behavior, market trends, and algorithm trading decisions. Automating this task can help institutions monitor public/company sentiment in real-time. Key to automating is understanding which approaches will increase the performance of sentiment analysis benchmarks.

Technical Challenge

FSA poses many challenges. First, there is openness in consensus of what is a positive, negative, or neutral sentiment depending on individual perspectives. One analyst could view a temporary drop in revenue as a good sign if it is coupled with larger revenue streams in the future, if a company is shifting its product strategy to pursue larger markets in the future. Others will likely assume a negative sentiment of lower revenues. Second, financial text is rarely an expressive language as other forms of written text. Often financial text aims to minimize tonality and just report figures and provide information. There is much less context to work with as in other sentiment analysis domains such as social media. Finally, financial documents heavily include numbers, and the specifics of those numbers in relation to other numbers is a critical to assess if something is a positive or negative sentiment.

Research Goals

We will evaluate the performance of a general vs domain specific models, then compare the impact of fine-tuning on general models to achieve better performance compared to the domain specific models, accounting for the previous challenges.

RELATED WORK:

Survey of LLMs in the Financial Domain

The evolution of large language models (LLMs) in the financial domain has been growing in complexity along-side the general development of LLMs. Lee et al (2024) captures the progression of publicly available domain specific models and their inheritance from general models.

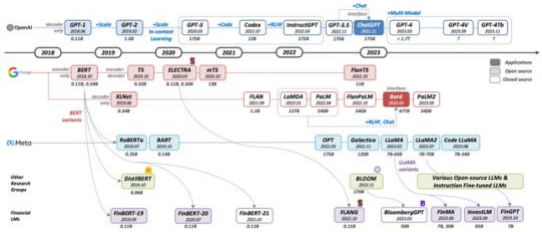

Figure 1: Evolution of Financial LLMs

Original BERT

BERT is a well-known encoder only model developed by Devlin et al (2019). Two separate models were developed, Bert-Base and Bert-Large of 110M and 340M parameters respectively. There were two variants of each: cased and uncased.

FinBERT

FinBERT was developed by Yang et al. (2020). It is one of the most widely adopted models in the space. It extended the Bert-base models with additional pre-training on 4.9 billion tokens of financial text including 10-Ks, 10-Qs, earnings calls, analyst reports. Yang et al constructed a domain-specific tokenizer FinVocab, comprising ~30K tokens to better capture financial terminology. They released multiple variants of FinBERT using both cased and uncased vocabularies and evaluated the model on benchmark datasets including the Financial PhraseBank, AnalystTone, and FiQA. Results demonstrated notable performance gains over baseline BERT models, with the cased variant improving accuracy by ~10%.

FinBERT Extended

A separate work by Araci (2019) took a complementary approach by further pretraining FinBERT on the TRC2 Reuters corpus, filtering the data for financial relevance. This study compared FinBERT with traditional and semi-contextual models like LSTM, ULMFit, and ELMo, and found FinBERT consistently outperformed them on accuracy, loss, and F1-score. They also evaluated different pretraining and fine-tuning strategies, finding that additional domain-specific pretraining yielded modest gains, suggesting the base FinBERT was already strong for sentiment classification.

FinGPT (Instruction Tuning)

Expanding beyond BERT-based models, Zhang et al. (2023) introduced Instruct-FinGPT, which applied instruction tuning to general-purpose models like LLaMA-7B for financial sentiment analysis. By reframing the classification task as a generative problem and fine-tuning on instruction-response pairs derived from labeled financial datasets, their model outperformed both FinBERT and untuned LLaMA on accuracy and F1 metrics. Notably, this approach leveraged the inherent contextual capabilities of large language

models (LLMs), capturing nuances such as earnings surprises more effectively.

InvestLM
A broader application of instruction tuning is explored by Yang et al. (2023) in InvestLM, a financial domain LLM fine-tuned from LLaMA-65B using Low-Rank Adaptation (LoRA) and Linear Rope Scaling to support longer contexts. The model was benchmarked against GPT-3.5, GPT-4, and Claude 2 using a human-evaluated dataset of 30 financial questions. InvestLM was preferred by financial experts in most cases, suggesting the effectiveness of instruction tuning for general financial reasoning, not just sentiment.

Together, these papers illustrate a spectrum of techniques—from domain-specific pretraining to instruction tuning—each offering unique advantages. While FinBERT remains a strong baseline for financial sentiment tasks, other approaches are gaining traction that rely on general models with more emphasis on instruction tuning.

DATA:

The main dataset used in this paper was the Financial PhraseBank from Malo. Et al. (2014). It consists of 4840 English sentences from financial news found on the LexisNexis database. The sentiment annotations were done by 16 people with backgrounds in finance, accounting, and economics. The dataset has four possible configurations based on agreement level by the annotators. Table 1 shows the distribution of samples by agreement level. For our analysis, we selected the >=50% agreement dataset to maximize the amount of data we have for training and testing. The distribution of positive, negative, sentiment labels are as follows:

Dataset Analysis
- Positive: 1363 sentences.
- Negative: 604 sentences.
- Neutral: 2879 sentences

Table 1: Financial PhraseBank Configuration

Agreement level	Positive (%)	Negative (%)	Neutral (%)	Count
100%	25.2%	13.4%	61.4%	2264
>=75%	25.7%	12.2%	62.1%	3453
>=66%	27.7%	12.2%	60.1%	4217
>=50%	28.1%	12.4%	59.4%	4846

We split the dataset into training and testing following the conventional 80-20 split. This resulted in 3876 training samples with the same distribution of positive, negative, and neutral sentiments, and 970 for testing.

We observe heavy bias in the number of neutral sentences and must account for this in our results. Additionally, we must remap labels to sentiment to ensure consistency between the datasets and the models.

METHODOLOGY

Model selection
Computational limitations restricted the selection of models in this paper. Google Collab provides a single T4 GPU for quick demonstrations and contains 16GB of RAM. The billion parameter models require up to 28GB of RAM, with some quantization methods that can reduce it to 12GB. In this paper, we focus on demonstration purposes

and select from a range of encoder only models in the 110M parameter range.

The generic model selected was Bert-based-uncased 110M. Through the work of Yang et al. (2020), the uncased models perform better than the cased models. Since there is no additional pre-training in this effort, so the most performant off the shelf model was selected.

For the financial domain-specific model, FinBERT-tone 110M parameter was used. FinBERT-tone is a base FinBERT model with additional fine-tuning for sentiment classification based on 10K manually annotated examples of positive, neutral, and negative.

Tools / Libraries
The following libraries were used:

- Transformers – hugging face library for interfacing with LLMs
- Datasets – hugging face library to interface with shared datasets
- Torch – open-source ML library with tensor GPU computing functionality
- Various data science libraries: scikit, matplotlib, etc.

Preprocessing
Preprocessing steps consisted of a few phases:

- Splitting testing and training datasets
- Remapping category labels to appropriate sentiment
- Configuring Bert-based-uncased model with 3 predictors.

Baseline Experiment:
The baseline experiment was to compare the Bert-based-uncased and FinBERT-tone model on training dataset. Since we were not modifying this dataset, we used the whole dataset to gather the initial baseline results.

Fine-Tuning Bert-Based-Uncased
The fine-tuning process consisted of downloading a fresh bert-based-uncased model and creating five different fine-tuned models, each with different amounts of training. The models were trained with successively larger datasets for 3 epochs per round. The models were then tested on the testing dataset to compare to performance over the baseline and FinBERT-tone models.

The five models were:

- Bert-Based-Uncased-10%
- Bert-Based-Uncased-25%
- Bert-Based-Uncased-50%
- Bert-Based-Uncased-75%
- Bert-Based-Uncased-100%

RESULTS:

We evaluate accuracy, macro F1 scores and F1 scores for each label. F1 scores provide insight into the trade-off between precision and recall.

Baseline
The baseline sentiment classification of Bert-based-uncased and FinBERT-tone had accuracies of 44% and 79% respectively, indicating a better performance with FinBERT-tone. When looking at the F1 scores at the macro and category level, the level of improvement is more drastic, where FinBERT-tone can maintain at least a 68% F1 score for all three categories, lowest being positive sentiment. Bert-base-uncased on the other hand was not able to predict one positive score. It seemed the model learned that predicting neutral would yield it a

sufficient score, which is an artifact of the over representation of neutral examples in the dataset.

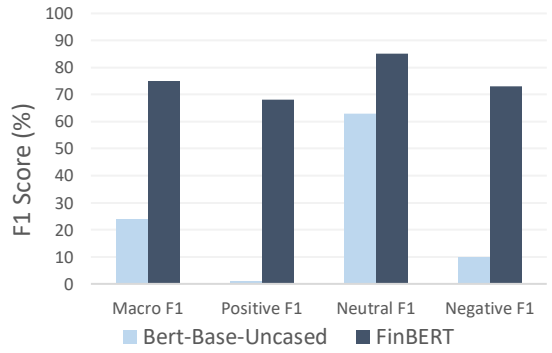

Figure 2 FinBERT vs BERT-base-uncased Baseline F1 Scores

It seems both models generally struggle with positive sentiment examples.

Fine Tuning Performance

After performing fine-tuning on vanilla Bert-based-uncased models, improvements in overall accuracy are observed from 44% to 85%, but this isn't the whole picture.

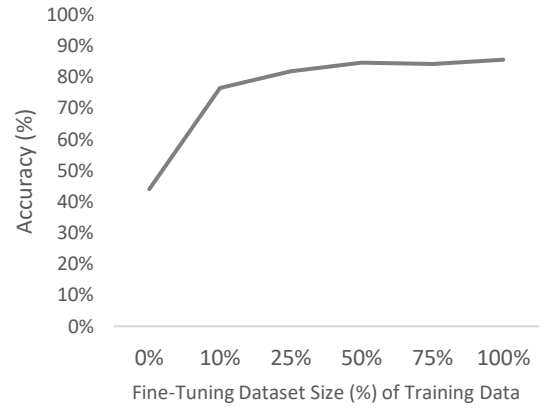

Figure 3 Accuracy of Bert-Based-Uncased Through Progressive Fine-Tuning

Not only do we see improvement in accuracy, but we also see a better F1-scores for each label of Positive Negative and Neutral, implying the model is getting better at predicting the sentiment and it is not merely an artifact of the data. Figure 4 shows the that with just 10% of the dataset being used for fine-tuning, we can achieve a significant performance enhancement in a general model sentiment classification ability.

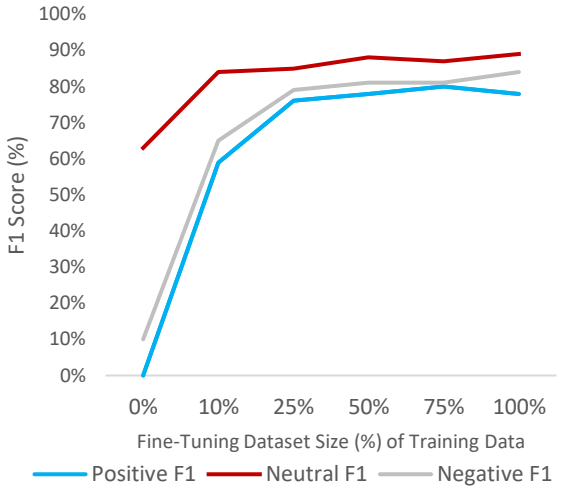

Figure 4 Impact of Fine-Tuning on F1 Scores with Increasing Datasets

We see substantial improvement across categories with slight fine-tuning, however, results appear to plateau around the 80-90% mark with about 25% of the training data. It seems much more data will be required to break past the 80%.

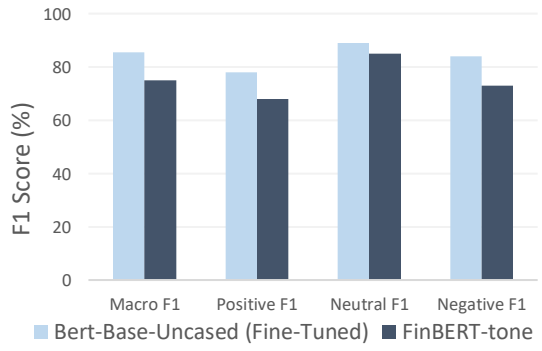

Figure 5 Bert-Base-Uncased fine-tuned exceeds native FinBERT performance.

A final view, Figure 5, comparing back to our baseline, we see that the Fine-Tuned Bert-Based-Uncased was more performant than the FinBERT-tone model in all F1 measures for each category.

DISCUSSION

General vs Domain Specific Models
Evaluating the accuracy and F1 scores of the general bert-based-uncased and the domain-specific FinBERT-tone revealed an unsurprising performance advantage of the FinBERT model in all metrics. However, the BERT-based-uncased, after a minimal amount of fine-tuning, performed on par, even better in some cases, to the vanilla FinBERT-tone model. In this experiment, we can see multiple paths to achieve a desire outcome. The flexibility of language models to accomplish specific tasks and be more beneficial than off the shelf domain models.

Holistic Evaluation with Unbalanced Data
The initial focus was on accuracy, but that quickly proved to be insufficient. The Bert-Base-Uncased baseline score lacked any ability to predict a correct positive label, despite a better than random performance of 44% across the dataset. Upon investigating the F1 scores to evaluate the tradeoff between precision and recall, we saw just how poorly the Bert-Base-Uncased model performed during baselines.
As the fine-tuning continued, we saw dramatic improvements in not only accuracy, but the F1 scores of all the values. This points to real improvement in the model's ability to categorize financial sentiment.

Difficulties in Financial Sentiment Analysis
Regardless of the F1 scores, we see limitations of both fine-tuned and domain specific models reaching past the 80-90% performance range. This is likely due to the challenges discussed in the introduction on difficulties on extracting tone and insight based on numbers alone.

CONCLUSIONS

There are a few conclusions to draw from this experiment:

Domain-specific pretraining works:
FinBERT's training on financial text gives it an edge in understanding the nuances of sentiment in financial language—especially jargon and numerical cues (e.g., "beats", "misses", "guidance"). The work done by Yang et al. 2020 provides sufficient performance for sentiment classification.

Fine-tuning works for specific tasks
Often a domain-specific model is unavailable. Given a good dataset, fine-tuning applied to a base general LLM has shown to rival performance of more domain specific models. It is a viable strategy if performance on a singular task is the end goal.

Positive recall is still the weakest point:
In both efforts, positive recall suffered the most. This indicates that more examples or a more balanced dataset may help with more performance.

There are various avenues to extend this effort:

- Test with additional datasets (FiQA, Twitter Datasets, etc.)
- Explore larger models.
- Explore encoder-decoder models.
- Instruction tuning of encoder-decoder models to accomplish similar results.

BIBLIOGRAPHY

[1] Lee, Jean, Nicholas Stevens, and Soyeon Caren Han. "Large Language Models in Finance (FinLLMs)." Neural Computing and Applications (2025): 1-15.

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.

[3] Yang, Yi, Mark Christopher Siy Uy, and Allen Huang. "Finbert: A pretrained language model for financial communications." arXiv preprint arXiv:2006.08097 (2020).

[4] Araci, Dogu. "Finbert: Financial sentiment analysis with pre-trained language models." arXiv preprint arXiv:1908.10063(2019).

[5] Zhang, Boyu, Hongyang Yang, and Xiao-Yang Liu. "Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models." arXiv preprint arXiv:2306.12659 (2023).

[6] Yang, Yi, Yixuan Tang, and Kar Yan Tam. "Investlm: A large language model for investment using financial domain instruction tuning." arXiv preprint arXiv:2309.13064 (2023).

[7] Malo, Pekka, et al. "Good debt or bad debt: Detecting semantic orientations in economic texts." Journal of the Association for Information Science and Technology 65.4 (2014): 782-796.