IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 26, 2020, accepted August 23, 2020, date of publication September 29, 2020, date of current version October 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027622

Pretraining Financial Text Encoder Enhanced by Lifelong Learning

ZHUANG LIU, DEGEN HUANG, (Member, IEEE), AND KAIYU HUANG
Dalian University of Technology, Dalian 116024, China

Corresponding author: Degen Huang (huangdg@dlut.edu.cn)

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 61672127 and Grant U1936109.

ABSTRACT As the number of financial literature grows rapidly, Financial text mining is becoming important increasingly. In recent years, extracting valuable information from financial documents, namely financial text mining, gained significant popularity within research communities. Although Deep Learning-based financial text mining has achieved remarkable progress recently, in financial fields it still suffers from issues of lack of task-specific labeled training data. To alleviate these issues, we present a pretraining financial text encoder, named F-BERT, a domain-specific language model pretrained on large-scale financial corpora. Different from original BERT, proposed F-BERT is trained continually on both general corpus and financial domain corpus, and four pretraining tasks can be pretrained through lifelong learning, which can enable our F-BERT to continually capture language knowledge and semantic information. The experimental results demonstrate that proposed F-BERT achieves strong results on several financial text mining tasks. Extensive experimental results show the effectiveness and robustness of F-BERT. The source code and pretrained models of F-BERT are available online.

INDEX TERMS BERT, fintech, lifelong learning, pretraining, transfer learning.

I. INTRODUCTION

In finance and economics, various financial text data are used to analyze, predict future financial 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, because the amount of labeled training data that can be used for financial text mining tasks is too small, the financial text mining model cannot use deep learning technology directly.

In the paper, we present F-BERT model addressing the issue by leveraging unsupervised Transfer Learning and Lifelong Learning. Word embedding, such as word2vec is a method of extracting knowledge from an unsupervised data and has invoked the rapid advancements in NLP taks. But, because of the special language used in the financial field, these simple word embedding approaches are not effective enough. pretrained Language Models (PLMs), such as original BERT [1], pretrained on a large-scale unsupervised data (such as Wikipedia) to improve contextualized representations more effectively. In financial text mining tasks, because of the huge differences in vocabulary and expression between general domain datasets and financial datasets, they still cannot be effectively applied to financial text. Furthermore, the pretraining of PLMs usually focuses on training the model through a few simple tasks. For example, BERT uses MaskLM and NSP as pretraining 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 seven pretraining tasks covering more knowledge,

The associate editor coordinating the review of this manuscript and approving it for publication was Zhanyu Ma.

and train F-BERT through lifelong learning on training data. Specifically, proposed F-BERT differs from standard PLMs pretraining methods. It constructs seven pretraining tasks, simultaneously trained on general corpora and financial domain corpora, to help F-BERT better capture language knowledge and semantic information.

In summary, our contributions include:

- Proposed F-BERT model differs from standard BERT in the training objectives. We construct four self-supervised pretraining tasks (subsection III-A), which can be learned through lifelong learning, capable of continually capturing language knowledge and semantic information in large-scale pretraining corpora.
- We conduct extensive experiments on three financial datasets. Experimental results show that proposed F-BERT achieves strong results on these financial text mining tasks, including financial Sentence Boundary Detection task and financial Sentiment Analysis task (subsection IV-D). Besides, our proposed F-BERT has also been successfully applied in our online system and now stably serve various scenarios (section VI).
- We implemented our F-BERT on Horovod framework using mixed precision training methodology (subsubsection IV-A2). We make the source code and pretrained models publicly available. We have proposed a pretraining financial text encoder with minimal task-specific architecture modifications, which is capable of being effectively applied to various financial text mining tasks.

II. BACKGROUND: BERT

As an advanced pretrained language model, BERT [1] has achieved great success in NLP tasks, built on bidirectional encoder representations transformers [2]. In BERT architecture, there are two successive phases: pretraining phase and fine-tuning phase.

- **Pretraining phase:** BERT constructs two self-supervised pretraining tasks, and are trained on English Wikipedia and BooksCorpus that are two large-scale unlabeled general domain corpus. Both pretraining objectives are Masked Language Model (MLM) task and Next Sentence Prediction (NSP) task, respectively. In MLM task, its goal is to predict masked input tokens randomly; In NSP task, its aims to predict whether two segments are consecutive.
- **Fine-tuning phase:** This process mainly uses labeled data from downstream tasks (such as sentiment analysis, text classification) to fine-tune all parameters initialized during pretraining.

BERT has two models, namely BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$. BERT$_{\text{LARGE}}$ has 24 layers (transformer blocks), 16 attention heads, and 340 million parameters; BERT$_{\text{BASE}}$ has 12 layers, 12 self-attention heads, and 110 million parameters.

A. NOTATION

Given an input word or subword sequence $X = (x_1, \ldots, x_m)$, where m is the length of the sequence. As a text encoder, for each token BERT gets a contextualized embeddings: $\mathsf{x}_1, \ldots, \mathsf{x}_m = encode(x_1, \ldots, x_m)$, Since the encoder is implemented via a deep transformer, it uses positional embeddings $\mathsf{p}_1, \ldots, \mathsf{p}_m$ to mark the absolute position for each token in the sequence. If X is packed by a sentence pair (X_1, X_2), we separate the two sentences with a special token [SEP]: `[CLS]`, $x_1, \ldots x_{X_1 len1}$, `[SEP]`, $x_1, \ldots x_{len2}$, `[EOS]`

B. ARCHITECTURE

BERT is built on the popular Transformer architecture [2], and we willnot review the Transformer in detail in this paper. In BERT, it uses bidirectional transformer with different layers (L), hidden dimension (H) and self-attention heads (A). Specifically, BERT has two parameter settings:

- BERT$_{\text{BASE}}$: 12 layers($L = 12$), 768 hidden dimensions ($H = 768$, FFN inner hidden size $= 3072$) and 12 attention heads ($A = 12$, attention head size $= 64$) with the total number of parameters, 110M;
- BERT$_{\text{LARGE}}$: 24 layers ($L = 24$), 1024 hidden dimensions ($H = 1024$, FFN inner hidden size $= 4096$) and 16 attention heads ($A = 16$, attention head size $= 64$) with the total number of parameters, 340M.

C. PRETRAINING

During the pretraining phase, BERT has two self-supervised pretraining tasks: masked language model (MLM) and next sentence prediction (NSP).

1) MASKED LANGUAGE MODEL (MLM) TASK

For MLM objective, BERT is trained via a cross-entropy loss to predict 15% of the input tokens, selected at random. To prevent the model from cheating, 80% of these selected tokens are replaced by a special `[MASK]` symbol in the input, 10% are replaced by a random token from the vocabulary, and 10% are left unchanged;

2) NEXT SENTENCE PREDICTION (NSP) TASK

NSP pre-training is a binary classification task, whose aims to predict whether two sentences are consecutive (whether they follow each other). This training task was first proposed in original BERT. BERT takes two text sequences X_1 *and* X_2 as input, and finally predicts whether X_2 is a direct continuation of X_1 in the original text.

D. FINE-TUNING

During fine-tuning phase, we first initialize BERT using the pretrained parameters, then further fine-tune it with supervised data from downstream target tasks (such as sentiment analysis, text classification). The post-trained model can be effectively adapted to downstream different tasks by fine-tuning with task-specific labeled data.

III. PROPOSED MODEL: F-BERT

As shown in Figure 1, proposed F-BERT is built based on the standard BERT architecture [1] based on the two-stage

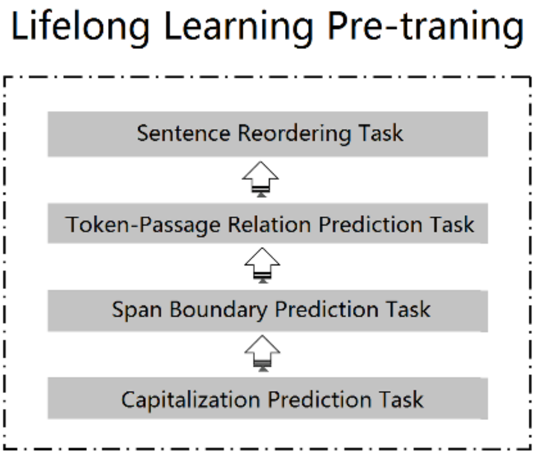
Lifelong Learning Pre-traning
Sentence Reordering Task
Token-Passage Relation Prediction Task
Span Boundary Prediction Task
Capitalization Prediction Task

FIGURE 1. Lifelong Learning pretraining Architecture. Proposed F-BERT are trained continually on both general corpus and financial domain corpus, and four pretraining tasks can be pretrained through lifelong learning, then it is fine-tuned using task-specific supervised data to adapt to downstream target tasks.

'*Pretraining*'-then-'*Fine-tuning*' pretraining language model approach, which recently become enormously popular in NLP. During pretraining phase, the F-BERT model differs from standard BERT architecture pretraining methods in that, instead of training with MaskLM and Next Sentence Prediction (NSP) pretraining objectives, it constructs a large variety of pretraining objectives to enable proposed F-BERT better capture language knowledge and semantic information. On top of that, F-BERT keeps updating the pretrained model through lifelong learning. Meantime, compared with traditional pretraining models, F-BERT is simultaneously trained on a general corpus and a financial domain corpus. During fine-tuning phase, we first initialize F-BERT using the pretrained parameters, and then further fine-tune it with downstream task-specific supervised data.

In the section, we will briefly introduce F-BERT in our proposed framework.

A. LIFELONG LEARNING

The choice of unsupervised pretraining objective plays an important role in applying to pretraining stage by continuously gains general knowledge. We will modify and combine multiple common unsupervised pretraining tasks [1], [3]–[9] fit our framework. Specifically, in this layer, we construct seven unsupervised pretraining tasks to learn different level knowledge from the training corpora. The seven pretraining tasks can be pretrained through lifelong learning. As shown in Figure 1, Pretraining tasks: Capitalization Prediction pretraining task, Span Replace Prediction pretraining task, Token-Passage Prediction pretraining task, and Sentence Reordering pretraining task. Next, we introduce these four self-supervised pretraining tasks in detail.

1) CAPITALIZATION PREDICTION (CAP) 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 [9], we do so by introducing a capitalization words prediction objective that involves predicting whether the word is capitalized or not.

2) SPAN REPLACE PREDICTION (SRP) TASK

Inspired by spanBERT [3] and T5 [8], we constructed a self-supervised pretraining task. Its aims to sample and drop out 20% in the input text randomly. We use an 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, instead of replacing each token with a masked token. Last, we predict the replaced (masked) span by the tokens that are observed at the span boundary.

3) TOKEN-PASSAGE PREDICTION (TPP) TASK

We constructed another pretraining task, whose aims to identify the key tokens of a passage that appears in the segment, which can enable F-BERT 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 [9], we also do so by introducing a token-passage prediction pretraining objective to predict whether the token appears in segments of the original passage.

4) SENTENCE REORDERING (SER) TASK

We constructed a sentence reordering pretraining objective which is used e.g. in ERNIE2 [9] and T5 [8], 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 pretrain F-BERT to reorganize the permuted segments which are modeled with a multi-class ($\sum n!$) classification task.

IV. EXPERIMENTS

A. PRETRAINING F-BERT

1) PRETRAINING DATA

BERT just uses general domain corpus (English Wikipedia and BooksCorpus, totaling 3.3 billion words), for pretraining. In order to better pretrain our proposed F-BERT model, we further collected a large variety of financial data automatically crawled from financial websites, such as financial news, financial articles and financial reports. Specifically, we consider four English-language financial corpora of varying domains and sizes. The statistics of our pretraining corpora for F-BERT are reported in Table 1.

- English Wikipedia[1] and BooksCorpus (Zhu *et al.*, 2015), which are the original pretraining data that are used to train original BERT (totaling 13GB plain text, 3.31B words). Both are general domain corpora. Following

[1] https://dumps.wikimedia.org/enwiki/

TABLE 1. List of pretraining corpora used for F-BERT.

Corpus	size(G)	# of words(B)	Domain
English Wikipedia	9	2.50	General
BooksCorpus	4	0.81	General
Financial News	26	6.93	Financial
Annual Financial Reports	7	2.15	Financial

Notes: We pretrain F-BERT with four English corpora of varying sizes and domains, totaling over 46GB plain text (12.39 billion words): general domain corpora (English Wikipedia + BooksCorpus, totaling 3.31 billion words) and financial domain corpora (Financial News + Annual Financial Reports, totaling 9.08 billion words).

BERT [10], we use both corpora as a part of our training data;

- Financial News (totaling 26GB plain text, 6.93B words), which we collect from the CommonCrawl News dataset[2] between May 2013 and October 2019, containing 13 million financial news, together with automatically crawled financial articles from FINWEB[3] (removing markup, removing non-textual content and filtering out redundant data, 15G financial news after filtering and 11G financial articles after filtering);
- Annual Financial Reports (totaling 7GB plain text, 2.15B words), which we crawl annual reports from SEC website[4] between May 2013 and October 2019. In business and finance company reports, company reports (especially annual financial reports) are very important text data. From the financial reports, we can get a brief overview of the financial and business status of a company. These dataset are publicly available and can be accesses from web site.

2) PRETRAINING SETTINGS

Our proposed F-BERT (including F-BERT$_{\text{BASE}}$ and F-BERT$_{\text{LARGE}}$) has the same settings of transformer and pretraining parameters as original BERT. Proposed four pretraining tasks also have great requirements on computing power. During 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 [10], using parallelized training approach. Different from TensorFlow [11], we implemented synchronous and data-parallel distributed training strategy with Horovod library, to reduce the training time. Horovod is an open source library[5] developed by Uber [9], based on baidu-allreduce[6] and [12]. The Horovod architecture utilizes a distributed optimizer strategy that is an optimizer wrapping *tf.Optimizer* essentially, 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, the performance loss of architecture is much smaller than that based on standard distributed TensorFlow [11], and the training speed can reach much three or four times. Compared to the distributed Tensorflow framework, the Horovod architecture can guarantee a very stable acceleration ratio on the scale of hundreds of GPU cards, and has better scalability.

Mixed Precision Training: Inspired by [13], who proposed mixed precision training methodology to train deep neural networks, which can reduce the memory consumption and spent-time in memory and arithmetic operations of DNN, we also utility mixed precision training method for self-supervised pretraining. Specifically, we train our F-BERT model by half-precision format FP16 that is used for storage and arithmetic. We store activations, gradients and weights in FP16, and update the copy of weights in FP32. And we maintain a single-precision copy of weights, after each optimizer step which accumulates the gradients. We use loss-scaling to preserve gradient values with small magnitudes and use FP16 arithmetic that accumulates into single-precision outputs, converted to FP16 before storing to memory.

B. FINE-TUNING F-BERT

Typically, BERT [1] has two successive steps, one during the pretraining stage and one during the fine-tuning stage. It firsts conducts unsupervised pretraining on the large corpus during the pretraining phase, and then conducts supervised fine-tuning on downstream NLP tasks during the fine-tuning phase. Following BERT, we pretrain F-BERT from scratch on four unsupervised corpus (subsection IV-A), and fine-tune it to various downstream supervised financial text mining tasks. Next, we briefly describe three supervised financial tasks.

1) 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, which is a dataset that was created for IJCAI19 FinNLP challenge. In FinSBD-2019, they provide training data with boundary labels (*beginning* boundary vs. *ending* boundary) for each token. Appendix subsection A has more details.

2) FINANCIAL SENTIMENT ANALYSIS (SA)

Financial SA is one of the most fundamental financial text mining tasks. Among financial text mining tasks, financial sentiment analysis (SA) is one of the most basic tasks. In financial SA task, its goal is to detect the target aspects mentioned in the financial text for the text data in the given financial field, and predict the sentiment score of each target. Sentiment scores ranged in [-1, 1]. In this work, financial SA datasets used are Financial PhraseBank [14] and FiQA Task 1 [15]. FiQA is a dataset created for WWW18 financial opinion mining challenge. In this work, we use the dataset

[2] http://commoncrawl.org/

[3] https://www.finweb.com/

[4] http://www.sec.gov/edgar.shtml

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

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

of FiQA Task 1, namely "Aspect-based financial sentiment analysis". Appendix subsection B has more details.

C. EXPERIMENTAL SETUP

Experimental settings of detailed fine-tuning on financial sentence boundary detection and financial sentiment analysis tasks are included in Table 10 for reference.

D. EXPERIMENTAL RESULTS

1) FINANCIAL SENTENCE BOUNDARY DETECTION (SBD)

Table 2 shows our results on financial SBD. The *ending* (**ES**) and *beginning* (**BS**) tokens of the sentence are separately evaluated, and 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 F-BERT pretrained on both general domain dataset and financial domain dataset is highly effective. Both F-BERT$_{\text{BASE}}$ and F-BERT$_{\text{LARGE}}$ achieved strong results. F-BERT$_{\text{LARGE}}$ outperformed BERT-SBD [16] by 0.03, 0.01, 0.026 in **ES**, **BS**, **MEAN**, respectively.

TABLE 2. Experimental results on test set for FinSBD English task.

Model	ES	BS	MEAN
Deep-Att [17]	0.83	0.91	0.875
BERT-SBD [16]	0.88	0.89	0.885
Ours: F-BERT$_{\text{BASE}}$	0.87	0.88	0.875
Ours: F-BERT$_{\text{LARGE}}$	**0.92**	**0.91**	**0.915**

Notes: BERT-SBD is AIG1 system in Leaderboard and obtains the state-of-the-art performance.

2) 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. As shown in experiments, it is apparent that our F-BERT$_{\text{BASE}}$ and F-BERT$_{\text{LARGE}}$ consistently significantly outperforms all baseline models on PhraseBank sentiment dataset and FiQA sentiment dataset, achieving comparable performance. Overall, these results highlight the importance of the financial domain-specific corpora pretrained design.

TABLE 3. Performance on the test set for the Financial PhraseBank sentiment dataset.

Model	Accuracy	F1
LPS [14]	0.71	0.71
FB-SA [18]	0.86	0.84
FINBERT [18]	0.87	-
Ours: F-BERT$_{\text{BASE}}$	0.89	0.87
Ours: F-BERT$_{\text{LARGE}}$	**0.93**	**0.92**

Notes: Bold face indicates best result in the corresponding metric.

V. ABLATION STUDY AND ANALYSES

In this section, we conduct ablation studies and further compare proposed F-BERT with existing PLMs models and present the results along with experimental details.

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

	headline		*post*	
Model	**MSE**	**R^2**	**MSE**	**R^2**
IIT-Dehi $^\natural$	0.20	0.18	0.10	0.08
Inf-UFG $^\natural$	0.21	0.17	0.10	0.16
FB-SA [18]	0.07	0.55	-	-
Ours: F-BERT$_{\text{BASE}}$	0.17	0.50	0.18	0.19
Ours: F-BERT$_{\text{LARGE}}$	**0.26**	**0.59**	**0.28**	**0.24**

Notes: FiQA sentiment dataset includes two discourses: *financial microblogs*, *financial news headlines*. The evaluation of FiQA sentiment dataset mainly includes: aspect-sentiment attachment, sentiment classification, and aspect classification. In aspect classification, it is evaluated on F1-score, recall and precision, respectively; In sentiment prediction, it is evaluated on R Squared(R^2) and MSE.

A. EFFECT OF PRETRAINING ON THE PERFORMANCE

We further evaluate the effect of pretraining on the performance of our proposed F-BERT. Specifically, we compare three models: **i)** No further pretraining model, namely Vanilla BERT; **ii)** Further pretraining on financial FiQA Aspect category classification dataset (denoted by BERT$_{task}$, F-BERT$_{task}$, respectively). Models are evaluated with *Accuracy*, *Precision* and *Recall* scores on the test set. The results are shown Table 5. As shown, our F-BERT$_{task}$ obtains better performance than that of all other models. Specially, although BERT$_{task}$ is further pretrained on the financial FiQA Aspect category classification training set, F-BERT achieves best performance, outperforming another models, 0.10% higher than Vanilla BERT and 0.03% higher than BERT$_{task}$, in terms of *Accuracy* score. Overall, this experimental result shows the strength of proposed F-BERT, and also shows that F-BERT learned domain-specific knowledge from a large number of financial domain corpora during the pretraining phase.

TABLE 5. Experimental results on the test set for the financial classification dataset (FiQA Aspect category classification task).

Model	Accuracy	Precision	Recall
Vanilla BERT	0.78	0.33	0.30
BERT$_{task}$	0.85	0.51	0.45
F-BERT$_{task}$	**0.88**	**0.56**	**0.51**

Notes: Bold face indicates best result in each metric.

B. TRAINING DATASET CONTRIBUTION

We train different proposed F-BERT on different datasets, separately. Table 6 reports the performance on different training datasets across PhraseBank task. As clearly shown in Table 6, F-BERT trained on financial corpora achieves significant improvements, which indicates that combining additional financial communication corpus can better to capture language knowledge and semantic information, benefit from additional pretraining. In these two financial corpora, although the Analyst Reports data set has fewer words, it performed better on downstream financial text mining task (i.e., PhraseBank task).

TABLE 6. Experimental results of pretraining on different training dataset.

corpus	Accuracy	F1
Wikipedia+BooksCorpus	0.82	0.81
Financial News	0.84	0.82
Annual Financial Reports	**0.85**	**0.85**

Notes: Here reports the performance on different training datasets on downstream PhraseBank task. Bold face indicates best result in the corresponding metric.

C. PRETRAINING WITH SMALL TRAINING DATA

Deep learning models require a lot of supervised pretraining data. However, existing methods suffers from the lack of a large number of labeled financial text data in the financial text mining field, in many applications in financial fields, large training corpora may not be available. To further demonstrate the advantages of our F-BERT, we conducted another experiment, which was based on a simulated small corpus to pretrain BERT and F-BERT respectively. Specifically, we randomly select 1/5 size of the entire financial data set as a simulated small corpus. Then, we pretrain both models based on this corpus, and use the previous experiment to test the same tasks. Table 2 shows the performance of proposed F-BERT and original BERT based on financial sentence boundary detection task and financial sentiment analysis task, respectively. As shown in Table 7, proposed F-BERT model outperform BERT on all both financial text mining tasks constantly. Considering that both models are trained on small corpus, the experimental results are highly encouraging, which shows that proposed F-BERT can provide stable and clear enhancement when trained on financial corpora of different sizes. Overall, this experiment simulates that our F-BERT model can better encode financial text in the case of limited data, proving that F-BERT still has great advantages under the small training data scenarios in financial domain.

TABLE 7. The performance of BERT and F-BERT on three financial tasks (SBD and SA) when they are trained on a small corpus.

	SBD	SA	
Model	MEAN	Accuracy	F1
BERT	0.86	0.84	0.82
F-BERT	**0.90**	**0.85**	**0.84**

Notes: During ablation, we use the PhraseBank sentiment dataset because it and FiQA task1 dataset are similar sentiment analysis datasets.

D. SIZE OF THE CORPUS FOR PRETRAINING

We further evaluate the performance of our F-BERT on different size of the dataset. The performance of all F-BERT models on different size are reported in Table 8. We first control the training data, and then we can observe that F-BERT is better than the original result. We then combine the data with the several additional datasets that are described in subsection IV-A. We further train F-BERT model separately by combining the training data and the number of training steps. We pretrained over 46GB of text totally. As shown in Table 8, as data size and diversity change, the performance of the model continues to further improve, which validates during pretraining stage the importance of data size and diversity. In the meantime, we also pretrain model with further longer (more training steps). We pretrain F-BERT with the number of pretraining steps, from 200K to 500K, to 800k, further to 1M last. As clearly shown in Table 8, F-BERT achieves significant improvements on downstream three tasks performance. Furthermore, we also observe that even the most trained F-BERT (with most training dataset and longest training time) may not obtain the better performance.

VI. ON-LINE EVALUATION: STOCK PRICE PREDICTION

We also test the effectiveness of F-BERT in our practical scenarios, Stock Price Prediction. In the stock market, stock price prediction plays a very important role in stock investment. In particular, short-term predictions using financial news articles have been promising in recent years.

We first downloaded historical S&P 500 component stocks data from the Yahoo Finance. We selected the data over a period of from 01/01/2018 to 05/25/2018. Due to the relationship between stock prices and financial news articles, we collected news articles from the financial domain within the same time period as stock data. We use the first 80% of the entire dataset as training dataset, and the remaining 20% for test dataset. Similarly, to predict the next value, we will shift the rolling window and add the next new value to the last, and so forth. Schematic diagram of rolling window is shown in Figure 2.

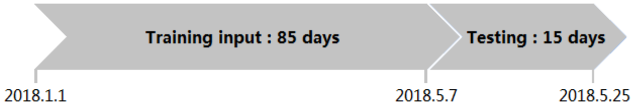

FIGURE 2. The diagram of rolling window.

Following [19] and [20], we also use the mean prediction accuracy (*MPA*) to evaluate the methods, MPA $= 1 - \frac{1}{N}\sum_{j=1}^{N}\frac{|P_{t,j}-\hat{P}_{t,j}|}{P_{t,j}}$, where $P_{t,j}$ is the real stock price of the j-th stock on the t-th day, N is the number of stocks, and $\hat{P}_{t,j}$ is the prediction result.

We reimplement BiLSTM, BERT for comparison. Figure 3 shows the performance of all three methods on S&P 500 index price datasets. We compare our F-BERT with

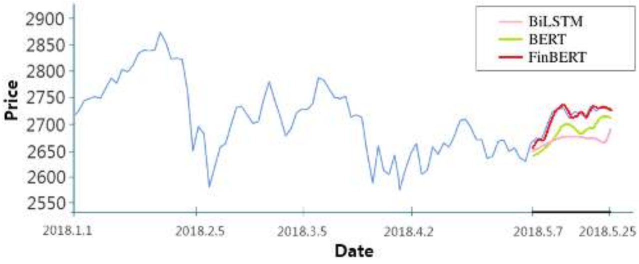

FIGURE 3. Prediction result of models based on price.

TABLE 8. The performance of F-BERT on different size of the corpus.

				Financial SBD	Financial SA	
	data size	**batch size**	**steps**	**MEAN**	**Accuracy**	**F1**
F-BERT$_{\text{LARGE}}$						
w/ Wiki+BookS	13G	128	200K	0.89	0.88	0.87
+ additional data	46G	128	200K	0.90	0.89	0.88
+ pretrain longer	46G	128	500K	0.92	0.91	0.89
+ pretrain a little longer	46G	128	800K	**0.93**	0.91	0.90
+ pretrain even longer	46G	128	1M	0.92	**0.91**	**0.90**

Notes: We pretrain F-BERT with more data (training datasize: 13GB → 46GB) and for longer (training steps: 200K → 500K → 800K → 1M). For the Financial SA task, we also use the PhraseBank sentiment dataset during ablation test.

BiLSTM, BERT. As clearly shown in Figure 3, the prediction results of our F-BERT and real index prices almost coincide, F-BERT is far ahead. Furthermore, we further evaluate the *Mean MPA* results for the prediction prices. Table 9 lists the *Mean MPA* results, which show F-BERT notably outperforms both competitive baseline models, demonstrating proposed F-BERT model can make the prediction result more accuracy and effectiveness.

TABLE 9. The accuracy comparison (*Mean MPA*) of all models.

Model	**Mean MPA**
BiLSTM	0.97167
BERT	0.97840
F-BERT	**0.99452**

VII. RELATED WORK

A. UNSUPERVISED TRANSFER LEARNING FOR LANGUAGE REPRESENTATION

pretrained 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 and ELMo [21] is an approach of extracting knowledge from a large-scale unlabeled data and has become one of the major advances in NLP. References [3]–[5], [16], [17], [22] presented models for financial domain tasks. However, because of the differences in vocabulary and expression between the general text and the financial domain text and the special language that is used in the financial field, these PLMs models are not effective enough.

B. LIFELONG LEARNING

Lifelong learning, sequential learning and multi-task learning are closely related to each other. Lifelong learning mainly focuses on using new data to train and adjust the model, without forgetting the knowledge that has been learned before [23], [24]. For Multi-task learning, its aims to learn more different tasks simultaneously [25]–[29], for example, incremental learning [30]–[32],in which different tasks usually have different classes. Also, for different tasks, they usually only access a part of old data or new data. From the perspective of training objectives, multi-task learning tackles task-specific problems: path traversal between multi tasks, parameter sharing in multi tasks, and attention for switching different tasks. In different experimental environments, their contribution solves the challenges, such as how to allocate data in different tasks, how to sort tasks and how to evaluate performance among a series of different tasks. Compared with multi-task learning, our lifelong learning [33] is a continual learning paradigm designed to train models with multiple tasks in order to remember the previously learned tasks when learning new tasks. The point is to continually train and adjust the model without forgetting the knowledge that has been learned. In proposed F-BERT, through lifelong learning, due to the knowledge acquired in previous training, our model is able to perform on new tasks well.

VIII. CONCLUSION

In the paper, we presented F-BERT model that is a pretrained language model for financial text mining. By constructing four pretraining tasks covering more knowledge, we continually trained F-BERT on general corpora and financial domain corpora through lifelong learning, which enabled F-BERT model effectively to capture language knowledge and semantic information. Also, we implemented our F-BERT on Horovod framework using mixed precision training methodology. Extensive experimental results demonstrated the effectiveness and robustness of F-BERT.

APPENDIX A
FINANCIAL DATASETS

A. FINANCIAL SENTENCE BOUNDARY DETECTION DATASET

In written language, sentences are the basic unit. In many NLP applications (such as POS tagging, information extraction), detecting the beginning of sentences, end of sentences, and sentence boundary detection is often a first step task. So far, for the sentence boundary detection task where automatically extracted from financial documents (PDF files generally), there is still no effective solution to solve the problem. Financial SBD dataset used in this paper is FinSBD Shared Task, which is a dataset that was created for IJCAI19 conference FinNLP challenge. 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 for FinSBD-2019 data. It's available online.[7] In this task, its aims to well extract segmented sentences from PDF documents about investment funds and financial prospectuses to detect the beginning and ending boundaries.

B. FINANCIAL SENTIMENT ANALYSIS DATASET

1) PHRASEBANK DATASET [14]

were annotated by people with background in finance and business. It has 4,845 english sentences that are randomly selected from financial articles and news found on LexisNexis database. The dataset publically available online.[8]

2) FiQA SENTIMENT DATASET [15]

was released in WWW18. It is created for financial text opinion mining challenge. FiQA sentiment dataset has two discourses: financial microblogs and financial news headlines, with manually annotated aspects and sentiment scores. In financial news headlines dataset, it has 436 samples for the training dataset and 93 samples for the dataset (529 labeled samples totally); In financial microblogs dataset, it has 675 samples for the training dataset and 99 samples for the test dataset (774 labeled samples totally). The dataset publically available online.[9] The evaluation of FiQA sentiment dataset mainly includes: aspect-sentiment attachment, sentiment classification, and aspect classification. In aspect classification, it is evaluated on F1-score, recall and precision, respectively; In sentiment prediction, it is evaluated on R Squared(R^2) and MSE.

APPENDIX B
FINE-TUNING HYPERPARAMETER

See Table 10.

TABLE 10. Fine-tuning Hyperparameter.

Hyperparam	FinBSD	PhraseBank	FiQA sentiment
Learning Rate	{2e-5,3e-5}	2e-5	{2e-5,3e-5}
Batch Size	32	{32,48}	48
Warmup ratio	0.08	0.1	0.1

Notes: Hyperparameters for fine-tuning F-BERT$_{\mathrm{LARGE}}$ on FinBSD, PhraseBank, FiQA sentiment datasets. We used a linear learning rate decay, and train the model for a max of 10 epochs.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their helpful comments and suggestions to improve the quality of the article.

[7] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/

[8] https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10/

[9] https://sites.google.com/view/fiqa/home/

REFERENCES

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423/

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need

[3] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: Improving pre-training by representing and predicting spans," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 64–77, Jul. 2020. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/1853

[4] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon, "Unified language model pre-training for natural language understanding and generation," in *Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst.*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds. Vancouver, BC, Canada, Dec. 2019, pp. 13042–13054. [Online]. Available: https://arxiv.org/pdf/1905.03197.pdf

[5] X. Liu, P. He, W. Chen, and J. Gao, "Multi-task deep neural networks for natural language understanding," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4487–4496, doi: 10.18653/v1/p19-1441.

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Represent., ICLR*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*. [Online]. Available: http://arxiv.org/abs/1907.11692

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified Text-to-Text transformer," 2019, *arXiv:1910.10683*. [Online]. Available: http://arxiv.org/abs/1910.10683

[9] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," 2019, *arXiv:1907.12412*. [Online]. Available: http://arxiv.org/abs/1907.12412

[10] A. Sergeev and M. Del Balso, "Horovod: Fast and easy distributed deep learning in TensorFlow," 2018, *arXiv:1802.05799*. [Online]. Available: http://arxiv.org/abs/1802.05799

[11] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," 2016, *arXiv:1603.04467*. [Online]. Available: http://arxiv.org/abs/1603.04467

[12] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," 2017, *arXiv:1706.02677*. [Online]. Available: http://arxiv.org/abs/1706.02677

[13] P. Micikevicius, S. Narang, J. Alben, G. F. Diamos, E. Elsen, D. García, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," in *Proc. 6th Int. Conf. Learn. Represent.s, ICLR*, 2019, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=r1gs9JgRZ

[14] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala, "Good debt or bad debt: Detecting semantic orientations in economic texts," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, pp. 782–796, Apr. 2014, doi: 10.1002/asi.23062.

[15] P.-A. Champin, F. Gandon, L. Médini, P. Chairs, M. Lalmas, P. G. Ipeirotis, in *Proc. WWW*, M. Maia, S. Handschuh, A. Freitas, B. Davis, R. Mcdermott, and M. Zarrouk, Eds., 2018, doi: 10.1145/3184558.

[16] J. Du, Y. Huang, and K. Moilanen, "Sentence boundary detection through sequence labelling and BERT fine-tuning," in *Proc. FinNLP IJCAI*, Macao, China, Aug. 2019, pp. 81–87. [Online]. Available: https://www.aclweb.org/anthology/W19-5513

[17] K. Tian and Z. J. Peng, "Sentence boundary detection in noisy texts from financial documents using deep attention model," in *Proc. FinNLP IJCAI*, Macao, China, Aug. 2019, pp. 88–92. [Online]. Available: https://www.aclweb.org/anthology/W19-5514

[18] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*. [Online]. Available: http://arxiv.org/abs/1908.10063

[19] X. Li, Y. Li, H. Yang, L. Yang, and X.-Y. Liu, ''DP-LSTM: Differential privacy-inspired LSTM for stock prediction using financial news,'' 2019, *arXiv:1912.10806*. [Online]. Available: http://arxiv.org/abs/1912.10806

[20] Y. Li and Y. Pan, ''A novel ensemble deep learning model for stock prediction based on stock prices and news,'' 2020, *arXiv:2007.12620*. [Online]. Available: http://arxiv.org/abs/2007.12620

[21] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, ''Deep contextualized word representations,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., (Long Papers)*, vol. 1, 2018, pp. 2227–2237, doi: 10.18653/v1/n18-1202.

[22] Z. Liu, K. Huang, D. Huang, and J. Zhao, ''Semantics-reinforced networks for question generation,'' in *Proc. ECAI 24th Eur. Conf. Artif. Intell.* in Frontiers in Artificial Intelligence and Applications, vol. 325, G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, and J. Lang, Eds. Santiago de Compostela, Spain: IOS Press, Aug. 2020, pp. 2078–2084, doi: 10.3233/FAIA200330.

[23] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, ''Continual lifelong learning with neural networks: A review,'' *Neural Netw.*, vol. 113, pp. 54–71, May 2019, doi: 10.1016/j.neunet.2019.01.012.

[24] M. Kayser, R. D. Soberanis-Mukul, A.-M. Zvereva, P. Klare, N. Navab, and S. Albarqouni, ''Understanding the effects of artifacts on automated polyp detection and incorporating that knowledge via learning without forgetting,'' 2020, *arXiv:2002.02883*. [Online]. Available: http://arxiv.org/abs/2002.02883

[25] Z. Liu, D. Huang, K. Huang, and J. Zhang, ''DIM reader: Dual interaction model for machine comprehension,'' in *Proc. Chin. Comput. Linguistics Natural Lang. Process. Based Naturally Annotated Big Data 16th China Nat. Conf., CCL 5th Int. Symp., NLP-NABD*, Nanjing, China, Oct. 2017, pp. 387–397, doi: 10.1007/978-3-319-69005-6_32.

[26] N. Jin, J. Wu, X. Ma, K. Yan, and Y. Mo, ''Multi-task learning model based on multi-scale CNN and LSTM for sentiment classification,'' *IEEE Access*, vol. 8, pp. 77060–77072, 2020, doi: 10.1109/ACCESS.2020.2989428.

[27] W. Zhao, H. Gao, S. Chen, and N. Wang, ''Generative multi-task learning for text classification,'' *IEEE Access*, vol. 8, pp. 86380–86387, 2020, doi: 10.1109/ACCESS.2020.2991337.

[28] P. Henderson, W.-D. Chang, F. Shkurti, J. Hansen, D. Meger, and G. Dudek, ''Benchmark environments for multitask learning in continuous domains,'' 2017, *arXiv:1708.04352*. [Online]. Available: http://arxiv.org/abs/1708.04352

[29] Z. Liu, K. Xiao, B. Jin, K. Huang, D. Huang, and Y. Zhang, ''Unified generative adversarial networks for multiple-choice oriented machine comprehension,'' *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 25:1–25:20, 2020, doi: 10.1145/3372120.

[30] Y. Xiang, Y. Miao, J. Chen, and Q. Xuan, ''Efficient incremental learning using dynamic correction vector,'' *IEEE Access*, vol. 8, pp. 23090–23099, 2020, doi: 10.1109/ACCESS.2019.2963461.

[31] F. Feng, R. H. M. Chan, X. Shi, Y. Zhang, and Q. She, ''Challenges in task incremental learning for assistive robotics,'' *IEEE Access*, vol. 8, pp. 3434–3441, 2020, doi: 10.1109/ACCESS.2019.2955480.

[32] L. Guo, G. Xie, X. Xu, and J. Ren, ''Exemplar-supported representation for effective class-incremental learning,'' *IEEE Access*, vol. 8, pp. 51276–51284, 2020, doi: 10.1109/ACCESS.2020.2980386.

[33] B. Liu, ''Lifelong machine learning: A paradigm for continuous learning,'' *Frontiers Comput. Sci.*, vol. 11, no. 3, pp. 359–361, Jun. 2017, doi: 10.1007/s11704-016-6903-6.