What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis

Xiang Deng*
The Ohio State University
Columbus, USA
deng.595@osu.edu

Vasilisa Bashlovkina
Google Research
NYC, USA
vasilisa@google.com

Feng Han
Google Research
NYC, USA
bladehan@google.com

Simon Baumgartner
Google Research
NYC, USA
simonba@google.com

Michael Bendersky
Google Research
Mountain View, USA
bemike@google.com

ABSTRACT

Market sentiment analysis on social media content requires knowledge of both financial markets and social media jargon, which makes it a challenging task for human raters. The resulting lack of high-quality labeled data stands in the way of conventional supervised learning methods. Instead, we approach this problem using semi-supervised learning with a large language model (LLM). Our pipeline generates weak financial sentiment labels for Reddit posts with an LLM and then uses that data to train a small model that can be served in production. We find that prompting the LLM to produce Chain-of-Thought summaries and forcing it through several reasoning paths helps generate more stable and accurate labels, while using a regression loss further improves distillation quality. With only a handful of prompts, the final model performs on par with existing supervised models. Though production applications of our model are limited by ethical considerations, the model's competitive performance points to the great potential of using LLMs for tasks that otherwise require skill-intensive annotation.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

Sentiment Analysis, Social Media, Finance, Large Language Model

ACM Reference Format:
Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. 2023. What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587324

1 INTRODUCTION

Social media platforms such as Reddit and Twitter contain insights about financial markets, for example in the form of posts that express financial expectations for a particular company. We define the financial sentiment of a post about a company as positive (bullish) if the author of the post has a favorable outlook for the company, negative (bearish) if their outlook for the company is negative, and neutral otherwise. Financial (market) sentiment analysis aims to automatically extract the financial performance expectations conveyed in the text.

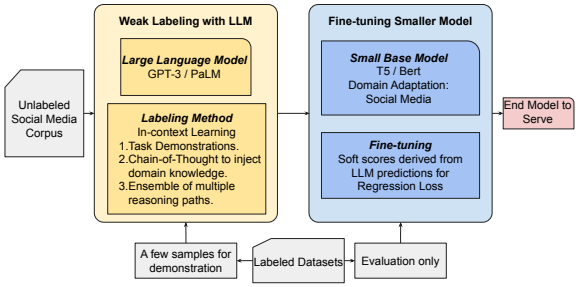

Figure 1: Our overall pipeline to bootstrap a market sentiment model with a LLM.

One particular challenge for market sentiment analysis on social media is the lack of high-quality labeled data, which arises because the annotation requires both finance domain knowledge and an understanding of social media jargon. Previous study has found that "bearish" or "bullish" tags selected by the authors of the posts themselves are often not accurate, and even finance experts may hold different opinions during annotation [5]. Our in-house annotation effort revealed the same issue, with raters only agreeing with each other around 70% of the time.

In the meantime, large language models (LLMs) such as GPT-3 [3] and PaLM [6] gained popularity in recent years for their impressive aptitude for in-context learning [3, 15]. An LLM can perform textual tasks with just a few examples demonstrating what needs to be done, often achieving results similar to those of state-of-the-art supervised models in a wide range of applications.

Inspired by this development, we investigate the use of LLMs to bootstrap a market sentiment analysis model for social media content with minimal human annotation efforts. We select Reddit as the target social media platform, which, unlike other platforms used in existing works (Twitter and Stocktwits), has a broader coverage of topics, ranging from market events to analysis and user investing actions, with both short user comments and long "due diligence" posts. To annotate Reddit posts with weak financial sentiment labels, we use LLM in-context learning [3, 6] with Chain-of-Thought [16] reasoning and repeated generation [14] for more stable predictions. Since the LLM is too large and slow to be used in a production setting, we distill it into a smaller student model. We find that if we aggregate multiple predictions for a single example into a soft score and use a regression loss, we can make use of more data and get a

*Work done while interning at Google.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04.
https://doi.org/10.1145/3543873.3587324

smoother precision-recall curve. Even though we are able to serve our student model online and control its precision by setting the prediction threshold, the model's application is limited by ethical considerations stemming from the high-stakes nature of investment decisions. Nevertheless, compared with models fine-tuned on existing market sentiment datasets, our model trained with only weakly-labeled Reddit data not only improves on the challenging in-domain testing data from Reddit, but also generalizes well across datasets and performs on par with the supervised counterparts.

2 RELATED WORK

Market Sentiment Analysis. Existing work roughly falls into two categories: lexicon-based methods that associate individual words with sentiment labels [4, 10], and machine learning methods that train a supervised model with labeled data [1, 17]. However, they either show unsatisfactory performance or demand huge amounts of labeled data that is hard to acquire in practice. In this work, we conduct an exploratory study of leveraging the in-context learning ability of LLMs to overcome the data challenge.

In-context Learning with LLM. Large language models have demonstrated the capability to perform tasks by making predictions conditioned on a few input-output examples without updating any model parameters [3]. Many recent works have studied the underlying mechanism of in-context learning [12, 15], and how to improve the few-shot performance [14, 16, 18]. We adapt some of these techniques for market sentiment analysis and design a pipeline that puts them into practice.

3 METHODOLOGY

3.1 In-context Learning with LLM

Though our target task is analyzing the author's overall financial outlook for a company, in social media, users normally discuss financial performance in terms of stock price movement. We use the domain-adapted proxy task of extracting stock price movement expectations in our LLM prompt. As shown in Figure 2, it is composed of ① **Task description**, which simulates the task setting and familiarizes the LLM with the target domain, and ② **Demonstrations**, which illustrate the task via multiple input-output examples (③, ⑤). The output ⑤ is verbalized to make it similar to the examples seen by the model during pre-training and then converted to actual categorical labels during post-processing.

Our preliminary study shows that while this prompt design can already yield reasonable results, the prediction is unstable and sensitive to the exact wording of the prompt. In particular, we notice that the results vary a lot if we simply shuffle the order of demonstrations, which means that the model struggles to truly understand the user's opinion [12]. To counter this instability, we incorporate **Chain-of-Thought** (COT) [16] reasoning into our setting. COT was originally designed to improve the multi-step reasoning ability of LLMs by explicitly instructing the model to generate intermediate reasoning steps. While we do not need multi-step reasoning for market sentiment analysis, we use COT to make the LLM summarize the author's finance-related arguments TL;DR-style (④), thus implicitly forcing it to recall relevant financial domain knowledge before drawing a conclusion (⑤). Because users often cite multiple,

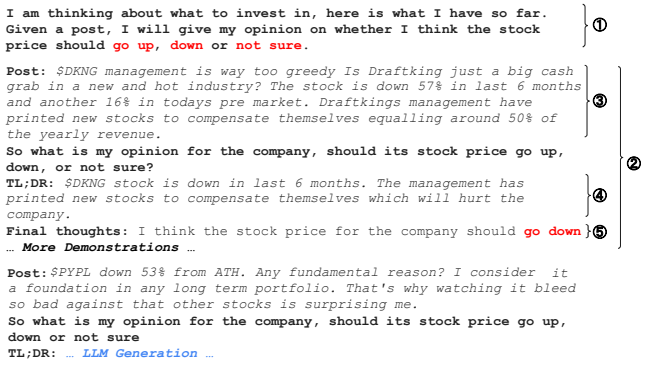

Figure 2: Prompt template for in-context learning.

sometimes conflicting arguments in their posts, we use temperature sampling [7] instead of greedy decoding during generation and repeat it multiple times to produce varying reasoning paths, giving the model a chance to focus on different lines of argument. As a result, each example gets assigned multiple, potentially inconsistent labels [14]. We use majority voting to get the final prediction for in-context learning, and describe how to better leverage the multiple predictions for distillation in Section 3.2.

3.2 Bootstrapping a Market Sentiment Model with LLM

While in-context learning with LLMs has shown impressive results during offline evaluation [3], it is impractical to serve such large models in production. A commonly used compression method is to first generate a large weakly-labeled dataset using the larger teacher model and then train a smaller student model in a supervised fashion [8]. We do notice that in some cases there are complex or ambiguous posts where the LLM assigns the weak label incorrectly. It is also the case that for those hard examples, the LLM makes inconsistent predictions when exploring different reasoning paths. A straightforward way to leverage this pattern would be to filter out weakly-labeled examples that have any inconsistency among the labels assigned via different LLM reasoning paths. However, such filtering may cost us many potentially useful examples and cause the student model to overfit to the remaining easy cases. Instead, we view the agreement ratio between the multiple labels of a single example as a soft score of sentiment polarity and train the student model to predict this score with a regression loss.

4 EXPERIMENTAL RESULTS

4.1 Experimental Setup

Problem Formulation. We study market sentiment analysis as a three-way classification task. The market sentiment of a post about a particular company is defined as positive (bullish) if the author's outlook for it is favorable, negative (bearish) if their outlook is negative, and neutral otherwise.

Datasets.

For both distillation and evaluation, we use Reddit posts labeled as finance-related by a proprietary topic classifier. We filter posts based on the popularity of the mentioned stock in an internal system and randomly sample 20,000 posts for distillation. Since there are

Table 1: Data Statistics.

	FiQA News	FiQA Post	Reddit Testing
# Total	370	674	100
% Neg / Neu / Pos	34/-/66	35/-/65	39/42/19
Avg. Length	9.7	13.4	83.0

no existing datasets for market sentiment analysis on Reddit, we sample another 100 posts for evaluation, which are annotated by three in-house experts who have both knowledge of investing terms and experience with Reddit.

We also experiment with the widely used FiQA benchmark [11], which contains two subtasks: FiQA-News with news headlines, and FiQA-Post with microblogs from Twitter and Stocktwits. We convert it to a binary classification task with the original sentiment scores.[1] Since the original testing set is private, we split the original training set into training, validation, and testing following an 80/10/10 ratio.

Baselines. Our backbone model is Charformer [13] (CF). We further pre-train CF on social media content. We consider the following baselines: (i) our backbone model fine-tuned on FiQA, (ii) PaLM in-context learning with COT and majority-vote aggregation over 8 reasoning paths, and (iii) two widely used existing market sentiment models: FinBERT-HKUST [17] and FinBERT-ProsusAI [1][2].

Implementation Details. We use PaLM-540B [6] as the LLM for in-context learning and weak labeling. We randomly select 6 examples as demonstrations and remove them from the test set when evaluating PaLM. For Reddit, we select two examples for each sentiment category and manually write the COT reasoning. For FiQA, we select three examples for each category, and use the "Aspect Snippet" in the original dataset as COT reasoning. For each input, we run the generation 8 times with a temperature of 0.5 to produce different reasoning paths and predictions. For distillation, we keep weakly-labeled examples for which the LLM makes 5 or more consistent predictions, and fine-tune the CF model on 17K Reddit posts labeled with soft scores aggregated from the 8 PaLM predictions. We use a learning rate of 1e-4 and a batch size of 64. We apply a regression head to the final CF encoder layer and drop the decoder. The final CF model in Table 2 has 102M parameters while both FinBERT models have 110M parameters. Ablations on the number reasoning paths, filtering, and fine-tuning objectives can be found in Figure 3, 4 and Table 3.

4.2 Results

Overall. Table 2 summarizes the main results. First, we can see that the Reddit dataset is more challenging than the FiQA datasets, likely due to longer posts and multifaceted user opinions. The PaLM model performs very well on all datasets with only 6 demonstration examples. Our student model fine-tuned on weakly labeled Reddit data is able to effectively transfer the knowledge from the LLM and outperform all supervised baselines on the Reddit dataset. At the same time, our model generalizes well to the FiQA dataset despite only being fine-tuned on Reddit posts. In summary, the experiments show promising results of leveraging LLMs for market sentiment analysis. With only a handful of labeled examples for demonstration, we can bootstrap a small student model that performs on par with or better than existing state-of-the-art models of servable size.

[1]Examples with sentiment scores greater than 0 are considered positive, and negative otherwise. We remove the few examples with a sentiment score of 0 and those mentioning multiple stocks. This drops 66 examples for News and 1 example for Post.
[2]We use the models released on Huggingface ModelHub: https://huggingface.co/ProsusAI/finbert, https://huggingface.co/yiyanghkust/finbert-tone

Table 2: Accuracy on benchmark datasets.

	FiQA News	FiQA Post	Reddit
CF - FiQA News	75.7	69.1	42.0
CF - FiQA Post	86.5	85.3	40.0
FinBERT-ProsusAI [1]	81.1	73.5	48.0
FinBERT-HKUST [17]	75.7	67.6	50.0
PaLM *COT* × 8	97.3	95.6	72.0
CF - Distilled PaLM	83.8	77.9	69.0

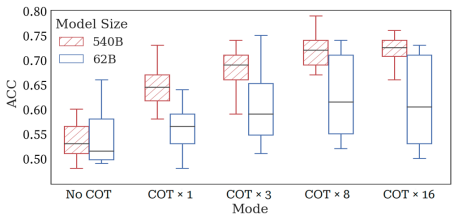

Figure 3: Ablation for In-context Learning. Here we compare LLMs of different sizes, effect of COT, and the use of multiple reasoning paths. *COT* × *N* stands for generating *N* reasoning paths and doing majority voting to aggregate the predictions.

Ablation on In-context Learning. Figure 3 demonstrates the importance of using chain-of-thought (COT) reasoning and repeating the generation for in-context learning. Here we use the same demonstration examples but shuffle their order to get different prompts. First, we see that in-context learning is sensitive to the prompt design: even with only the order of examples changed, the final performance varies a lot. Second, using COT to have PaLM summarize the post's main arguments as a TL;DR greatly improves the performance. Asking the LLM to generate multiple reasoning paths and aggregating the predictions further boosts the performance, as it allows the model to explore different aspects of the user's opinion. Finally, model size influences the effectiveness of COT reasoning. While the 62B and 540B PaLM models have similar performance with the base prompt, the 540B model benefits much more from COT, likely because its superior generational ability allows it to produce more useful intermediate thinking steps.

Ablation on Distillation Methods. We compare filtering the PaLM-labeled data with different intra-label agreement thresholds. As we can see from Table 3, exposing the student model to examples with inconsistent labels hurts its performance even though it gets to see more training data that way. We don't include a full ablation on the student model backbone but we experiment with its loss function. Figure 4 shows that using a regression loss instead of classification is advantageous for two reasons: it better leverages the soft scores from the examples with inconsistent labels and it produces a slightly smoother precision-recall curve. The latter is important for production applications because the smoother curve allows us to pick an operating point with the desired precision.

Error Analysis. We conduct error analysis over the Reddit testing set for our final model (CF - Distilled PaLM). The majority of errors are between neutral and the other two labels (29 out of 31), which is less severe than positive/negative errors. We notice that the model struggles when the input contains contradictory

Table 3: Average precision for Positive and Negative labels. Here we compare using classification (CLS) and regression (RGR) loss at different intra-label agreement thresholds.

Agreement		8	7	6	5
# Examples		6,240	10,474	14,152	17,456
Pos	CLS	80.5	75.8	71.4	76.9
	RGR	74.2	78.5	81.7	84.2
Neg	CLS	68.0	64.0	47.4	57.9
	RGR	54.3	61.8	61.7	65.5

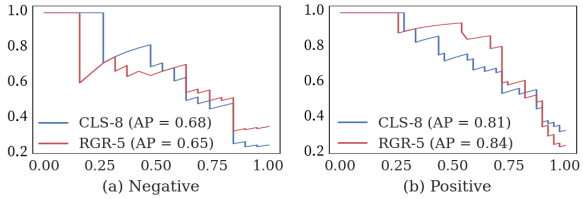

Figure 4: Precision-Recall curve. Here we compare models that achieve the best average precision (as shown in Table 3) using Classification (CLS-8) and Regression (RGR-5) loss.

arguments or discusses advanced investing actions. Better handling such complicated posts and dynamically incorporating relevant finance domain knowledge could be a subject of future work.

5 ETHICAL CONSIDERATIONS

Applying our model to social media content can make its wealth of financial information more accessible to users. For example, bullish/bearish tags for individual posts can help novice investors orient themselves in the language of r/wallstreetbets. However, the model's output should not be used to make investment decisions due to the associated risks. First, the model predicts the wrong sentiment more than 30% of the time. Second, even if the model doesn't make a mistake, the social media posts it is applied to may convey sentiment that prompts the user to make bad financial decisions. Prior research has found that investors are susceptible to social media advice [9] even though the sentiment it carries is a poor predictor of stock prices [2]. Finally, aggregating financial sentiment from social media may amplify malicious behavior like market manipulation. In fact, our model detected a negative sentiment spike for Pfizer in March 2021 when there seemed to be a coordinated effort to promote a rumor that Pfizer shares were getting delisted from NYSE[3]. These risks need to be thoroughly addressed and mitigated to ensure that the likely benefits from deploying our model substantially outweigh the foreseeable downsides.

6 CONCLUSION

We study financial sentiment analysis on Reddit with an LLM distilled into a production-friendly student model. With minimal human-annotated data, our classifier performs on par with existing supervised models and generalizes well across other datasets. The application of our model does pose a product challenge: how can we incorporate the model's output responsibly, delivering value to users without misleading them or inadvertently amplifying malicious behavior? Nevertheless, our investigation highlights the promise of in-context learning with LLMs for tasks that are hard for human raters to annotate. Can human raters, instead of simply labeling the data, help design domain-knowledge-injected prompts teaching the LLM to perform the task, or otherwise "collaborate" with the LLM? How can automatic prompt-tuning further optimize the human-engineered prompt? Exploring the answers to these questions would be a compelling direction for future work.

[3]https://factcheck.afp.com/doc.afp.com.328D4BT

REFERENCES

[1] Dogu Araci. 2019. *Finbert: Financial sentiment analysis with pre-trained language models.* arXiv:1908.10063

[2] Daniel Bradley, Jan Hanousek, Russell Jame, and Zicheng Xiao. 2021. *Place your bets? The market consequences of investment advice on Redditâ€™s Wallstreetbets.* MENDELU Working Papers in Business and Economics 2021-76.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS 2020.*

[4] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. NTUSD-Fin: a market sentiment dictionary for financial social media data applications. In *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2018).* 37–43.

[5] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and Perspectives from 10,000 Annotated Financial Social Media Data. In *LREC.* 6106–6110.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. *PaLM: Scaling Language Modeling with Pathways.* arXiv:2204.02311

[7] Jessica Ficler and Yoav Goldberg. 2017. Controlling Linguistic Style Aspects in Neural Language Generation. In *Proceedings of the Workshop on Stylistic Variation.*

[8] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *IJCV* 129, 6 (2021), 1789–1819.

[9] Kathryn Kadous, Molly Mercer, and Yuepin Zhou. 2017. *Undue Influence? The Effect of Social Media Advice on Investment Decisions.* WorkingPaper.

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

[11] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. *Companion Proceedings of the The Web Conference 2018* (2018).

[12] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *EMNLP 2022.*

[13] Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast Character Transformers via Gradient-based Subword Tokenization. In *ICLR.*

[14] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. *Self-Consistency Improves Chain of Thought Reasoning in Language Models.* arXiv:abs/2203.11171

[15] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. *Emergent Abilities of Large Language Models.* arXiv:2206.07682

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS 2022.*

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

[18] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models.* arXiv:2205.10625