Research on Application of Financial Large Language Models

Mingting Du *

School of mathematics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China

* Corresponding Author Email: 42234005@smail.swufe.edu.cn

Abstract. With the increasing use of large language models such as chatgpt, it is not difficult to apply their capabilities to the research of natural language processing in the financial field, including but not limited to text extraction, sentiment analysis, etc. This paper analyzes the construction ideas and applications of three financial big language models, including BloombergGPT, PIXIU and FinBERT, and concludes that the current application of big language models in the financial field is possible, multi-faceted and suitable, but there are still shortcomings in ethics, data processing and other aspects. The application of large language models in the field of finance is still something to look forward to. Through this study and the comparative exploration of various models, we hope to provide valuable modeling experience for practitioners in the field of finance or computer. At the same time, it is hoped that each researcher can follow the ideas of these model-making teams to make up for the shortcomings in their own models and make their own financial big language models better.

Keywords: Large Language Models, BloombergGPT, PIXIU, FinBERT, natural language processing.

1. Introduction

GPT-3, released in 2020, demonstrates the power of in-context learning and is a significant milestone for LLMs, with large language models still under development gaining great attention in recent years [1]. Practitioners in the financial field began to think about whether big language models could be applied in the financial field, such as stock trend prediction, corporate financial report interpretation and other scenarios that require a lot of data processing power.

This paper selects three mature financial big language models BloombergGPT, PIXIU and FinBERT, and analyses their model-building ideas, application scenarios, experimental results and shortcomings respectively, so as to draw conclusions on the application research of big language models in the financial field based on their similarities.

2. Brief introduction of large language models

To predict the probability distribution of lexical sequences, a statistical model known as a language model is trainable on a large corpus database. The original model is called the n-gram model, which represents the sequence of terms as Markov processes, assuming that the probability of the next term depends only on the previous term [2]. Subsequently, language models based on Recurrent Neural networks (RNN) such as LSTM and GRU were born.

Then, in 2017, the Transformer architecture was unveiled, which sparked the development of several more language models that, for instance, perform better at translation tasks than recurrent neural networks. For effective training on big data sets, the Transformer architecture models word associations using a "self-attention" method. Currently, a famous Transformer-based models is Generative Pre-trained Transformer Model, which is also called GPT, created by OpenAI. It is a sizable language model that only has encoders and uses location embedding and a self-focusing technique to get dependencies in text [3]. With a deep bidirectional architecture and the ability to learn context representations, Bidirectional Encoder Representations from Transformers, known as BERT, is a decoder-only framework [4]. FinBERT is a big language model that focuses on finance and is based on BERT. Its sentiment analysis capability is improved by pre-training BERT financial

texts. Furthermore, BLOOM was developed using the Tansformer architecture and a vast array of parameters that were trained on around 366 billion (1.6TB) tokens [5]. The open-source BLOOM model supports several languages, and its financial application version is known as BLOOMGPT.

3. Large language models in finance

3.1. BloombergGPT

Many financial tasks are supported by the 50 billion parameter language model of BloombergGPT [6]. In order to maintain competitive performance on conventional LLM benchmarks while achieving best-in-class results on financial benchmarks, the BloombergGPT team developed a model. The group achieved this by utilizing Bloomberg's current capabilities for data development, gathering, and maintenance to create the largest domain-specific dataset to date.

3.1.1. Data construction of FinPILE

FinPILE is an extensive collection of financial papers in the English language that were taken from the Bloomberg archive, including news, documents, press releases, financial documents that were crawled from the internet, and social media [6].

BloombergGPT's training expertise on a large, clean, well-collated domain-specific dataset may offer valuable insights into developing large language models for finance, even if the team has said that FinPILE would not be made public.

For almost four decades, the bloomberg terminal has given a broad and diversified set of organized and unstructured financial data and analytics. Using this extensively gathered and updated dataset, Bloomberg experts assembled a collection of financial papers to develop FinPile, a collection of corporate records, financial news, and other information pertaining to the financial markets. FinPILE contains web data made out of The majority of FinPile is made up of online material that BloombergGPT finds and gathers, which includes hundreds of transcripts of Bloomberg TV news; news data, which is a collection of news items authored by Bloomberg journalists; financial statements, largely consisting of the files in this dataset originate from EDGAR, the SEC's web database; news, which generally consists press releases made publicly by financially relevant firms. Press releases, as opposed to filings, are written in a manner like to news stories. Bloomberg is the primary source for news as well as other papers the company has produced, including opinion and analysis. "Bloomberg News" and "Bloomberg First Word" are the primary sources.

The PILE, the data set used in GPT-neo, GPT-J, and GPT-neox (20B), has been used to successfully train LLMs and includes training that diversifies data and may even support financial data; C4, known as Colossal Clean Crawled Corpus, a frequently used data for train LLMs, was one of the three well-known and publicly available datasets used in the model. The techniques used for cleaning and processing C4 data are varied; Wikipedia, particularly the most recent entries, may enhance the model's legitimacy.

3.1.2. Model construction

The BloombergGPT, based on BLOOM, is a decoder-only causal language model. The size of this model is based on Chinchilla scaling laws [7]. Due to the limited data, it is necessary to select a model as large as possible, and the parameter of the final model is 50B [6]. A total of 139,200 steps (about 53 days) were trained, with 569B tokens of 709B tokens available in training [6].

3.1.3. Evaluation

For both financial and general reasons, the effectiveness of BloombergGPT was assessed across two major job areas. The notion that training on high quality financial particular data would result in greater performance on the financial problem is tested in part by the Financial particular task. The General Task looks at whether the model's output may be directly compared to results that have already been published.

Compared with broader NLP tasks, NLP tasks in finance possess unique qualities and difficulties while working with financial data.

External Financial tasks include Any news that might be favorable or bad for investors is classified as positive or negative in the Financial Phrasebank Dataset (FPB), which contains the sentiment classification job of financial news phrases; otherwise, it is classified as neutral; FiQA SA is based on the 2018 Financial Q&A and Opinion Mining Challenge and is used to anticipate the sentiment of particular features in English financial news and Weibo [8]; Determines if a headline in the gold commodities sector has specific information using the binary classification task Headline; Credit risk assessments for the named entity recognition (NER) task on financial data are derived from financial agreements filed with the SEC; ConvFinQA was tasked with responding to conversational questions requiring numerical reasoning based on S&P 500 earnings report inputs, which included text and more than one table containing financial data.

According to the final trial results, BloombergGPT comes in second in NER and performs better than all other models for four of the five tasks, which include ConvFinQA, FiQA SA, FPB, and Headline [6].

One of the internal financial tasks is Equity News Sentiment, which assesses how readers will feel about a company based on news stories; a "positive," "negative," or "neutral" note indicates that readers' long-term belief in the company may rise, fall, or remain unchanged; Similar to "Equity News Sentiment," but using English-language social media material pertaining to finance instead of news, is Equity Social Media Sentiment, as reported by BloombergGPT; Similar to "Equity News Sentiment," however instead of news, the team used a transcript from a press conference held by the firm; Country News Sentiment, which differs from different sentiment tasks in that its aim is to forecast the emotions conveyed for a country in articles, and the dataset includes Bloomberg's news stories in English, premium news, and web content [6]. ES News Sentiment, which is to predict sentiment on specific aspects of a company expressed in news reports; the aim is to aviod indicating the impact on investor’s belief.

This section's baseline definition is the goal for which the model can produce the right answer given the information in the rendered input text. The data sets include BoolQ, OpenBookQA, RACE, and some other data sets.

The final experiment result is that except for OpenBookQA, the performance of BloombergGPT is the highest among BLOOM176B, GPT-NeoX, and OPT66B [6].

A situation that isn't immediately related to a user-facing program is called a linguistic job. These responsibilities include assessing implication, syntax, and disambiguation. The purpose of these challenges is to evaluate the model's language comprehension directly.

The data sets include Recognizing Textual Entailment, Adversarial NLI, Commitment Bank, Choice of Plausible Alternatives, Words in Context, Winograd, Winograd, HellaSWAG, Story Cloze.

The final experiment result is that BloombergGPT falls slightly behind GPT-3 and outperforms the other models [6].

3.1.4. Advantages of BloombergGPT

In the comparison, BloombergGPT fared better than the other models with the approximate size of parameters. Furthermore, it can match or even outperform bigger models under some situations. Even yet, the model improved its capacity to handle generic data, beating models of a comparable size and, occasionally, equal or surpassing bigger models.

3.2. PIXIU

PIXIU is a thorough framework that includes FinMA, the first financial LLM, which is built on optimizing LLMa using data from multimodal and multitasking teaching. The researchers developed distinct task-specific instructions written by domain experts for each task using publicly available training data from a variety of tasks, including financial sentiment analysis, news headline classification, named entity recognition, question and answer sets, and stock movement predictions.

3.2.1. Raw Data Processing

Create a financial instruction tuning dataset (FIT) for a range of financial NLP and forecasting applications using open-source data.

Analyzing the emotional data entered into financial texts is the goal of the long-standing financial sentiment analysis endeavor. The Financial Phrase Bank (FPB) dataset and FiQA-SA are the two datasets used in this model [8].

The purpose of the news headline categorization task is to assess various types of information, including changes in price in financial language. The algorithm makes use of a dataset of golden news headlines that spans the years 2000 to 2019 and includes news stories about "gold" together with the nine associated tags (Resembling the external financial task's headline part in Bloomberg GPT). Sorting each label in each data sample into a binary class is the task.

Named Entity Recognition (NER) tasks identify important financial entities, such as individuals, businesses, and locations, that may be utilized to create a financial knowledge map. The model leverages the FIN dataset, which comprises manually annotated entity types from LOCATION, organization, and PERSON.

The process of automatically providing financial answers to inquiries based on the data supplied is known as question answering. FinQA and Con-vFinQA are the two datasets that the researchers used.

Predicting the movement of stocks is a fundamental financial job that has significant practical significance in areas like investment strategy. The price change is allocated to the positive sample “1” if it exceeds 0.55% and to the negative sample “-1” if it falls below -0.5%. Three widely used data sets—BigData22, ACL18, and CIKM18—are employed in the model.

3.2.2. Instruction Construction

The model constructs its instruction tuning samples using the following template for the FPB, FiQA-SA, Headline, NER, BigData22, ACL18, and CIKM18 datasets: "[task prompt] Text: [input text] Response: [output]” [9]. The prompt for each data is denoted by "[task prompt]", the input financial data from each data is denoted by "[input text]", and the corresponding output for each data is indicated by "[output]", such as the label of the input text from “Netural”, “Positive” and “Negative”.

The following template is built by this model for FinQA and ConvFinQA. The task prompt is nearly the same but includes a part of “Context Input”. For every data sample, the input contextual information is "[input context]". For instance, the text in the FinQA files can be used to populate the input context.

3.2.3. FinMA – Financial large language model

The researchers used FIT fine-tuning of LLaMA to create FinMA, and they fine-tuned LLaMA 7B and 30B checkpoints to acquire instruction tuning data pertaining to NLP tasks to create FINMA-7B and FINMA-30B. By fine-tuning LLaMA 7B to get comprehensive instruction tuning data, FinMA-7B-full is constructed [9].

3.2.4. Experiments and Results

In this section, Investigators compare FIT-fine-tuned FinMA with BloombergGPT, GPT-4, ChatGPT, BLOOM, GPT-NeoX, OPT-66B, Vicuna-13B [10-13].

On the FPB, FiQA-SA, and Headline datasets, the experiment's end result shows that the fine-tuning model FinMA performs noticeably better than other LLMS, highlighting the significance of domain-specific instruction tweaking for enhancing domain-specific large language model performance [9].

3.3. FinBERT

Owing to the specific terminology employed in the financial domain, generic models are not effective enough. The researchers introduced FinBERT, a Bert-based language model, to handle NLP tasks in the financial field.

3.3.1. Preliminaries

The following techniques were employed in the study to get the model ready for training: long short-term memory, a kind of recurrent neural network that uses "forget" and "update" gates to let long-term dependencies in the sequence remain in the network; ELMo embedding, which is contextualized word representation because surrounding words affect word representation [14]; ULMFit, a downstream NLP task transfer learning model pre-trained with language models [15]; Transformer is an attention-based sequence information model architecture. BERT, essentially a language model, consists of a set of Transformer encoders stacked on top of each other.

3.3.2. Data sets

1.8 million news articles published by Reuters between 2008 and 2010 make up TRC2-financial; 4845 English sentences selected at random from financial news found on the LexisNexis database [16] make up Financial PhraseBank; and FiQA Sentiment, a dataset created for conference financial opinion mining and question answering challenge, are among the data sets used in the experiment. 1,174 financial news items and tweets with related sentiment scores are among the data used by the researchers [17].

3.3.3. Conclusions

Three indicators were employed in the assessment process: the macro F1 average, the cross entropy loss, and the accuracy. FinBERT outperforms the models published in other studies (LPS, HSC, FinSSLX) as well as the approaches we constructed ourselves (LSTM and ULMFit) for every tested metric [17].

4. Challenges and Limitations

4.1. Ethic Level

Finance is a delicate area of technology, and guaranteeing accurate, reliable information is crucial for goods, consumers, and a company's reputation in the marketplace. Similarly, as makers of models, they take extra care with whatever they produce, whether from humans or machines. You need to keep an eye on the model's tendency to generate inappropriate content, and make it cleaner and contain less obviously biased or offensive content output.

4.2. Data Level

Handle financial data with high dimensions: Compared with ordinary text, financial data contains more dimensions, including time series and other information, and the higher the dimension, the more complex the processing. Processing information in high-dimensional financial data sets is a challenge to computing power and the choice of processing methods.

Ensure data purity and diversity: There are many ways to obtain information today, and the data capture process for large language models needs further supervision to ensure the reliability of data sources and prevent the capture of data from misinformation from affecting experimental results. In addition, obtaining data from multiple aspects is crucial to the establishment of financial big language model, and ensuring the diversity of data helps to further enhance the data processing capability of financial big language model.

5. Conclusion

Starting from the big language model, this study introduces BloombergGPT, PIXIU and FinBERT financial big language models based on big language models, and introduces their data set construction ideas, model processing ability, experimental process results and shortcomings. Through efficient processing speed, in terms of obtaining financial information and analyzing market sentiment, financial big language model greatly improves the information collection and processing ability of practitioners in the financial industry, simplifies many tedious information processing processes, improves the acquisition speed of financial information, and helps users to obtain market information after analysis in a timely manner.

However, on the moral level, the output content of the model, as well as the objectivity and rationality of the judgment given by the model, deserve further supervision and consideration. At the data level, the processing of high-dimensional data, as well as the purity and diversity of the data sets used for training, require further thinking and changes to optimize the capabilities and output of the financial big language model.

At present, with the further development of artificial intelligence and the continuous training of large language model data sets, this paper believes that financial large language model still has a very large space for development, and it is hoped to achieve more intelligent data analysis, processing and output capabilities, and further provide users with better quality and more helpful results.

References

[1] T. Brown, B. Mann, N. Ryder, et al. Language models are few-shot learners. NeurIPS, 33: 1877 – 1901. (2020).

[2] T. Almutiri, F. Nadeem, Markov models applications in natural language processing: a survey. Int. J. Inf. Technol. Comput. Sci 2, 1 – 16 (2022).

[3] Tom B. Brown et al. Language Models are Few-Shot Learners, arXiv:2005.14165 (2020).

[4] J. Devlin, M. Chang, K. Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805 (2019).

[5] T. Scao, A. Fan, C. Akiki, et al. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv: 2211.05100. (2022).

[6] S. Wu, O. Irsoy, S. Lu, et al. BloombergGPT: A Large Language Model for Finance, arXiv: 2303.17564v3 (2023).

[7] J. Hoffmann, S. Borgeaud, A. Mensch et al. (2022). An empirical analysis of compute-optimal large language model training. Adv. neural inf. process. syst, 35, 30016 - 30030 (2022).

[8] M. Maia, S. Handschuh, A. e Freitas, et al. financial opinion mining and question answering. The Web Conference 2018, pages 1941 – 1942 (2018).

[9] Q. Xie, W. Han, X. Zhang, et al. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance, arXiv: 2306.05443v1 (2023).

[10] OpenAI. GPT-4 Technical Report. arXiv: 2303.08774 (2023).

[11] T. Le Scao, A. Fan, C. Akiki, et al. Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv: 2211.05100 (2022).

[12] S. Zhang, S. Roller, N. Goyal, et al. Opt: Open pre-trained transformer language models. arXiv preprint arXiv: 2205.01068 (2022).

[13] Z. Zhang, H. Zhang, K. Chen, et al. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. arXiv preprint arXiv: 2110.06696 (2021).

[14] M. E Peters, M. Neumann, M. Iyyer, et al. Deep contextualized word representations. https: //doi.org/10.18653/v1/N18- 1202 arXiv: 1802.05365 (2018).

[15] J. Howard, S. Ruder. Universal Language Model Fine- tuning for Text Classification. arXiv: 1801.06146 http://arxiv.org/abs/ 1801.06146 (2018).

[16] P. Malo, A. Sinha, P. Korhonen, et al. good debt or bad debt: Detecting semantic orientations in economic texts. J. Assoc. Inf. Sci. Technol. 65, 4, 782 – 796 (2014).

[17] D. Tan Araci, FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv preprint arXiv: 1908.10063 (2019).