FinBERT: A Pretrained Language Model for Financial Communications

Yi Yang   Mark Christopher Siy UY   Allen Huang

School of Business and Management, Hong Kong University of Science and Technology

`{imyiyang,acahuang}@ust.hk, mcsuy@connect.ust.hk`

Abstract

Contextual pretrained language models, such as BERT (Devlin et al., 2019), have made significant breakthrough in various NLP tasks by training on large scale of unlabeled text resources. Financial sector also accumulates large amount of financial communication text. However, there is no pretrained finance specific language models available. In this work, we address the need by pretraining a financial domain specific BERT models, FinBERT, using a large scale of financial communication corpora. Experiments on three financial sentiment classification tasks confirm the advantage of FinBERT over generic domain BERT model. The code and pretrained models are available at https://github.com/yya518/FinBERT. We hope this will be useful for practitioners and researchers working on financial NLP tasks.

1 Introduction

The growing maturity of NLP techniques and resources is drastically changing the landscape of finanical domain. Capital market practitioners and researchers have keen interests in using NLP techniques to monitor market sentiment in real time from online news articles or social media posts, since sentiment can be used as a directional signal for trading purposes. Intuitively, if there is positive information about a particular company, we expect that company's stock price to increase, and vice versa. For example, Bloomberg, the financial media company, reports that trading sentiment portfolios outperform the benchmark index significantly (Cui et al., 2016). Prior financial economics research also reports that news article and social media sentiment could be used to predict market return and firm performance (Tetlock, 2007; Tetlock et al., 2008).

Recently, unsupervised pre-training of language models on large corpora has significantly improved the performance of many NLP tasks. The language models are pretained on generic corpora such as Wikipedia. However, sentiment analysis is a strongly domain dependent task. Financial sector has accumulated large scale of text of financial and business communications. Therefore, leveraging the success of unsupervised pretraining and large amount of financial text could potentially benefit wide range of financial applications.

To fill the gap, we pretrain **FinBERT**, a finance domain specific BERT model on a large financial communication corpora of 4.9 billion tokens, including corporate reports, earnings conference call transcripts and analyst reports. We document the financial corpora and the FinBERT pretraining details. Experiments on three financial sentiment classification tasks shows that FinBERT outperforms the generic BERT models. Our contribution is straightforward: we compile a large scale of text corpora that are the most representative in financial and business communications. We pretrain and release FinBERT, a new resource demonstrated to improve performance on financial sentiment analysis.

2 Related Work

Recently, unsupervised pre-training of language models on large corpora, such as BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), ULM-Fit (Howard and Ruder, 2018), XLNet, and GPT (Radford et al., 2019) has significantly improved performance on many natural language processing tasks, from sentence classification to question answering. Unlike traditional word embedding (Mikolov et al., 2013; Pennington et al., 2014) where word is represented as a single vector representation, these language model returns

contextualized embeddings for each word token which can be fed into downstream tasks.

The released language models are trained on general domain corpora such as news articles and Wikipedia. Even though it is easy to fine tune the language model using downstream task, it has been shown that pre-training a language model using large-scale domain corpora can further improve the task performance than fine-tuning the generic language model. To this end, several domain-specific BERT models are trained and released. BioBERT (Lee et al., 2019) pretrains a biomedical domain-specific language representation model using large-scale biomedical corpora. Similarly, ClinicalBERT (Huang et al., 2019) applies BERT model to clinical notes for hospital readmission prediction task, and (Alsentzer et al., 2019) applies BERT on clinical notes and discharge summaries. SciBERT (Beltagy et al., 2019) trains a scientific domain-specific BERT model using a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks. We are the first to pre-train and release a finance domain specific BERT model.

3   Financial Corpora

We compile a large financial domain corpora that are most representative in finance and business communications.

Corporate Reports 10-K & 10-Q The most important text data in finance and business communication is corporate report. In the United States, the Securities Exchange Commission (SEC) mandates all publicly traded companies to file annual reports, known as Form 10-K, and quarterly reports, known as Form 10-Q. This document provides a comprehensive overview of the company's business and financial condition. Laws and regulations prohibit companies from making materially false or misleading statements in the 10-Ks. The Form 10-Ks and 10-Qs are publicly available and can be accesses from SEC website.[1]

We obtain 60,490 Form 10-Ks and 142,622 Form 10-Qs of Russell 3000 firms during 1994 and 2019 from SEC website. We only include sections that are textual components, such as Item 1 (Business) in 10-Ks, Item 1A (Risk Factors) in both 10-Ks and 10-Qs and Item 7 (Managements Discussion and Analysis) in 10-Ks.

Earnings Call Transcripts Earnings calls are quarterly conference calls that company executives hold with investors and analysts to discuss firm overall performance. During an earnings call, executives such as CEOs and CFOs read forward-looking statements and provide their information and interpretation of their firms performance during the quarter. Analysts also have the opportunity to request managers to clarify information. Institutional and individual investors listen to the earnings call and spot the tones of executives that portend good or bad news for the company. We obtain 136,578 earnings conference call transcripts of 7,740 public firms between 2004 and 2019. The earnings call transcripts are obtained from the website Seeking Alpha[2].

Analyst Reports Analyst reports are another useful source of information for institutional and individual investors (sri International, 1987). An analyst report typically provides several quantitative summary measures, including a stock recommendation, an earnings forecast, and sometimes a target price. It also provides a detailed, mostly textual analysis of the company. Institutional investors spend millions of dollars annually to purchase the full content of analyst reports to read the written textual analysis. We obtain analyst reports in the Investext database issued for S&P firms during the 1995-2008 period, which yields a set of 488,494 reports.

Overall Corpora Statistics The total size of all 4 corpora is approximately 4.9 billion tokens. We present the pretraining financial corpora statistics in Table 1. As a comparison, BERT's pre-training corpora consists of two textual corpora with a total of 3.3 billion tokens.

Corpus	# of tokens
Corporate Reports 10-K & 10-Q	2.5B
Earnings Call Transcripts	1.3B
Analyst Reports	1.1B

Table 1: Size of pretraining financial corpora.

4   FinBERT Training

Vocabulary We construct FinVocab, a new Word-Piece vocabulary on our financial corpora using the SentencePiece library. We produce both cased and uncased versions of FinVocab, with sizes of

28,573 and 30,873 tokens respectively. This is very similar to the 28,996 and 30,522 token sizes of the original BERT cased and uncased BaseVocab. The resulting overlap between between the original BERT BaseVocab, and FinVocab is 41% for both the cased and uncased versions.

FinBERT-Variants We use the original BERT code [3] to train FinBERT on our financial corpora with the same configuration as BERT-Base. Following the original BERT training, we set a maximum sentence length of 128 tokens, and train the model until the training loss starts to converge. We then continue training the model allowing sentence lengths up to 512 tokens. In particular, we train four different versions of FinBERT: cased or uncased; BaseVocab or FinVocab.

FinBERT-BaseVocab, uncased/cased: Model is initialized from the original BERT-Base uncased/cased model, and is further pretrained on the financial corpora for 250K iterations at a smaller learning rate of $2e^{-5}$, which is recommended by BERT code.

FinBERT-FinVocab, uncased/cased: Model is trained from scratch using a new uncased/cased financial vocabulary FinVocab for 1M iterations.

Training The entire training is done using a NVIDIA DGX-1 machine. The server has 4 Tesla P100 GPUs, providing a total of 128 GB of GPU memory. This machine enables us to train the BERT models using a batch size of 128. We utilize Horovord framework (Sergeev and Del Balso, 2018) for multi-GPU training. Overall, the total time taken to perform pretraining for one model is approximately 2 days. With the release of FinBERT, we hope financial practitioners and researchers can benefit from FinBERT model without the necessity of the significant computational resources required to train the model.

5 Financial Sentiment Experiments

Given the importance of sentiment analysis in financial NLP tasks, we conduct experiments on financial sentiment classification datasets.

5.1 Dataset

Financial Phrase Bank is a public dataset for financial sentiment classification (Malo et al., 2014). The dataset contains 4,840 sentences selected from financial news. The dataset is manually labeled by 16 researchers with adequate back-

ground knowledge on financial markets. The sentiment label is either positive, neutral or negative.

AnalystTone Dataset is a dataset to gauge the opinions in analyst reports, which is commonly used in Accounting and Finance literature (Huang et al., 2014). The dataset contains randomly selected 10,000 sentences from analyst reports in the Investext database. Each sentence is manually annotated into one of three categories: positive, negative and neutral. This classification yields a total of 3,580 positive, 1,830 negative, and 4,590 neutral sentences in the dataset.

FiQA Dataset is an open challenge dataset for financial sentiment analysis, containing 1,111 text sentences [4]. Given an English text sentence in the financial domain (microblog message, news statement), the task of this challenge is to predict the associated numeric sentiment score, ranged from -1 to 1. We convert the original regression task into a binary classification task for consistent comparison with the above two datasets.

We randomly split each dataset into 90% training and 10% testing 10 times and report the average. Since all dataset are used for sentiment classification, we report the accuracy metrics in the experiments.

5.2 Fine-tune Strategy

We follow the same fine-tune architecture and optimization choices used in (Devlin et al., 2019). We use a simple linear layer, as our classification layer, with a softmax activation function. We also use cross-entropy loss as the loss function. Note that an alternative is to feed the contextualized word embeddings of each token into a deep architectures, such as Bi-LSTM, atop frozen BERT embeddings. We choose not to use this strategy as it has shown to perform significantly worse than fine-tune BERT model (Beltagy et al., 2019).

5.3 Experiment Results

We compare FinBERT with original BERT-Base model (Devlin et al., 2019), and we evaluate both cased and uncased versions of this model. The main results of financial sentiment analysis tasks are present in Table 2.

FinBERT vs. BERT The results show substantial improvement of FinBERT models over the generic BERT models. On PhraseBank dataset, the best model uncased FinBERT-FinVocab achieves the

---

	BERT		FinBERT-BaseVocab		FinBERT-FinVocab	
	cased	uncased	cased	uncased	cased	uncased
PhraseBank	0.755	0.835	0.856	0.870	0.864	**0.872**
FiQA	0.653	0.730	0.767	0.796	0.814	**0.844**
AnalystTone	0.840	0.850	0.872	0.880	0.876	**0.887**

Table 2: Performance of different BERT models on three financial sentiment analysis tasks.

		10-Ks/10-Qs	Earnings Call	Analyst Reports	All
BaseVocab	PhraseBank	0.835	0.843	0.845	0.856
	FiQA	0.707	0.731	0.744	0.767
	AnalystTone	0.845	0.862	0.871	0.872
FinVocab	PhraseBank	0.847	0.860	0.861	0.864
	FiQA	0.766	0.778	0.796	0.814
	AnalystTone	0.858	0.870	0.872	0.876

Table 3: Performance of pretraining on different financial corpus.

accuracy of 0.872, a 4.4% improvement over uncased BERT model and 15.4% improvement over cased BERT model. On FiQA dataset, the best model uncased FinBERT-FinVocab achieves the accuracy of 0.844, a 15.6% improvement over uncased BERT model and a 29.2% improvement over cased BERT model. Lastly, on the Analyst-Tone dataset, the best model uncased FinBERT-FinVocab improves the uncased and cased BERT model by 4.3% and 5.5% respectively. Overall speaking, pretraining on financial corpora, as expected, is effective and enhances the downstream financial sentiment classification tasks. In financial markets where capturing the accurate sentiment signal is of utmost importance, we believe the overall FinBERT improvement demonstrates its practical utility.

FinVocab vs. BaseVocab We assess the importance of an in-domain financial vocabulary by pretraining different FinBERT models using BaseVocab and FinVocab. For both uncased and cased model, we see that FinBERT-FinVocab outperforms its BaseVocab counterpart. However, the performance improvement is quite marginal on PhraseBank and AnalystTone task. Only do we see substantial improvement on FiQA task (0.844 vs. 0.796). Given the magnitude of improvement, we suspect that while an in-domain vocabulary is helpful, FinBERT benefits most from the financial communication corpora pretraining.

Cased vs. Uncased We follow (Devlin et al., 2019) in using both the cased model and the uncased model for all tasks. Experiments result suggest that uncased models perform better than cased models in all tasks. This result is consistent with

prior work of Scientific domain and Biomedical domain BERT models.

Corpus Contribution We also train different FinBERT models on three financial corpus separately. The performance of different FinBERT models (cased version) on different tasks are present in Table 3. It shows that FinBERT trained on all corpora achieves the overall best performance indicating that combining additional financial communication corpus could improve the language model quality. Among three datasets, Analyst Reports dataset appears to perform well in three different tasks, even though it only has 1.1 billion word tokens. Prior research finds that corporate report such as 10-Ks and 10-Qs contains redundant content, and that a substantial amount of textual volume contained in 10-K reports is attributable to managerial discretion in how firms respond to mandatory disclosure requirements (Cazier and Pfeiffer, 2016). Does it suggest that Analyst Reports data contains more information content than corporate reports and earnings call transcripts? We leave it for future research.

6 Conclusion

In this work, we pre-train a financial-task oriented BERT model, FinBERT. The FinBERT model is trained on a large financial corpora that are representative of English financial communications. We show that FinBERT outperforms generic BERT models on three financial sentiment classification tasks. With the release of FinBERT, we hope practitioners and researchers can utilize FinBERT for a wider range of applications where the prediction target goes beyond sentiment, such as

financial-related outcomes including stock returns, stock volatilities, corporate fraud, etc.

Acknowledgments

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *Proceedings of EMNLP*.

Richard A Cazier and Ray J Pfeiffer. 2016. Why are 10-k filings so long? *Accounting Horizons*, 30(1):1–21.

X Cui, D Lam, and A Verma. 2016. Embedded value in bloomberg news and social sentiment data. *Bloomberg LP*.

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

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings ACL*, pages 328–339.

Allen H Huang, Amy Y Zang, and Rong Zheng. 2014. Evidence on the information content of text in analyst reports. *The Accounting Review*, 89(6):2151–2180.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

sri International. 1987. *Investor information needs and the annual report*. Financial Executives Research Foundation.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alexander Sergeev and Mike Del Balso. 2018. Horovod: fast and easy distributed deep learning in tensorflow. *arXiv preprint arXiv:1802.05799*.

Paul C Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467.