FinancialBERT – A Pretrained Language Model for Financial Text Mining

Preprint · February 2022

DOI: 10.13140/RG.2.2.34032.12803

1 author:

Ahmed Rachid Hazourli
VMware
4 PUBLICATIONS **32** CITATIONS

SEE PROFILE

FinancialBERT - A Pretrained Language Model for Financial Text Mining

Ahmed Rachid Hazourli
ahmedrachidhazourli@yahoo.fr

Abstract

Textual data in the financial domain is becoming increasingly important as the number of financial documents rapidly grows. With the progress in natural language processing (NLP), extracting valuable information has gained popularity among researchers, deep learning has boosted the development of effective financial text mining models and made significant breakthroughs in various Natural Language Processing tasks.

State-of-the-art models such as BERT (Devlin et al., 2019) model developed by Google pre-trained on a large scale of unlabeled texts from Wikipedia, has shown its effectiveness by achieving good results on general domain data. However, these models are not effective enough on finance-specific language and semantics, limiting the accuracy that financial data scientists can expect from their NLP models. In this paper, we introduce FinancialBERT, a domain-specific language representation model pre-trained on large-scale financial corpora that can enhance NLP research in the financial sector. With almost the same architecture across tasks, FinancialBERT largely outperforms BERT and other state-of-the-art models in Sentiment Analysis task when pre-trained on financial corpora.

Our pre-trained model FinancialBERT is freely available at: `https://huggingface.co/ahmedrachid/FinancialBERT`.

Keywords: Natural Language Processing, BERT, Language Model, Pretrained Model, Sentiment Analysis, Financial Language Modelling

1. Introduction

In recent years, Deep Neural Networks have revolutionized the development of intelligent systems in many fields especially in Natural Language Processing using state-of-the-art neural networks architectures that significantly improved many NLP tasks. These results are achieved thanks to unsupervised pre-training of language models on large text collections based on deep learning techniques such as Long Short-Term Memory (LSTM), Transformers...

As the amount of textual content generated in the financial domain is growing at an exponential rate, natural language processing is becoming a strategic tool for financial analysis. Such textual data is a valuable source of knowledge, however, applying state-of-the-art models to financial text mining has limitations. Firstly, word embeddings or representations such as ELMO (Peters et al., 2018), Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019) are trained on general domain texts, it is then hard to estimate their performance on financial datasets. Also, the word distributions are different between general and financial domains.

BERT achieves great results on various NLP tasks, adapting it for the financial domain could potentially achieve high performance by building a model capable of understanding financial language, producing more accurate word embeddings and ultimately can improve the performance of downstream tasks such as text classification, topic modelling, automatic summarization and sentiment analysis.

2. Related Work

Unlike traditional word embedding where a word is represented as a single vector representation, language models such as BERT (Devlin et al., 2019), ELMO (Peters et al., 2018) return contextualized embeddings for each word token which can be fed into downstream tasks. These models are trained on general domain corpora and are easy to fine tune for downstream tasks.

The goal of this work is to test the hypothesized advantages of using fine-tuning pre-trained language models. Thus, we pre-train FinancialBERT, a finance domain-specific BERT model on a large financial communication corpora including financial news, corporate reports and earning calls.

The main contributions of this paper are the following:

1. Introduce and release FinancialBERT, a new finance domain-specific BERT-base model. We achieve state-of-the-art results on Financial PhraseBank dataset.
2. Perform extensive experimentation to investigate the performance of fine-tuning versus task-specific architectures atop frozen embeddings, and the effect of having an in-domain vocabulary. Then, evaluate on a financial corpus for sentiment analysis to show the effectiveness of our approach.
3. Most importantly, we make publicly available both the pre-trained FinancialBERT and our fine-tuned Sentiment Analysis model . We expect these resources to boost NLP research and applications for finance, since fine-tuning pre-trained Transformer-based language models for particular downstream tasks is the state-of-the-art.

3. Methods

In this section, we will present our FinancialBERT implementation that has the same structure as BERT, after giving a brief background on relevant neural architectures. Then, we describe in detail the pre-training and fine-tuning process of FinancialBERT.

3.1. Background

With the advent of deep learning and its application in NLP, researchers began applying Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) for text classification. The current state-of-the-art in text classification typically involves a purely attentional architecture, the Transformer architecture (Vaswani et al., 2017).

3.2. BERT

Bidirectional Encoder Representations from Transformers (BERT) model architecture (Devlin et al., 2019) is based on a multilayer bidirectional Transformer. It is pre-trained on large textual corpora in an unsupervised way. The attention mechanism (Vaswani et al., 2017) of the Transformer allows obtaining contextual word embeddings, BERT (Devlin et al., 2019) was trained on two parallel tasks:

1. *Masked Language Modeling (MLM):* instead of predicting the next word given previous ones, BERT (Devlin et al., 2019) masks a randomly selected 15% of all tokens and learns to predict them, and hence can be used for learning bidirectional representations. Thus, it learns to produce token-level embeddings.
2. *Next Sentence Prediction (NSP):* the model predicts whether or not these two actually follow each other. It learns whether the second sentence is the next one or not from the embeddings of the special token CLS (class) and produces sentence-level embeddings.

3.3. Model Architecture

The original English BERT was pre-trained on two generic corpora, English Wikipedia and Books Corpus with a total of 3,5B words. BERT (Devlin et al., 2019) has two versions:

1. *BERT-BASE:* with 12 layers of stacked Transformers, each of 768 hidden units, 12 attention heads, 110M parameters (L=12, H=768, A=12, Total Parameters=110M)
2. *BERT-LARGE:* with 24 layers, each of 1024 hidden units, 16 attention heads, 340M parameters (L=24, H=1024, A=16, Total Parameters=340M).

Both architectures were trained on "cased" texts that keep character casing or "uncased" that convert all text to lowercase.

4. Pre-training FinancialBERT

In this section, we first describe our financial corpora, the details of the BERT training procedure, and finally the specific task we examine.

4.1. Financial Corpora

As a general purpose language representation model, BERT was pre-trained on English Wikipedia and BooksCorpus.

However, financial domain texts contain a considerable number of new domain-specific terms. In this work, we pre-train FinancialBERT on a large corpora of representative financial texts:

1. *TRC2-financial*[1]*:* Thomson Reuters Text Research Collection (TRC2) corpus comprises 1,800,370 news stories that were published by Reuters covering the period between 2008 and 2010.
2. *Bloomberg News*[2]*:* 400,000 financial articles published by Bloomberg between 2006 and 2013.
3. *Corporate Reports:*[3] a rich source of information as they often disclose new important statements and provide a comprehensive overview of the company's business and financial condition. These documents are available on EDGAR database as the Securities Exchange Commission (SEC) mandates all publicly traded companies to file annual reports (10K) and quarterly reports (10Q). We retrieved 154,354 documents of the 10-K reports from 1996 to 2015 and 37,646 quarterly reports 10-Q. Then, we filtered on sections and decided to use only "Risk Factors" (Section 1A) and "Management Discussion and Analysis of Financial Conditions and Results of Operations" (Section 7).
4. *Earnings Call Transcripts:* we obtained 42,156 earnings call transcripts. They are teleconferences, or webcasts between the management of a public company, analysts, investors, and the media to discuss the company's financial results during a given reporting period, such as a quarter or a fiscal year. An earnings call is usually preceded by an earnings report, which contains summary information on financial performance for the period.

Corpus	Number of words	Domain
English Wikipedia	2.5B	General
BooksCorpus	0.8B	General
TRC2-financial	0.29B	Financial
Bloomberg News	0.2B	Financial
Corporate Reports	2.2B	Financial
Earnings Call Transcripts	0.7B	Financial

Table 1: Size of text corpora.

The text corpora used for pre-training of FinancialBERT has a total size of 3.39 billion tokens and are listed above in Table 1. The description of the textual corpora are listed in Table 2.
For better performance, we initialized FinancialBERT with the pre-trained BERT [4] model provided by Devlin et al.

[1] https://trec.nist.gov/data/reuters/reuters.html

[2] https://www.bloomberg.com/

[3] https://www.sec.gov/edgar.shtml

[4] The pre-trained weights are made public by creators of BERT. The code and weights can be found here: https://github.com/google-research/bert

(2019) that was trained on Wikipedia + BooksCorpus corpora with a total of 3.3 billion tokens.

Model	Corpus
BERT	Wikipedia + BooksCorpus
FinancialBERT	TRC2 + Bloomberg News + Corporate Reports + Earnings Call Transcripts

Table 2: Description of pre-training text corpora.

4.2. Vocabulary

BERT uses WordPiece (Wu et al., 2016) with a 30,000 token vocabulary for unsupervised tokenization of the input text. With WordPiece tokenization, any new words can be represented by frequent subwords.
We found that using uncased vocabulary results in slightly better performances in downstream tasks.

4.3. Implementation Details

In our work, we use the Transformers library from Huggingface on Python. For pre-training we used mainly the BERT recommended parameters. We used the default BERT optimizer, AdamWeight decay optimizer, the recommended learning rate of 5e-5, a batch size of 32, a dropout rate of 0.1 and a maximum sequence length of 512.

Data preprocessing and training BERT on financial corpora took significant computational resources. Our entire model procedure took 23 days of computational runtime using a single Nvidia GeForce RTX 2060 6GB GPU. We believe that releasing our pre-trained model FinancialBERT will be useful to the financial researchers and use it on downstream tasks without the necessity of the significant computational resources.

5. Experimental Evaluation

In this section, we describe experiments on Sentiment Analysis task to evaluate the effectiveness of our pre-trained language model.

5.1. Sentiment Analysis

Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also studied in the financial domain. Financial sentiment analysis differs the general one, it is important to guess how the market will react to news and other textual data.

It can be performed by implementing one of the two different approaches using NLP models unsupervised or supervised. As it is known sentiments can be either positive or negative or neutral. NLP algorithms can be used to evaluate if a series of words reflect a positive or negative sentiment. Coming to unsupervised learning, it involves using a rule-based approach by counting the number of positive and negative words based on a dictionary such as Loughran and McDonald (2011). The supervised approach is a classification model that involves using traditional machine learning or deep learning methods.

5.2. Dataset

The main sentiment analysis dataset used in this paper is Financial PhraseBank [5] from (Malo et al., 2014).
Financial Phrasebank consists of 4845 english articles that were categorised by sentiment class and were annotated by 16 researchers with a financial background. The sentiment label is either *positive, neutral or negative*. However the dataset is available in four possible configurations depending on the percentage of agreement of annotators as you can see on the Table 3.

Agreement Level	Positive	Negative	Neutral	# of articles
100%	25.2%	13.4%	61.4%	2262
75% - 99%	26.6%	9.8%	63.6%	1191
66% - 74%	36.7%	12.3%	50.9%	765
50% - 65%	31.1%	14.4%	54.5%	627
Total	28.1%	12.4%	59.4%	4845

Table 3: Description of Financial PhraseBank dataset.

We chose to use the whole Data (at least *50% agreement*). 80% of them as training set, 10% as test set and 10% of the remaining as validation set as show in Table 4.

Dataset	Metric	Train	Dev	Test
Financial PhraseBank	Accuracy + F1	3876	484	485

Table 4: Sentiment Analysis task evaluation metrics, and train, dev, test sets sizes.

5.3. Fine-tuning FinancialBERT

Sentiment analysis is a natural language processing classification task, we train a model that predicts a sentiment label based on an article as input.

Typically, we have two successive steps, one during the pre-training FinancialBERT phase and one during the fine-tuning phase. We firstly conducted unsupervised pre-training on the large financial corpus and then applied supervised fine-tuning on down-stream NLP tasks.

In our work, we use the same fine-tuning architecture used in (Devlin et al., 2019) by adding a dense layer after the last hidden state of the [CLS] token. This is the recommended practice for using BERT for any classification task. Then, the classifier network is trained on the labeled sentiment dataset. We also use cross-entropy loss as the loss function. We used a batch size of 32, a maximum sequence length of 512, and a learning rate of 2e-5 and 5 epochs for fine-tuning our model.

5.4. Results

The following Table 5 presents the sentiment analysis results in a classification report on the test set.
Our fine-tuned FinancialBERT[6] clearly outperforms two common baselines, the BERT-base (Devlin et al., 2019) and

[5]The dataset can be found here: `https://www.researchgate.net/publication/251231364_FinancialPhraseBank-v10`

[6]Our fine-tuned model is available at: `https://huggingface.co/ahmedrachid/FinancialBERT-Sentiment-Analysis`

class	*precision*	*recall*	*f1-score*	*support*
negative	0.96	0.97	0.97	58
neutral	0.98	0.99	0.98	279
positive	0.98	0.97	0.97	148
macro avg	0.97	0.98	0.98	485
weighted avg	0.98	0.98	0.98	485

Table 5: Experimental Results on Financial PhraseBank test set.

FinBERT (Yang et al., 2020), a financial domain specific BERT.
FinancialBERT achieved better performance than the state-of-the-art model on the Financial PhraseBank, which demonstrates its effectiveness in sentiment analysis. We obtained the highest Accuracy (0.12 higher) and F1 score (0.13 higher) than the state-of-the-art model FinBERT as shown in Table 6.

Model	**Accuracy**	**F1-score**
BERT-base (Devlin et al., 2019)	0.84	0.83
FinBERT (Yang et al., 2020)	0.87	0.85
FinancialBERT (ours)	**0.99**	**0.98**

Table 6: Performance of different BERT models on three financial sentiment analysis task.

As expected, we should highlight the importance of pre-training on financial corpora approach which improves performance and enhances the downstream financial sentiment classification task.

6. Conclusion

We presented FinancialBERT, a new pre-trained language model for financial communications, which has been trained on a large corpora and can be fine-tuned for multiple NLP tasks. Requiring minimal task-specific architectural modification, our model achieves state-of-the-art performance on Sentiment Analysis task, significantly outperforming other compared models.

With the release of FinancialBERT, we hope financial practitioners and researchers can benefit from our model without the necessity of the significant computational resources required to train the model.

Future directions include: further exploration of domain-specific pre-training strategies and incorporating more tasks in financial NLP such as Named Entity Recognition (NER) and Question-Answering tasks.

7. Bibliographical References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation.

Yang, Y., UY, M. C. S., and Huang, A. (2020). Finbert: A pretrained language model for financial communications.