NFinBERT: A Number-Aware Language Model for Financial Disclosures

Hao-Lun Lin\textsuperscript{1} and Jr-Shian Wu\textsuperscript{2}
National Chengchi University
Taipei, Taiwan
\{106703027,106703026\}@nccu.edu.tw

Yu-Shiang Huang\textsuperscript{3}
National Taiwan University
Taipei, Taiwan
b05702095@ntu.edu.tw

Ming-Feng Tsai\textsuperscript{4}
National Chengchi University
Taipei, Taiwan
mftsai@nccu.edu.tw

Chuan-Ju Wang\textsuperscript{5}
Academia Sinica
Taipei, Taiwan
cjwang@citi.sinica.edu.tw

Abstract

As numerals comprise rich semantic information in financial texts, they play crucial roles in financial data analysis and financial decision making. We propose NFinBERT, a number-aware contextualized language model trained on financial disclosures. Although BERT and other contextualized language models work well for many NLP tasks, they are not specialized in finance and thus do not properly manage numerical information in financial texts. Therefore, we propose pre-training the language model on a large collection of “pre-processed” financial disclosures in which the numbers in reports are explicitly replaced with the knowledge and understanding of the financial and accounting functions of reports. Experimental results on two fine-tuning classification tasks show that language models pre-trained on financial specialized texts generally outperform BERT. Furthermore, the proposed number-aware NFinBERT significantly surpasses other models when the task becomes more difficult or number-sensitive.

1 Introduction

BERT (Devlin et al., 2018), a state-of-the-art language model, consists of a set of Transformer encoders (Vaswani et al., 2017) stacked on top of each other. In contrast to traditional language models’ prediction of the next token given previous tokens, BERT uses masked LMs (MLMs) and next sentence prediction (NSP) to pre-train the language model, defining language modeling in an unconventional manner. Due to the superior performance of BERT on various natural language processing tasks, numerous related studies and models, including RoBERTa (Liu et al., 2019) and ELECTRA (Clark et al., 2020), have been proposed to advance the state of the art. Other works attempt to adopt this powerful technique to different domains by pre-training language models on corpora from different target domains, such as finance and biomedical sciences (DeSola et al., 2019; Lee et al., 2020).

For applications in finance and accounting, in addition to pre-training domain-specific language models, recent work has focused on fine-tuning the pre-trained model for downstream tasks, including sentiment analysis (Sousa et al., 2019) and numeral category prediction (Wang et al., 2019). However, most such studies directly use the original design of BERT and thus do not properly manage numerical information in financial texts. However, in contrast to other domains, numbers in financial text such as financial disclosures, market commentary, and financial news are especially important for understanding the minutiae of such textual information. Moreover, financial documents usually contain relatively large amounts of numbers; for example, whereas only 0.98% of the tokens in the blog corpus (SCHLER, 2006) are numbers, the 10-K financial reports used here have a much higher proportion of number tokens: 4.79% of all tokens. Thus, properly addressing such numerical information when pre-training the language models is critical to raising the quality of the pre-trained model. For instance, the sentence “Q4 revenue raised by 4,000,000, which is 12.8% of the total amount in the year” is nonsensical if the numbers in it are not properly interpreted.

To this end, we propose NFinBERT,\textsuperscript{1} a number-
aware contextualized language model pre-trained on a large collection of “pre-processed” financial disclosures, for which we explicitly replace numbers in reports with the knowledge and understanding of the financial and accounting functions of reports. We conduct two downstream tasks to evaluate the proposed model: one is binary classification for risk sentence detection and the other is 12-class classification for sentence-level numeral category prediction. The results indicate that language models pre-trained on financial specialized texts generally outperform BERT. Furthermore, the proposed number-aware NFinBERT significantly surpasses other models on more difficult or numbersensitive tasks.

2 Pre-Training Models on Financial Reports

2.1 Data and Preprocessing

To pre-train the domain-specific language models on financial reports, we used the 10-K reports from 1996 to 2013 collected by (Loughran and McDonald, 2011). Moreover, following previous studies (Kogan et al., 2009; Tsai et al., 2016; Buehlmaier and Whited, 2018), we used only Section 7 “Management’s Discussion and Analysis of Financial Conditions and Results of Operations” (MD&A) in the experiments as it contains the most important forward-looking statements about the companies. The resultant corpus contains 183,115 MD&A sections from different companies in 18 years, with 45,126,776 sentences and 838,842,639 tokens in total.

To train NFinBERT, the number-aware language model, we identify the 11 common categories of numbers in financial reports listed in Table 1, with help from several domain experts in finance and accounting. Note that these tokens are usually not pure integers or floats, and may contain commas or parentheses due to the number formats used in accounting, making the preprocessing more complicated than that for normal numbers. For instance, one million is sometimes presented as “1,000,000” in financial reports, and “(1,000)” represents negative one thousand. For such complex preprocessing, we used both regular expressions and named entity recognition (NER) to recognize tokens containing numbers and slot them into one of the 11 classes. For example, $1,000,000 in the reports was masked as the token [MONEY], and 95% was masked as [RATIO]. Therefore, in addition to [CLS] and [SEP], BERT’s original masks, we here add 11 masks to train NFinBERT. The distribution of categories is listed in the last column of Table 1.

2.2 FinBERT and NFinBERT

BERT (Devlin et al., 2018) is a language model containing a set of Transformer encoders (Vaswani et al., 2017) stacked on top of each other; such a design defines language modeling in an unconventional manner. Following previous studies (Howard and Ruder, 2018; Araci, 2019), we pre-train language models on finance, the target domain, and experiment with two approaches: 1) Pre-training the model on a large collection of financial reports—the original corpus containing 92,402,863 sentences—and 2) Pre-training the model on a corpus in which all of the numbers have been replaced by the tokens listed in Table 1.

As in (DeSola et al., 2019), we pre-train the two language models using 10K warm-up steps, setting the max sentence length and batch size both equal to 128, the maximum predictions per sequence to 20, and the learning rate to $5 \times 10^{-4}$. The performance on MLM and NSP is summarized in Table 2 and is generally consistent with the results in (DeSola et al., 2019).

3 Experiments

In this section, we describe experiments on two fine-tuned classification tasks to evaluate the effectiveness of pre-trained language models. The first task (denoted as Task 1 hereafter) considers binary classification for identifying risk sentences in financial reports, and the second (Task 2) is multi-class classification regarding the types of numbers mentioned in sentences extracted from the reports.

3.1 Datasets

3.1.1 Task 1: Binary Classification for Risk Prediction

We conducted the experiments on 10K-Sentence, a sentence-level risk classification dataset (Lin et al., 2020), consisting of 2,432 sentences extracted from the 10-K reports from 1996 to 2013; each sentence in 10K-sentence is categorized as either risky or non-risky by annotators specializing in finance or
<table>
<thead>
<tr>
<th>Category</th>
<th>Explanation</th>
<th>Example</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>[MONEY]</td>
<td>monetary numbers</td>
<td>$600,000</td>
<td>11,616,433</td>
</tr>
<tr>
<td>[DATE]</td>
<td>dates</td>
<td>2020-02-02</td>
<td>21,480,578</td>
</tr>
<tr>
<td>[PHONE]</td>
<td>phone numbers</td>
<td>800-555-5555</td>
<td>3,525</td>
</tr>
<tr>
<td>[BOND]</td>
<td>bond ratings</td>
<td>Aaa3</td>
<td>34,519</td>
</tr>
<tr>
<td>[ORDINAL]</td>
<td>ordinal information</td>
<td>Note 4</td>
<td>2,950,815</td>
</tr>
<tr>
<td>[QUANTITY]</td>
<td>quantities</td>
<td>100,000 shares</td>
<td>1,476,558</td>
</tr>
<tr>
<td>[ADDRESS]</td>
<td>addresses</td>
<td>Rd. 3</td>
<td>13,577</td>
</tr>
<tr>
<td>[RATIO]</td>
<td>numbers related to ratios</td>
<td>1-to-5</td>
<td>18,246</td>
</tr>
<tr>
<td>[PERCENT]</td>
<td>percentages</td>
<td>95%</td>
<td>4,342,499</td>
</tr>
<tr>
<td>[TIME]</td>
<td>time unit smaller than a day</td>
<td>1 hour</td>
<td>53,049</td>
</tr>
<tr>
<td>[OTHER]</td>
<td>other numbers</td>
<td>G-8</td>
<td>8,491</td>
</tr>
</tbody>
</table>

Table 1: Categories of finance numbers

<table>
<thead>
<tr>
<th>Models</th>
<th>Steps</th>
<th>MLM acc.</th>
<th>NSP acc.</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>FinBERT 150K</td>
<td>73.66%</td>
<td>96.62%</td>
<td>1.2604</td>
<td></td>
</tr>
<tr>
<td>FinBERT 300K</td>
<td>75.98%</td>
<td>97.37%</td>
<td>1.1257</td>
<td></td>
</tr>
<tr>
<td>NFinBERT 150K</td>
<td>76.48%</td>
<td>97.62%</td>
<td>1.1080</td>
<td></td>
</tr>
<tr>
<td>NFinBERT 300K</td>
<td>77.55%</td>
<td>98.25%</td>
<td>1.0416</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results for pre-trained FinBERT and NFinBERT

3.1.2 Task 2: Multi-class Classification for Number Category Prediction

For the second task, we constructed a new dataset containing 25,261,147 sentences in total extracted from the 10-K reports from 1996 to 2013, each of which is labeled with one of the 11 categories listed in Table 1 plus a “[Nothing]” type. Specifically, the dataset is composed of all sentences in the 10-K reports from 1996 to 2013 containing exactly one number or no number, the former of which was labeled with one of the 11 categories and the latter of which was labeled with the [Nothing] type. For the following experiments, we performed 12-class classification for number category prediction on 1,403,397 randomly selected sentences, constituting 5.6% of the original dataset with the same category distribution of the original dataset, due to computational resource limitations.

3.2 Experimental Settings

In both tasks, we split the datasets into training, validation, and test sets at an 8:1:1 ratio, respectively. Moreover, to mitigate the label imbalance problem in Task 2, we down-sampled the training data to the median of the numbers of instances in each category, resulting in 177,473 sentences in total. The resulting category distribution for model training is illustrated in the first row of Figure 1. Note that only the training set was down-sampled; the validation and test sets retained the original category distribution. We used 15 epochs to fine-tune all BERT-based models, setting the max sentence length to 128 and the batch size to 32, and used the validation set to search learning rates in \{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}. The best learning rates for BERT, FinBERT, and NFinBERT were 10^{-5}, 10^{-4}, and 10^{-4}, respectively. Note that the results on both validation and test sets are the averaged results over five repetitions.

---

4Dataset details can be found in (Lin et al., 2020).
5As the sentences are from the reports of 1996 to 2013 (i.e., 18 years in total), we here simulate a one-year dataset by randomly sampling 1/18 \approx 5.6\% of the sentences from the original dataset.
6Note that we reduced the training instances only in categories for which the number of instances were higher than the median; we kept the rest categories unchanged.
7Note that in our experiments, as we found that using 2\% of the down-sampled sentences achieves satisfactory performance, we here used only 3,195 sentences for Task 2 model training due to computational resource limitations.
3.3 Results

For both tasks, we compared three BERT-based models with three baselines—TF-IDF bag-of-words (BOW) with logistic regression, convolution neural network (CNN) (Kim, 2014), and fastText (Joulin et al., 2016)—and summarize their performance in Table 3. As shown in the table, for Task 1, all three BERT-based models yield comparable performance, significantly better than the three baseline models.

On Task 2, which is more difficult than Task 1, both FinBERT and NFinBERT surpass BERT\(^8\) in terms of macro F1 by a significant amount.\(^9\) Figure 1 details the performance of each category in terms of F1 score with a heatmap for all three BERT-based models. From the figure, we observe that BERT is outperformed by FinBERT and NFinBERT for the categories with the fewest training instances ([OTHER], [PHONE], and [RATIO]); this is why BERT achieves better accuracy but a lower macro F1 score in Table 3. Moreover, for these three categories, NFinBERT yields more accurate prediction than FinBERT, suggesting that the number-aware pre-trained language model is beneficial for Task 2.

4 Conclusion

We introduce NFinBERT, a number-aware language model trained on financial disclosures, in which we identify 11 categories of numeral tokens with the knowledge and understanding of the financial and accounting functions of reports and replace them with additional masks to pre-train the model. The experimental results show that it is crucial to pre-train BERT on a finance-specific corpus for finance-related downstream tasks; moreover, the proposed NFinBERT outperforms other compared models for 12-class classification for sentence-level numeral category prediction.

References


Kevin Clark, Minh-Thang Luong, Quoc V Le, and


