Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models

Boyu Zhang[2*], **Hongyang (Bruce) Yang**[1*], **Xiao-Yang Liu**[1†]
[1]Columbia University; [2]Swinburne University of Technology
{HY2500, XL2427}@columbia.edu; boyu.zhang68@gmail.com

Abstract

Sentiment analysis is a vital tool for uncovering insights from financial articles, news, and social media, shaping our understanding of market movements. Despite the impressive capabilities of large language models (LLMs) in financial natural language processing (NLP), they still struggle with accurately interpreting numerical values and grasping financial context, limiting their effectiveness in predicting financial sentiment. In this paper, we introduce a simple yet effective instruction tuning approach to address these issues. By transforming a small portion of supervised financial sentiment analysis data into instruction data and fine-tuning a general-purpose LLM with this method, we achieve remarkable advancements in financial sentiment analysis. In the experiment, our approach outperforms state-of-the-art supervised sentiment analysis models, as well as widely used LLMs like ChatGPT and LLaMAs, particularly in scenarios where numerical understanding and contextual comprehension are vital.

1 Introduction

Financial sentiment analysis, the task of discerning investor sentiment from financial articles, news, and social media, is an essential instrument for comprehending and forecasting market movements. Conventional models often struggle with several difficulties, including insensitivity to numeric values, difficulties interpreting sentiment without explicit context, and the challenges associated with financial jargon, multilingual data, temporal dependency, insufficient labeled data, and the inherent noise in social media data.

Large Language Models (LLMs) have been pivotal in mitigating some of these challenges, demonstrating a significant contribution to the field of financial natural language processing (NLP). One distinguishing feature is LLMs' inherent general knowledge garnered during pre-training on vast and diverse corpora, including financial texts. However, LLMs do not have enough financial context. Their performance in interpreting numerical values is often inadequate and may struggle to accurately determine sentiment when the context is absent or ambiguous. These challenges underline the need for improved models that can adeptly understand the intricate nuances of financial sentiment analysis.

In response to these challenges, our study explores the potential of instruction tuning of general-purpose LLMs for sentiment analysis in the finance sector. In this study, we investigate two primary research questions: ***1) How to enable LLMs to address the issue of numerical sensitivity in financial sentiment analysis? and 2) What is the role of contextual understanding in improving financial sentiment analysis?***

We propose *Instruct-FinGPT* by instruction tuning [Wei *et al.*, 2022] a pre-trained LLM (namely LLaMA [Touvron *et al.*, 2023]). Through this approach, we transform the classification based sentiment analysis dataset into a generation task, thereby allowing LLMs to apply their extensive training and superior analytical capabilities more effectively. The ultimate goal of Instruct-FinGPT is to enhance the performance in financial sentiment analysis by minimizing the requirement of fine-tuning data and maximizing the contextual understanding and numerical sensitivity inherent to LLMs. By introducing this novel method, we aspire to push the boundaries of current methodologies, opening up promising avenues for future exploration in the realm of financial sentiment analysis.

The primary contributions of this paper are as follows:

- We design an instruction-tuned FinGPT model for financial sentiment analysis. This model surpasses both general-purpose LLMs and state-of-the-art supervised models in benchmark performance, despite utilizing only a small amount of instruction data and training resources.
- We address the critical issue of numerical sensitivity in financial sentiment analysis, a component often neglected by existing models, enhancing the model's ability to accurately interpret sentiment from financial news.
- We underscore the importance of contextual understanding in financial sentiment analysis, leveraging the inherent general knowledge of LLMs for improved performance in sentiment analysis, especially when the context is missing or vague.

Our study provides new insights into the application of

[*] Equal contribution.
[†] Corresponding author.

LLMs for financial sentiment analysis, offering potential solutions to some of the enduring challenges in the field.

2 Related Work

The task of sentiment analysis, particularly in the financial domain, has been a significant area of research in the field of Natural Language Processing (NLP). There are several works in literature [Xing *et al.*, 2018; Loughran and McDonald, 2011; Tai and Kao, 2013; Hamilton *et al.*, 2016; Day and Lee, 2016; Chan and Chong, 2017; Sohangir *et al.*, 2018; Araci, 2019; Mishev *et al.*, 2020] that utilize different methodologies for performing financial sentiment analysis, ranging from lexicon-based techniques to machine learning and deep learning approaches.

One noteworthy work is by Araci [Atkins *et al.*, 2018], which presents an ensemble of traditional machine learning algorithms for predicting the direction of stock market movement based on financial news articles. While this work made strides in using machine learning for financial sentiment analysis, it does not extensively address the challenges related to numerical sensitivity or contextual understanding.

In terms of deep learning approaches, the transformer-based model BERT [Kenton and Toutanova, 2019] has been widely used for sentiment analysis tasks due to its powerful context understanding capability. However, BERT and its derivatives typically require substantial amounts of labeled data for fine-tuning, which might be challenging to obtain in the financial domain.

More recently, FinBERT [Araci, 2019], a variant of BERT designed explicitly for the financial domain, was developed to address these issues. FinBERT has been fine-tuned on the financial text and has shown promising results in financial sentiment analysis. Nonetheless, it suffers from limitations such as insensitivity to numerical values and struggles with the context where the necessary information may be missing. FLANG [Shah *et al.*, 2022] additionally presents financial assessment benchmarks across five distinct NLP tasks within the financial sector, along with the incorporation of conventional benchmarks prevalent in prior research.

While BloombergGPT [Wu *et al.*, 2023] demonstrates impressive performance in sentiment analysis tasks, there are inherent challenges to its accessibility and applicability for broader usage. The model, proprietary to Bloomberg, was trained on a vast corpus of specialized financial data, which may not be readily available to others. Moreover, the resources required to train such a model are substantial (1.3M GPU hours, a cost of around $5M). This is in contrast to our approach, which demonstrates substantial effectiveness with a significantly smaller corpus and less computational resources (estimated around less than $300 per training), making it more feasible for wider deployment.

Our work stands distinct in its focus on leveraging the power of LLMs, their inherent general knowledge, and reasoning capabilities to perform sentiment analysis in the financial domain. We explore a novel instruction tuning approach and demonstrate its effectiveness in our experiments.

3 Our Method

Despite the pre-trained LLMs such as GPT-3 and LLaMA can acquire the general abilities for solving various tasks, increasing studies have shown that LLM's abilities can be further adapted according to specific goals. Our approach uses instruction tuning to adapt the general-purpose LLMs to financial sentiment analysis, enhancing their understanding of numerical values and context in this specific task. The process involves transforming the sentiment analysis task from a classification task to a text generation task, which aligns better with the capabilities of LLMs. Further, we use the transformed dataset to instruction finetune the LLMs in a supervised learning way. Last, we map the generated outputs into sentiment labels during inference.

3.1 Instruction Tuning

We adopt the instruction tuning method of an LLM on financial sentiment analysis datasets. This process is divided into three main steps:

Formatting Financial Sentiment Analysis Dataset into Instruction Tuning Dataset

The existing financial sentiment analysis datasets are formatted as text classification task where the **inputs** are the financial news or headlines and the **outputs** are integer-type labels representing *positive*, *negative* and *neutral* sentiments. Our first step is to formulate these classification datasets into instruction-formatted dataset.

Following [Zhao *et al.*, 2023], we create 10 human-written **instructions** describing the task of financial sentiment analysis, and formulate each sample from the original dataset by combining one randomly selected **instruction** with the **input** and **output** in the format of ”Human: [**instruction**] + [**input**], Assistant: [**output**]”. This process is shown in Fig 1.

Instruction Tuning LLaMA-7B

While pretrained LLMs possess capabilities such as reasoning, understanding numbers, world knowledge, and multilingualism, they struggle to effectively apply these abilities to specific tasks. This limitation hinders their ability to achieve state-of-the-art (SOTA) performance on specific tasks, thus restricting their application potential. For instance, [Wei *et al.*, 2022] found that the zero-shot performance of LLMs is significantly lower compared to their few-shot performance. In our scenario, we leverage instruction data, which typically includes numeric values, financial context, and financial jargon, to provide supervised signals. Through instruction tuning, we align the LLM's capabilities with the sentiment analysis labels, achieving a more precise and nuanced understanding of sentiments expressed in financial texts which enables it to outperform both pretrained LLMs and supervised models specifically designed for financial sentiment analysis.

We illustrate our approach using instruction tuning with the LLM model called LLaMA-7B as an example to validate our ideas. Instruction tuning involves fine-tuning pre-trained LLMs by leveraging a collection of formatted instances in natural language [Wei *et al.*, 2022]. It is a method closely aligned with supervised fine-tuning. During the training process, we specifically employ the formatted instances to fine-

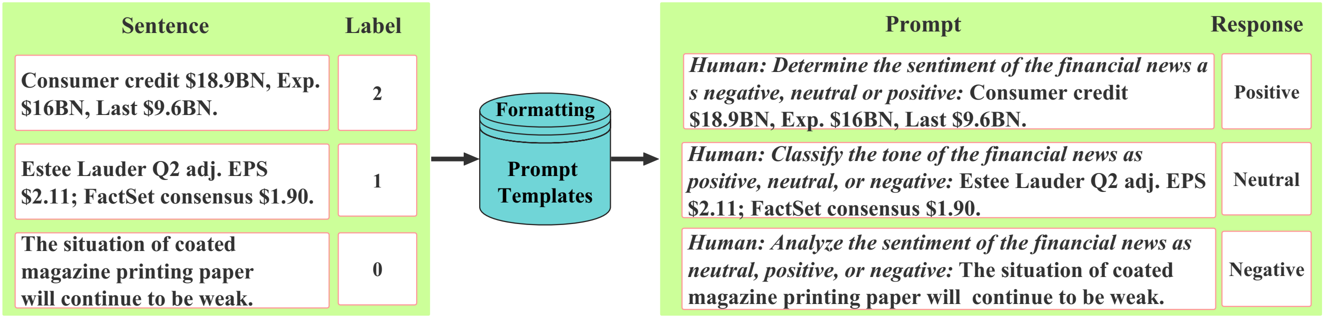
Sentence
Label
Consumer credit $18.9BN, Exp. $16BN, Last $9.6BN.
2
Estee Lauder Q2 adj. EPS $2.11; FactSet consensus $1.90.
1
The situation of coated magazine printing paper will continue to be weak.
0
Formatting
Prompt Templates
Prompt
Response
Human: Determine the sentiment of the financial news a s negative, neutral or positive: Consumer credit $18.9BN, Exp. $16BN, Last $9.6BN.
Positive
Human: Classify the tone of the financial news as positive, neutral, or negative: Estee Lauder Q2 adj. EPS $2.11; FactSet consensus $1.90.
Neutral
Human: Analyze the sentiment of the financial news as neutral, positive, or negative: The situation of coated magazine printing paper will continue to be weak.
Negative

Figure 1: Formatting sentiment analysis dataset into instruction tuning dataset.

tune the LLaMA-7B LLM using a supervised learning approach, i.e., training with a sequence-to-sequence loss. This choice allows us to showcase the effectiveness and applicability of instruction tuning in enhancing the financial sentiment analysis performance of LLMs like LLaMA-7B.

Mapping the Generated Outputs into Sentiment Labels Since the instruction finetuned LLaMA-7B is an autoregressive generative model, even though we train it using instruction templates to guide its output towards the desired sentiment judgments, it still has the possibility of generating free-style text. Therefore, we need to map the model's output back to the specified three emotions for proper evaluation. Our approach is as follows: if the model's output contains ”positive,” ”negative,” or ”neutral” terms, we map it to the corresponding label; otherwise, we consider it as ”neutral” sentiment.

3.2 Comparison Between LLMs and FinBERT for Sentiment Analysis

Our approach employs LLMs and compares their efficacy in sentiment analysis with the well-established FinBERT model. The comparison is based on three pivotal aspects:

- **Contextual understanding**: LLMs have an advantage due to their large-scale pretraining on diverse data. This provides them with a more comprehensive general knowledge, enabling a superior understanding of the context compared to FinBERT. The diversity and richness of the training datasets of LLMs are unmatched, providing them with a well-rounded knowledge that outshines FinBERT's capability.
- **Numerical sensitivity**: Financial texts often incorporate significant numerical data, which plays a crucial role in conveying the sentiment. LLMs, with their inherent numerical sensitivity, exhibit an enhanced capacity for interpreting the sentiment implied by numerical fluctuations. Refer to certain scholarly reports for in-depth studies on this characteristic of LLMs.
- **Decoder-only vs encoder-only models**: FinBERT is an encoder-only model which encodes the input sequence into a representation and relies on a separate classifier to make predictions based on the encoded representation. On the other hand, the employed LLM is a decoder-only model which can generate the entire output sequence, including the class label, directly from a latent representation or fixed-length vector. This character allows the LLMs easily adapt to various tasks without modifying the model structure while the encoder-only models require the development of task-specific classifiers, which can be more labor-intensive.

4 Performance Evaluation

In this section, we evaluate the effectiveness of our proposed method from three perspectives: general sentiment analysis, numerical understanding, and general knowledge supplementing. To validate our method's performance, we compare it against state-of-the-art sentiment analysis model, FinBERT, and the general-purpose LLM, ChatGPT.

Our experimental results validate the effectiveness of our approach. With only a small amount of fine-tuning data, our model consistently achieves superior performance in sentiment analysis compared to FinBERT and ChatGPT.

4.1 Datasets

Our training data is an amalgamation of the Twitter Financial News dataset [Magic, 2022] and FiQA dataset [Maia *et al.*, 2018], resulting in a comprehensive collection of 10, 501 samples.

Training Datasets

- **Twitter financial news sentiment training**: This dataset is a corpus of news tweets that pertain to the financial sector and is exclusively in English. Its primary purpose is the classification of financial sentiment within the context of Twitter discussions. The dataset comprises 9,540 samples for training, each annotated with one of three labels: Bearish, Bullish, or Neutral.
- **FiQA dataset**: This dataset, which is readily accessible via HuggingFace, includes 961 samples. Each sample has been annotated with one of three labels: positive, neutral, or negative, denoting the sentiment conveyed in the corresponding text.

Testing Datasets

- **Twitter financial news sentiment validation (Twitter Val)**: This dataset, accessible through Hugging-

Datasets			Models		
Name	**Size**	**Metrics**	**FinBERT**	**LLaMA-7B**	**Instruct-FinGPT-7B**
Twitter Val	2388	Acc	0.725	0.54	**0.880**
		F1	0.668	0.36	**0.841**
Testing Time			18 seconds (1 GPU)	498 seconds (8 GPUs)	498 seconds (8 GPUs)
Numerical	117	Acc	0.633	0.60	**0.837**
		F1	0.630	0.42	**0.795**
Contextual	20	Acc	0.50	0.55	**0.80**
		F1	0.22	0.34	**0.63**

Table 1: Experimental results on the Twitter financial news sentiment validation, numerical, and contextual datasets

Face, contains 2,390 samples annotated with three labels: Bearish, Bullish, or Neutral.

- **Numerical sensitivity dataset (numerical)**: This dataset, which we automatically filtered from Twitter Val, includes 117 samples. These samples contain at least two numerical values related to financial indicators without strong indication words such as 'raise', 'fall', 'increase', 'decrease'.
- **Contextual understanding dataset (contextual)**: This dataset, which we randomly selected from Twitter Val, includes 20 samples. These samples lack the essential contexts to make a sentiment prediciton.
- **Financial PhraseBank (FPB) dataset**: This dataset [Malo *et al.*, 2014] comprises 4,840 samples randomly extracted from financial news articles available on the LexisNexis database. The samples were carefully annotated by a team of 16 annotators with backgrounds in finance and business, ensuring high quality annotations.

4.2 Model Training

The training parameters are given in Table 2. For our Instruct-FinGPT-7B model, we initialize it with LLaMA-7B model and perform instruction tuning over 10 epochs. The training process utilizes the AdamW optimizer [Loshchilov and Hutter, 2017], with a batch size of 32, an initial learning rate of $1e^{-5}$, and a weight decay of 0.1. To maintain efficiency, we set a maximum input text length of 512 tokens. We utilize DeepSpeed [Rasley *et al.*, 2020] for the fine-tuning process on 8 A100 (40GB) GPUs, resulting in a total training time of 58 minutes.

Parameter	Value
Learning rate	1e-5
Weight Decay	0.1
Batch size	32
Training epochs	10
LR Scheduler	CosineAnnealing
Num warmup Steps	0
Max Token Length	512
GPUs	8 A100 (40GB)

Table 2: Training parameters.

4.3 Baseline Models

LLaMA-7B [Touvron *et al.*, 2023] We obtained the LLaMA-7B[1] model from Meta and use it for inference, keeping the same inference setting as our Instruct-FinGPT-7B.

FinBERT We obtained the FinBERT model from the Hugging Face Model Hub. The FinBERT model is used for sentiment analysis after pre-processing raw data, which includes tokenizing the text and padding or truncating it to fit the model's max input length. Once pre-processed, the data is run through FinBERT for inference, providing sentiment analysis results (positive, negative, or neutral) for each text input.

ChatGPT The utilization of OpenAI's ChatGPT API for sentiment analysis comprises a streamlined four-step process:

1. **API setup**: This involves setting up the OpenAI Python client, which serves as an interface to interact with the ChatGPT API.
2. **Data preparation**: The Instruction Tuning dataset as shown in Figs. 1 is employed for the inference with the ChatGPT model.
3. **API call**: Due to existing limitations, the GPT-3.5 API is used for requests. The GPT-4.0 version is currently unavailable for programmatic access and can only be interacted with via a web interface.
4. **Response interpretation**: The response from the API includes the sentiment of the text directly. This direct sentiment output simplifies the task of sentiment analysis.

4.4 Evaluation and Analysis

To evaluate the performance of our model, we test it on a benchmark financial sentiment analysis dataset and contrast the results with those of FinBERT. The key evaluation metrics center around the model's capability to manage numerical values and comprehend sentiment within various contexts.

Performance Metrics The primary performance metrics for our model include accuracy, and F1-score. Accuracy measures the proportion of correct predictions, and the F1-score represents the harmonic mean of precision and recall.

[1] We use LLaMA-7B for research and education purposes.

News	True Value	FinBERT	ChatGPT 3.5	ChatGPT 4.0	Instruct-FinGPT
Pre-tax loss totaled euro 0.3 million, compared to a loss of euro 2.2 million in the first quarter of 2005.	**Positive**	Negative	Negative	**Positive**	**Positive**
Madison Square Garden Q2 EPS $3.93 vs. $3.42.	**Positive**	Negative	**Positive**	**Positive**	**Positive**
Consumer credit $18.9BN, Exp. $16BN, Last $9.6BN.	**Positive**	Neutral	**Positive**	**Positive**	**Positive**
Estee Lauder Q2 adj. EPS $2.11; FactSet consensus $1.90.	**Neutral**	**Neutral**	Positive	Positive	**Neutral**

Table 3: Examples and results on the numerical sensitivity dataset.

News	True Value	FinBERT	ChatGPT-3.5	ChatGPT-4.0	Instruct-FinGPT
The situation of coated magazine printing paper will continue to be weak.	**Negative**	Neutral	**Negative**	**Negative**	**Negative**
Boeing announces additional order for 737 MAX planes.	**Neutral**	Positive	Positive	Positive	Positive
Boeing: Deliveries 24 Jets in November.	**Positive**	Neutral	**Positive**	Neutral	**Positive**
PPD's stock indicated in early going to open at $30, or 11% above $27 IPO price.	**Neutral**	Positive	Positive	Positive	**Neutral**

Table 4: Examples and results on the contextual understanding dataset.

Overall Performance Based on the evaluation results in Table 1, our instruction tuned LLaMA-7B (Instruct-FinGPT-7B) consistently outperforms both FinBERT and LLaMA-7B across all three datasets in terms of accuracy and F1 score. Especially, comparing our Instruct-FinGPT-7B with the original LLaMA-7B model (without instruction tuning), it is evident that the instruction tuning method significantly improves the model's performance on financial sentiment analysis.

Analysis of Numerical Sensitivity Numerical data plays a crucial role in financial sentiment analysis, as it often reflects important financial indicators. In Table 3, we assess the models' ability to comprehend and interpret sentiment associated with numbers.

- Example 1: This is an example from FinBERT, where FinBERT failed in this case. However, ChatGPT 4.0 and Instruct-FinGPT correctly recognize the substantial decrease in the loss from 2.2 million to 0.3 million, indicating a positive sentiment.
- Example 2: The increase in EPS is correctly identified as a positive sentiment by all models except FinBERT.
- Example 3: The exceeding of consumer credit expectations and the previous value is recognized as a positive sentiment by all models except FinBERT.
- Example 4: The statement about Estee Lauder and FactSet consensus is neutral, as it merely states the facts without indicating a positive or negative sentiment.

Our model demonstrates varying levels of effectiveness in understanding and interpreting sentiment associated with numerical data.

Analysis of Contextual Understanding The ability of our model to interpret sentiment in different contexts is an important aspect of its performance evaluation. Financial news can be nuanced, and a statement that may appear negative in one context could be neutral or even positive in another. We assess the models' performance in contextual understanding based on the examples provided in Table 4.

- Example 1: This is an example from FinBERT, where FinBERT failed in this case. But ChatGPT and Instruct-FinGPT recognized that the situation of coated magazine printing paper is expected to remain weak, indicating a negative outlook for the industry. LLMs' language understanding capabilities and knowledge of financial contexts enable them to accurately interpret such statements and predict the sentiment.
- Example 2: In this specific case, indicating that Boeing has received more orders for their aircraft. It looks like positive news. However, without further context, it's challenging to determine the sentiment accurately. The lack of specific details about the order, the customer, or any potential implications can make it difficult to assess the sentiment correctly.
- Example 3: The sentiment of the financial news is positive. The statement highlights that Boeing delivered 24 jets in November, indicating a successful and productive month for the company.
- Example 4: All of the models failed on this one. The opening price of a stock is higher than the IPO price doesn't necessarily indicate the stock is rising from its current market price.

Performance	ChatGPT 3.5	LLaMA-7B	Ours-7B
FPB (ACC)	0.64	0.60	**0.76**
FPB (F1)	0.51	0.40	**0.74**

Table 5: Zero-shot evaluation between ChatGPT and Instruct-FinGPT on the entire dataset of financial phaseBank.

Overall, our model demonstrates a better understanding of the contextual sentiment in these examples compared to FinBERT and ChatGPT. It successfully recognizes the negative sentiment in Example 1 and accurately identifies the neutral sentiment in Example 2 and the positive sentiment in Example 3. These results highlight the importance of contextual understanding in financial sentiment analysis and the variations in performance across different models.

Zero-Shot Generalization to Other Financial Datasets Finally, we evaluate the zero-shot ability of our model, which refers to how well the model can generalize to other unseen financial datasets. A model with strong zero-shot capabilities can provide more robust and versatile results in real-world applications. We compare our Instruct-FinGPT with ChatGPT3.5 and LLaMA-7B on the full FPB dataset. Here we do not compare with FinBERT because it uses FPB as the training set.

The evaluation results are shown in Table 5. Based on these results, it can be concluded that the instruction tuned LLaMA-7B model performs the best among the three, achieving the highest accuracy and F1 score. The fine-tuning process with sentiment instruction data seems to have improved the model's ability to capture sentiment in financial phrases, resulting in better zero-shot performance compared to both ChatGPT and the original LLaMA-7B model.

5 Conclusion and Future Work

In this paper, we have presented an innovative approach for financial sentiment analysis by harnessing the general knowledge and reasoning capabilities of LLMs. Our method represents a substantial contribution to the field of sentiment analysis, demonstrating that instruction tuning of an LLM can yield superior performance with a small amount of task-specific data. Our findings pave the way for future research into the potential of LLMs for a broad range of financial tasks.

Disclaimer: We are sharing codes for academic purposes under the MIT education license. Nothing herein is financial advice, and NOT a recommendation to trade real money. Please use common sense and always first consult a professional before trading or investing.

References

[Araci, 2019] Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. In *arXiv preprint arXiv:1908.10063*, 2019.

[Atkins *et al.*, 2018] Adam Atkins, Mahesan Niranjan, and Enrico Gerding. Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2):120–137, 2018.

[Chan and Chong, 2017] Samuel WK Chan and Mickey WC Chong. Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64, 2017.

[Day and Lee, 2016] 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.

[Hamilton *et al.*, 2016] William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 2016, page 595. NIH Public Access, 2016.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 2017.

[Loughran and McDonald, 2011] Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

[Magic, 2022] Neural Magic. Twitter financial news sentiment. http://precog.iiitd.edu.in/people/anupama, 2022.

[Maia *et al.*, 2018] Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra. Balahur. Www '18: Companion proceedings of the the web conference 2018. In *International World Wide Web Conferences Steering Committee*, Republic and Canton of Geneva, CHE, 2018.

[Malo *et al.*, 2014] Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

[Mishev *et al.*, 2020] 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.

[Rasley *et al.*, 2020] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Association for Computing Machinery*, KDD '20, page 3505–3506, New York, NY, USA, 2020.

[Shah *et al.*, 2022] Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of* *the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2022.

[Sohangir *et al.*, 2018] 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.

[Tai and Kao, 2013] Yen-Jen Tai and Hung-Yu Kao. Automatic domain-specific sentiment lexicon generation with label propagation. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, pages 53–62, 2013.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wei *et al.*, 2022] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[Wu *et al.*, 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.

[Xing *et al.*, 2018] Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023.