# FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining

**Zhuang Liu**[1,*] , **Degen Huang**[1] , **Kaiyu Huang**[1] , **Zhuang Li**[2] and **Jun Zhao**[2]
[1]Dalian University of Technology, Dalian, China
[2]Union Mobile Financial Technology Co., Ltd., Beijing, China

## Abstract

There is growing interest in the tasks of financial text mining. Over the past few years, the progress of Natural Language Processing (NLP) based on deep learning advanced rapidly. Significant progress has been made with deep learning showing promising results on financial text mining models. However, as NLP models require large amounts of labeled training data, applying deep learning to financial text mining is often unsuccessful due to the lack of labeled training data in financial fields. To address this issue, we present FinBERT (BERT for Financial Text Mining) that is a domain specific language model pre-trained on large-scale financial corpora. In FinBERT, different from BERT, we construct six pre-training tasks covering more knowledge, simultaneously trained on general corpora and financial domain corpora, which can enable FinBERT model better to capture language knowledge and semantic information. The results show that our FinBERT outperforms all current state-of-the-art models. Extensive experimental results demonstrate the effectiveness and robustness of FinBERT. The source code and pre-trained models of FinBERT are available online.

## 1 Introduction

In finance and economics, various financial text data are used to analyze and predict future market trends. Whether for analyst reports or official company announcements, financial texts mining play a crucial role in financial technology. The volume of financial text data continues to rapidly increase. An unprecedented number of such texts are created every day, so for any single entity, manually analyzing these texts and gaining actionable insights from them is almost an extremely difficult task. Advances in machine learning technology have made financial text mining models in FinTech possible. However, in financial text mining tasks, constructing supervised training data is prohibitively expensive as this requires the use of expert knowledge in finance fields. Therefore, due to the small amount of labeled training data that can be used for financial text mining tasks, most financial text mining models cannot directly utilize deep learning technology.

In this paper, we propose FinBERT model addressing the issue by leveraging unsupervised Transfer Learning and Multi-task Learning. Word embedding, such as word2vec [Mikolov *et al.*, 2013] is a method of extracting knowledge from an unsupervised data and has become one of the major advancements in natural language processing (NLP). However, because of the special language used in the financial field, these simple word embedding approaches are not effective enough. Pre-trained Language Models (PLMs), such as BERT [Devlin *et al.*, 2019], pre-trained on a large-scale unsupervised data (such as Wikipedia) to improve contextualized representations more effectively. However, in the task of financial text mining, due to the large differences in vocabulary and expression between the financial corpus and the general domain corpus, they still cannot be effectively applied to financial data. Furthermore, the pre-training of PLMs usually focuses on training the model through a few simple tasks. For example, BERT uses MaskLM and NSP as pre-training objectives. However, in fact, vocabulary, semantics, sentence order and proximity between sentences, all of which can enable the PLMs to learn more language knowledge and semantic information in the training corpus. Especially for financial text data, for example, named entities like stock, bond type and financial institution names, contain unique vocabulary information. Therefore, in order to efficiently capture language knowledge and semantic information in large-scale training corpora, we construct six pre-training tasks covering more knowledge, and train FinBERT through multi-task learning on training data. Specifically, proposed FinBERT differs from standard PLMs pre-training methods. It constructs six pre-training tasks, simultaneously trained on general corpora and financial domain corpora, to help FinBERT better capture language knowledge and semantic information.

In summary, the main contributions of our paper are the following:

- FinBERT is the first domain specific BERT that is pre-trained through multi-task learning on financial corpora, to transfer knowledge from financial domain corpora.
- Our FinBERT model differs from standard BERT in the training objectives. We construct six self-supervised

[*]Corresponding author: liuzhuang.dlut@gmail.com

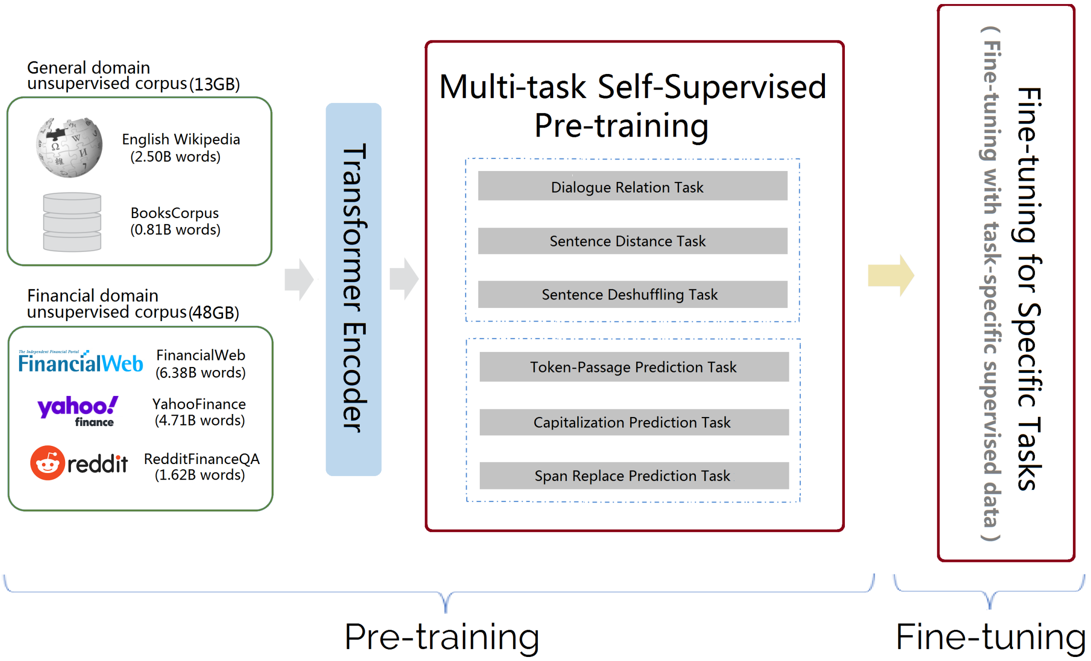
General domain unsupervised corpus (13GB)
English Wikipedia (2.50B words)
BooksCorpus (0.81B words)
Financial domain unsupervised corpus (48GB)
FinancialWeb
FinancialWeb (6.38B words)
yahoo! finance
YahooFinance (4.71B words)
reddit
RedditFinanceQA (1.62B words)
Transformer Encoder
Multi-task Self-Supervised Pre-training
Dialogue Relation Task
Sentence Distance Task
Sentence Deshuffling Task
Token-Passage Prediction Task
Capitalization Prediction Task
Span Replace Prediction Task
Fine-tuning for Specific Tasks
( Fine-tuning with task-specific supervised data )
Pre-training
Fine-tuning

Figure 1: An illustration of the architecture for FinBERT, where the model can be simultaneously trained on a general corpus and a financial domain corpus, and six pre-training tasks can be pre-trained through multi-task self-supervised pre-training learning, and last the pre-trained model is fine-tuned using task-specific supervised data to adapt to various language understanding tasks.

pre-training tasks (subsection 2.2), which can be learned through multi-task self-supervised learning, capable of efficiently capturing language knowledge and semantic information in large-scale pre-training corpora.

- We conduct extensive experiments on several financial benchmark datasets. Experimental results show that proposed FinBERT outperforms all previous state-of-the-art models in financial Sentence Boundary Detection, financial Sentiment Analysis, and financial Question Answering, respectively. (subsection 3.3).
- We implemented our FinBERT on Horovod framework using mixed precision training methodology (section 3.1). We make the source code and pre-trained models publicly available. With minimal task-specific architecture modifications, FinBERT can be used for various other downstream financial text mining tasks to help significantly boost overall performance.

## 2 Proposed Model: FinBERT

As shown in Figure 1, our proposed FinBERT model is built based on the standard BERT architecture [Devlin *et al.*, 2019] based on the two-stage '*Pre-training*'-then-'*Fine-tuning*' pre-training language model approach, which recently become enormously popular in NLP. During pre-training phase, the FinBERT model differs from standard BERT architecture pre-training methods in that, instead of training with MaskLM and Next Sentence Prediction (NSP) pre-training objectives, it constructs a large variety of pre-training objectives to help the FinBERT model better capture language knowledge and semantic information. On top of that, FinBERT keeps updating the pre-trained model through multi-task self-supervised pre-training learning. Meantime, compared with traditional pre-training models, FinBERT is simultaneously trained on a general corpus and a financial domain corpus. During fine-tuning phase, FinBERT is first initialized with the pre-trained parameters, and is later fine-tuned on task-specific supervised data.

In this section, we will briefly introduce FinBERT in our proposed framework.

### 2.1 Transformer Encoder

First, the input embedding module is responsible for converting each word into an embedding representation that can be fed into the Transformer Encoder. The Transformer [Vaswani *et al.*, 2017] encoder is based on self-attention mechanism. Self-attention mechanism can capture global context information by pairwise correlation. The multi-layer Transformer captures the context information of each token through stacking the self-attentions. Followed by BERT and ERNIE2 [Sun *et al.*, 2019b], the embedding representation is the sum of four parts: token embedding, segment embedding, position embedding, and task embedding. We use different task IDs for different tasks. Each task ID is assigned to a unique task embedding, ranging from 0 to 5. Then, FinBERT uses the multi-layer Transformer architecture as the encoder.

### 2.2 Multi-task Self-Supervised Pre-training

The choice of unsupervised pre-training objective plays an important role in applying to pre-training stage by continuously gains general knowledge. We will modify and combine multiple common unsupervised pre-training tasks [Joshi *et al.*, 2019; Dong *et al.*, 2019; Liu *et al.*, 2019a; Lan *et al.*, 2020; Devlin *et al.*, 2019; Liu *et al.*, 2019b; Raffel *et al.*, 2019; Sun *et al.*, 2019b] fit our framework. Specifically, in this layer, we construct six unsupervised pre-training tasks to learn different level knowledge from the training corpora. As shown in Figure 1, **i)** Basic-level Pre-training tasks: Span Replace Prediction pre-training task, Capitalization Prediction pre-training task and Token-Passage Prediction pre-training task. **ii)** High-level Pre-training tasks: Sentence Deshuffling pre-training task, Sentence Distance pre-training task and Dialogue Relation pre-training task. Next, we introduce these six self-supervised pre-training tasks in detail.

**Span Replace Prediction pre-training task.** Inspired by spanBERT [Joshi *et al.*, 2019] and T5 [Raffel *et al.*, 2019], we constructed a self-supervised pre-training objective that randomly samples and drop out 15% in the input text. Instead of replacing each token with a masked token, we use a unique masked token to replace all of each successive boundary, i.e., the continuous span of all consecutively discarded tokens will be replaced with a mask token. Last, we predict the masked span by the tokens observed at the span boundary.

**Capitalization Prediction pre-training task.** In finance and economics, capital words usually have specific semantic value in a sentence. The cased model exhibits certain advantages in tasks like financial named entity recognition, such as stock, bond type and financial institution name. Similar to ERNIE2 [Sun *et al.*, 2019b], we do so by introducing a capitalization words prediction objective that involves predicting whether the word is capitalized or not.

**Token-Passage Prediction pre-training task.** We constructed a pre-training task to identify the key words of a passage appearing in the segment, which can enable FinBERT to capture the topics of a passage. Empirically, in financial news, the words appearing in the passage many times are usually used words commonly and usually relevant with the main topics of the passage. Similar to ERNIE2 [Sun *et al.*, 2019b], we also do so by introducing a token-passage prediction pre-training objective to predict whether the token appears in segments of the original passage.

**Sentence Deshuffling pre-training task.** We also constructed a sentence reordering pre-training objective which is used e.g. in ERNIE2 [Sun *et al.*, 2019b] and T5 [Raffel *et al.*, 2019], to learn the relationships among sentences. We split a given paragraph into 1 to $n$ segments randomly shuffle it by a random permuted order, last we use the original deshuffled sequence as a training target, and pre-train the model to reorganize permuted segments which are modeled with a multi-class ($\sum n!$) classification task.

**Sentence Distance pre-training task.** Standard BERT [Devlin *et al.*, 2019] uses Next Sentence Prediction (NSP) as a training target, which is a binary classification pre-training task. We also constructed a self-supervised training target to predict sentence distance, inspired by BERT [Devlin *et al.*, 2019]. But differs in that we create a three-class classification task, rather than a binary classification. Specifically, we define a three-class classification problem of two sentences to classify as "00", "01" or "11". "00" means that they are in the same passage and adjacent, "01" means that they are in the same passage and not adjacent, and "11" means that they are not in the same passages.

**Dialogue Relation pre-training task.** We also constructed a self-supervised pre-training task to learn the semantic relevance among sentences using Question answering (QA) data. Empirically, QA data is important for semantic relation, since the corresponding question semantics of the same answer are usually very similar. Therefore, similar to ERNIE [Sun *et al.*, 2019a], we constructed the QA Relation pre-training task, which enables the model to learn implicit relationships and learn semantic relevance.

# 3 Experiments

## 3.1 Pre-training FinBERT

### Pre-training Data

Similar to that of BERT, training data in the English corpus are from BookCorpus and English Wikipedia. In order to apply the text mining model to financial texts, we further collected various financial data that are crawled on financial websites, such as financial news and dialogue. Specifically, we consider five English financial corpora of varying domains and sizes, totaling over 61 GB text:

- English Wikipedia[1] and BooksCorpus (Zhu et al., 2015), which are the original training data used to train BERT (totaling 13GB, 3.31B words);
- FinancialWeb (totaling 24GB, 6.38B words), which we collect from the CommonCrawl News dataset[2] between July 2013 and December 2019, containing 13 million financial news (15GB after filtering), together with web-crawled financial articles from FINWEB[3] (9GB after filtering);
- YahooFinance (totaling 19GB, 4.71B words), a dataset crawled from Yahoo Finance[4]. We crawled financial articles (published in the last four years) from Yahoo Finance, and performed data cleaning (removing markup, removing non-textual content and filtering out redundant data);
- RedditFinanceQA (totaling 5GB, 1.62B words), a corpus that contains automatically collected question-answer pairs about financial issues from Reddit[5] website with at least four up votes.

The statistics for all pre-training data are reported in Table 1. We have built and maintained an open repository of financial corpora[6] that can be accessed and analyzed by anyone.

### Pre-training Settings

Our models, FinBERT$_{\mathrm{LARGE}}$ and FinBERT$_{\mathrm{BASE}}$, have the same model settings of transformer and pre-training hyper-parameters as BERT. Like BERT, proposed multi-task self-

[1] https://dumps.wikimedia.org/enwiki/
[2] http://commoncrawl.org/
[3] https://www.finweb.com/
[4] https://finance.yahoo.com/
[5] https://www.reddit.com/
[6] https://drive.google.com/finbert/data/

| Corpus | size(G) | # of words(B) | Domain |
|---|---|---|---|
| English Wikipedia | 9 | 2.50 | General |
| BooksCorpus | 4 | 0.81 | General |
| FinancialWeb | 24 | 6.38 | Financial |
| YahooFinance | 19 | 4.71 | Financial |
| RedditFinanceQA | 5 | 1.62 | Financial |

***Notes:*** We pre-train FinBERT on five English corpora, totaling over 61GB text (16.02 billion words): general domain corpora (Wikipedia + BooksCorpus, totaling 3.31 billion words) and financial domain corpora (FinancialWeb + YahooFinance + RedditFinanceQA, totaling 12.71 billion words).

Table 1: List of text corpora used for FinBERT Corpus.

supervised pre-training also has great requirements on computing power. In our work, we use the architecture we developed for FinBERT pre-training. Our architecture is based on the flexible scheduling of hundreds of GPUs implemented by Apache Hadoop YARN, while providing a distributed training solution based on Horovod framework [Sergeev and Balso, 2018], using parallelized training approach. Horovod is an open source library[7] developed by Uber [9], which is mainly based on [Goyal *et al.*, 2017] and baidu-allreduce[8]. Our architecture (based on Horovod) essentially utilizes a distributed optimizer strategy that is an optimizer wrapping *tf.Optimizer*, and before applying the gradient to the model weights it uses an *allreduce* operation to average gradient values. Empirically, with the increase in the number of GPUs, Our architecture's performance loss is much smaller than that based on standard distributed TensorFlow [Abadi *et al.*, 2016], and the training speed can reach much three times. Moreover, compared to the distributed Tensorflow framework, Our architecture can still guarantee a very stable acceleration ratio on the scale of hundreds of GPU cards, and has good scalability.

[Micikevicius *et al.*, 2018] proposed mixed precision training methodology for training deep neural networks (DNN), which can reduce the memory consumption and time spent in memory and arithmetic operations of DNN. In our work, we use mixed precision training approach for multi-task self-supervised pre-training. Specifically, we train our FinBERT using half-precision format FP16, used for storage and arithmetic. We store activations, gradients and weights using in FP16, and update the copy of weights using in FP32. We maintain a single-precision copy of weights, after each optimizer step which accumulates the gradients. Followed by [Micikevicius *et al.*, 2018], we use loss-scaling to preserve gradient values with small magnitudes. We use FP16 arithmetic that accumulates into single-precision outputs, converted to FP16 before storing to memory.

## 3.2 Fine-tuning FinBERT

Typically, BERT [Devlin *et al.*, 2019] has two successive steps, one during the pre-training phase and one during the fine-tuning phase. It firsts conducts unsupervised pre-training on the large corpus during the pre-training phase, and then conducts supervised fine-tuning on down-stream NLP tasks during the fine-tuning phase. Similar to BERT, we pre-train our FinBERT model from scratch on these large unsupervised corpus (subsection 3.1), and fine-tune/apply it to various down-stream supervised financial text mining tasks. Next, we briefly describe three financial text mining tasks.

**Financial Sentence Boundary Detection (SBD)**

Financial SBD is a basic task for financial text mining, whose aim is to extracting well segmented sentences from financial prospectuses by disambiguating/detecting sentence boundaries of texts, i.e., *beginning* boundary and *ending* boundary. In our work, financial SBD dataset used is FinSBD Shared Task[9], which is a dataset that was created for IJCAI19 FinNLP challenge. FinSBD-2019 provides training data with boundary labels (*beginning* boundary vs. *ending* boundary) for each token.

**Financial Sentiment Analysis (SA)**

Financial SA is one of the most fundamental financial text mining tasks. Given a text in the financial domain, SA aims to detect the target aspects which are mentioned in the financial text and predict the sentiment score for each of the mentioned targets. Sentiment scores will be defined using continuous numeric values ranged from -1(negative) to 1(positive). In this work, financial SA datasets used are Financial PhraseBank[10] [Malo *et al.*, 2014] and FiQA[11] Task 1 [FiQ, 2018]. FiQA is a dataset that was created for WWW18 conference financial opinion mining and question answering challenge. Here we use the data of Task 1, "Aspect-based financial sentiment analysis".

**Financial Question Answering (QA)**

Question Answering (QA) of natural language text is a fundamental challenge in natural language processing (NLP), whose aims is to automatically provide answers to questions related to a given short text or passage. Financial QA involves answering client questions such as "*Why are big companies like Apple or Google not included in the Dow Jones Industrial Average (DJIA) index?*". Financial QA dataset used in

[7] https://github.com/horovod/horovod/

[8] https://github.com/baidu-research/baidu-allreduce/

[9] The FinSBD-2019 dataset contains financial text that had been pre-segmented automatically. There are 953 distinct beginning tokens and 207 distinct ending tokens in the training and dev sets of FinSBD-2019 data. It's available at https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/.

[10] Financial Phrasebank dataset consists of 4845 English sentences selected randomly from financial news found on LexisNexis database. These sentences then were annotated by 16 people with background in finance and business. It's available at https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10/.

[11] FiQA SA dataset includes two types of discourse: financial news headlines and financial microblogs, with manually annotated target entities, sentiment scores and aspects. The financial news headlines dataset contains a total 529 annotated headlines samples (436 samples for the training set and 93 samples for the test set) while the financial microblogs contains a total 774 annotated posts samples (675 samples for the training set and 99 samples for the test set). It's available at https://sites.google.com/view/fiqa/home/.

this paper is FiQA Task 2 [FiQ, 2018]. FiQA is a dataset that was created for WWW18 conference financial opinion mining and question answering challenge[12]. Here we use the data of Task 2, "Opinion-based QA over financial data".

## 3.3 Experimental Results

**Financial Sentence Boundary Detection (SBD)**

Table 2 shows our results on financial SBD. The *beginning* (**BS**) and *ending* (**ES**) tokens of the sentence are separately evaluated, the official evaluation metrics of FinSBD-2019 include: **i)** F1 scores, separately used to predict **BS** and **ES** ; **ii)** the mean of F1 scores. As shown in Table 2, proposed FinBERT pre-trained on both general domain dataset and financial domain dataset is highly effective. Both FinBERT$_{\mathrm{BASE}}$ and FinBERT$_{\mathrm{LARGE}}$ outperformed the current state-of-the-art model. FinBERT$_{\mathrm{LARGE}}$ outperformed BERT-S by 0.085 in terms of **MEAN** score.

| **Model** | **ES** | **BS** | **MEAN** |
|---|---|---|---|
| Rule-based [Fatima *et al.*, 2019] | 0.80 | 0.86 | 0.830 |
| BiLSTM-CRF [Du *et al.*, 2019] | 0.83 | 0.88 | 0.855 |
| Deep-Att [Tian and Peng, 2019] | 0.83 | 0.91 | 0.875 |
| BERT-S [Du *et al.*, 2019] | 0.88 | 0.89 | 0.885 |
| FinBERT$_{\mathrm{BASE}}$ (ours) | 0.91 | 0.92 | 0.915 |
| FinBERT$_{\mathrm{LARGE}}$ (ours) | **0.96** | **0.98** | **0.970** |

Table 2: Experimental results on test set for FinSBD English task.

**Financial Sentiment Analysis (SA)**

The results of PhraseBank sentiment dataset are shown in Table 3. The results of FiQA sentiment dataset are shown in Table 4. From the data in Table 3 and Table 4, it is apparent that our FinBERT$_{\mathrm{BASE}}$ and FinBERT$_{\mathrm{LARGE}}$ consistently significantly outperforms all baseline models on PhraseBank sentiment dataset and FiQA sentiment dataset, achieving state-of-the-art results. Overall, experimental results highlight the importance of the financial domain-specific corpora pre-trained design.

| **Model** | **Accuracy** | **F1** |
|---|---|---|
| LPS [Malo *et al.*, 2014] | 0.71 | 0.71 |
| HSC [Krishnamoorthy, 2018] | 0.71 | 0.76 |
| ULMFit [Raaci, 2019] | 0.83 | 0.79 |
| FB-SA [Raaci, 2019] | 0.86 | 0.84 |
| FinBERT$_{\mathrm{BASE}}$ (ours) | 0.91 | 0.89 |
| FinBERT$_{\mathrm{LARGE}}$ (ours) | **0.94** | **0.93** |

Table 3: Experimental results on the test set for the Financial PhraseBank sentiment dataset.

|  | *headline* |  | *post* |  |
|---|---|---|---|---|
| **Model** | **MSE** | **R**$^2$ | **MSE** | **R**$^2$ |
| CUKG [FiQ, 2018] | 0.13 | 0.46 | 0.10 | 0.09 |
| IIT-Dehi $^\natural$ | 0.20 | 0.18 | 0.10 | 0.08 |
| Inf-UFG $^\natural$ | 0.21 | 0.17 | 0.10 | 0.16 |
| NLP301 [FiQ, 2018] | - | - | 0.31 | -1.67 |
| SC-V [Yang *et al.*, 2018] | 0.08 | 0.40 | - | - |
| RCNN [Piao *et al.*, 2018] | 0.09 | 0.41 | - | - |
| FB-SA [Raaci, 2019] | 0.07 | 0.55 | - | - |
| FinBERT$_{\mathrm{BASE}}$ (ours) | 0.29 | 0.67 | 0.28 | 0.26 |
| FinBERT$_{\mathrm{LARGE}}$ (ours) | **0.38** | **0.77** | **0.37** | **0.36** |

***Notes:*** FiQA sentiment dataset includes two types of discourse: *financial news headlines* and *financial microblogs*. In order to evaluate the sentiment scores models, regression model evaluation measures were used during the experiments, Mean Squared Error (**MSE**), R Square (**R**$^2$). $^\natural$ indicates results are taken from WWW18 Shared Task.

Table 4: Experimental results on the test set for the FiQA sentiment dataset.

**Financial Question Answering (QA)**

Table 5 shows our results on financial QA. From the experiments, we can clearly see that both FinBERT$_{\mathrm{BASE}}$ and FinBERT$_{\mathrm{LARGE}}$ are much better than the previous models including the state-of-the-art model. As shown in Table 5, FinBERT$_{\mathrm{LARGE}}$ significantly outperformed BERT and the state-of-the-art model, and achieved a *nDCG* of 0.76 and a mean reciprocal rank (*MRR*) score of 0.68. Considering that it is prohibitively expensive to collect labeled training data for financial Question Answering (QA) task. Very often, we only have very small training data (only few hundreds of samples) or even no training data. Even FinBERT$_{\mathrm{BASE}}$ outperformed the state-of-the-art model by 2 in terms of both *nDCG* and *MRR* score. The experimental results are highly effective and encouraging.

On all the financial datasets, such as financial SBD, SA and QA, FinBERT achieved state-of-the-art results, which proves the effectiveness of proposed FinBERT.

| **Model** | **nDCG** | **MRR** |
|---|---|---|
| CUKG-TongJi [Champin *et al.*, 2018] | 0.17 | 0.10 |
| eLabour [Champin *et al.*, 2018] | 0.31 | 0.19 |
| FinBERT$_{\mathrm{BASE}}$ (ours) | 0.63 | 0.51 |
| FinBERT$_{\mathrm{LARGE}}$ (ours) | **0.76** | **0.68** |

***Notes:*** To evaluate financial question answering scores models, ranking evaluation measures were used during the experiments: Normalized Discounted Cumulative Gain (**nDCG**) and Mean reciprocal rank (**MRR**). Bold face indicates best result in the corresponding metric.

Table 5: Experimental results on the test set for the FiQA question answering dataset.

[12] Financial QA dataset is built by crawling Stack exchange posts under the Investment topic in the period between 2009 and 2017. The final dataset contains a KB of 57,640 answer posts with 17,110 question-answer pairs for training and 531 question-answer pairs for testing. It's available at https://sites.google.com/view/fiqa/home/.

# 4 Ablation Study and Analyses

In this section, We conduct ablation studies and further compare proposed FinBERT with existing PLMs models and present the results along with experimental details.

## 4.1 Effect of Pre-training on the Performance

We further measure the effect of pre-training on the performance of the classifier. We compare four models: **i)** No further pre-training (denoted by Vanilla BERT, Vanilla FinBERT), **ii)** Further pre-training on classification training set (denoted by BERT-task, FinBERT-task). As shown in Table 6, FinBERT-task achieves better performance than that of all other models. Specially, although BERT-task is further pre-trained on the financial classification training set (FiQA Aspect category classification task), Vanilla FinBERT outperforms two Vanilla BERT based models, Vanilla BERT and BERT-task, 0.09% higher than Vanilla BERT and 0.02% higher than BERT-task, in terms of *Accuracy* score. Overall, this experimental result shows the strength of proposed FinBERT, and also shows that FinBERT learned domain-specific knowledge from a large number of financial domain corpora during the pre-training phase.

| Model | Accuracy | Precision | Recall |
|---|---|---|---|
| Vanilla BERT | 0.78 | 0.33 | 0.30 |
| BERT-Task | 0.85 | 0.51 | 0.45 |
| Vanilla FinBERT | 0.87 | 0.56 | 0.51 |
| FinBERT-Task | **0.91** | **0.72** | **0.69** |

Table 6: Experimental results on the test set for the financial classification dataset (FiQA Aspect category classification task).

|  | SBD | SA |  | QA |  |
|---|---|---|---|---|---|
| **Model** | **MEAN** | **Acc.** | **F1** | **nDCG** | **MRR** |
| BERT | 0.86 | 0.84 | 0.82 | 0.39 | 0.27 |
| FinBERT | **0.94** | **0.89** | **0.90** | **0.73** | **0.64** |

Table 7: The performance of BERT and FinBERT on three financial tasks (SBD, SA and QA) when they are trained on a small corpus.

## 4.2 Pre-training with Small Training Data

Deep learning models require a lot of pre-training data. However, due to the lack of training data in the financial field, in many applications in financial fields, large corpora may not be available. To further demonstrate the advantages of our FinBERT, we conducted another experiment, which was based on a simulated small corpus to pre-train BERT and FinBERT respectively. Specifically, we randomly select 1/5 size of the entire financial data set as a simulated small corpus. Then, we pre-train both models based on this corpus, and use the previous experiment to test the same tasks. In Table 7, we report the results of our FinBERT and BERT based on the Financial SBD, Financial SA, and Financial QA, respectively. As shown in Table 7, FinBERT constantly outperform BERT on all three financial tasks. Considering that both models are trained on small corpus, the experimental results are highly encouraging, which shows that proposed FinBERT can provide stable and clear enhancement when trained on financial corpora of different sizes. Overall, this experiment simulates that our FinBERT model can better encode financial text in the case of limited data, proving that FinBERT still has great advantages under the small training data scenarios in financial domain.

# 5 Related Work

## 5.1 Unsupervised Transfer Learning for PLMs

Pre-trained Language Models (PLMs) has attracted extensive attention. Fine-tuning of PLMs has shown that it can effectively improve downstream NLP tasks. PLMs such as word2vec [Mikolov *et al.*, 2013] and ELMo [Peters *et al.*, 2018] is an approach of extracting knowledge from a large-scale unlabeled data and has become one of the major advances in NLP. [Du *et al.*, 2019; Tian and Peng, 2019; Liu *et al.*, 2020; Joshi *et al.*, 2019; Dong *et al.*, 2019; Liu *et al.*, 2019a] presented models to tackle financial domain tasks. However, because of the special language used in the financial field and the large differences in vocabulary and expression between the financial corpus and the general domain corpus, these PLMs approaches are not effective enough.

## 5.2 Multi-task Learning

[Raaci, 2019; Devlin *et al.*, 2019] usually focuses on training the model through a few simple tasks. For example, BERT uses MaskLM and NSP as pre-training objectives. [Du *et al.*, 2019] used word embeddings as input to the BiLSTM-CRF model for sentence boundary detection task. [Yang *et al.*, 2018] used a word and its surrounding context as input to neural network. However, all of these models are based on a single-task method and cannot use the information that is contained in multiple data as in Multi-Task Learning. Although models based on multi-task learning has been widely used in NLP research, the application of multi-task learning in financial text mining has not yet seen promising results.

# 6 Conclusion

In this paper, we proposed FinBERT that is a pre-trained language model for financial text mining. By constructing six pre-training tasks covering more knowledge, we simultaneously trained FinBERT on general corpora and financial domain corpora, which enabled FinBERT model effectively to capture language knowledge and semantic information. Also, we implemented our FinBERT on Horovod framework using mixed precision training methodology. Extensive experimental results demonstrated the effectiveness and robustness of FinBERT.

# Acknowledgments

We would like to thank the reviewers for their helpful comments and suggestions to improve the quality of the paper. The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China under (No.61672127, U1916109).

## References

[Abadi *et al.*, 2016] Martín Abadi, Ashish Agarwal, and Paul Barham. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. volume abs/1603.04467, 2016.

[Champin *et al.*, 2018] Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors. *In Proceedings of WWW*. ACM, 2018.

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

[Dong *et al.*, 2019] Li Dong, Nan Yang, and Wenhui Wang. Unified language model pre-training for natural language understanding and generation. In *Proceedings of NeurIPS*, pages 13042–13054, 2019.

[Du *et al.*, 2019] Jinhua Du, Yan Huang, and Karo Moilanen. Sentence boundary detection through sequence labelling and BERT fine-tuning. In *FinNLP IJCAI 2019*, pages 81–87, Macao, China, August 2019.

[Fatima *et al.*, 2019] Mehwish Fatima, Mark-Christoph Mueller, and Christoph Mark. Machine learning vs. rule-based sentence boundary detection. In *FinNLP IJCAI 2019*, pages 115–121, Macao, China, August 2019.

[FiQ, 2018] In Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross Mcdermott, and Manel Zarrouk, editors, *In Proceedings of WWW*, 2018.

[Goyal *et al.*, 2017] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017.

[Joshi *et al.*, 2019] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529, 2019.

[Krishnamoorthy, 2018] Srikumar Krishnamoorthy. Sentiment analysis of financial news articles using performance indicators. *Knowl. Inf. Syst.*, 56(2):373–394, 2018.

[Lan *et al.*, 2020] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. 2020.

[Liu *et al.*, 2019a] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of ACL*, pages 4487–4496, 2019.

[Liu *et al.*, 2019b] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, and Danqi Chen. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[Liu *et al.*, 2020] Zhuang Liu, Keli Xiao, Bo Jin, Kaiyu Huang, and Yunxia Zhang. Unified generative adversarial networks for multiple-choice oriented machine comprehension. *ACM Transactions on Intelligent Systems and Technology*, 11(3):1–20, 2020.

[Malo *et al.*, 2014] Pekka Malo, Ankur Sinha, Pekka J. Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *JASIST*, 65(4):782–796, 2014.

[Micikevicius *et al.*, 2018] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *Proceedings of ICLR*, 2018.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Computer Science*, 2013.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, and Christopher Clark. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237, 2018.

[Piao *et al.*, 2018] Guangyuan Piao, John G. Breslin, and Ross Mc-Dermott. Financial aspect and sentiment predictions with deep neural networks: An ensemble approach. In *Proceedings of WWW*, pages 1973–1977, 2018.

[Raaci, 2019] Dogu Raaci. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063, 2019.

[Raffel *et al.*, 2019] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[Sergeev and Balso, 2018] Alexander Sergeev and Mike Del Balso. Horovod: fast and easy distributed deep learning in tensorflow. *CoRR*, abs/1802.05799, 2018.

[Sun *et al.*, 2019a] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, and Danxiang Zhu. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223, 2019.

[Sun *et al.*, 2019b] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 2.0: A continual pre-training framework for language understanding. *CoRR*, abs/1907.12412, 2019.

[Tian and Peng, 2019] Ke Tian and Zi Jun Peng. Sentence boundary detection in noisy texts from financial documents using deep attention model. In *FinNLP IJCAI 2019*, pages 88–92, Macao, China, August 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008, 2017.

[Yang *et al.*, 2018] Steve Yang, Jason Rosenfeld, and Jacques Makutonin. Financial aspect-based sentiment analysis using deep representations. *CoRR*, abs/1808.07931, 2018.