NFinBERT: A Number-Aware Language Model for Financial Disclosures

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

Yu-Shiang Huang[3]
National Taiwan University
Taipei, Taiwan
b05702095@ntu.edu.tw

Ming-Feng Tsai[4]
National Chengchi University
Taipei, Taiwan
mftsai@nccu.edu.tw

Chuan-Ju Wang[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 "preprocessed" 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 numeral 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,[1] a number-

[1]The pre-trained model will be publicly available upon

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 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 number-sensitive 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).[2] 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)[3] 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 5×10^{-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

publication.

[2] https://sraf.nd.edu/textual-analysis/resources/

[3] SpaCy was used for NER.

Category	Explanation	Example	Amount
[MONEY]	monetary numbers	$600,000	11,616,433
[DATE]	dates	2020-02-02	21,480,578
[PHONE]	phone numbers	800-555-5555	3,525
[BOND]	bond ratings	Aaa3	34,519
[ORDINAL]	ordinal information	Note 4	2,950,815
[QUANTITY]	quantities	100,000 shares	1,476,558
[ADDRESS]	addresses	Rd. 3	13,577
[RATIO]	numbers related to ratios	1-to-5	18,246
[PERCENT]	percentages	95%	4,342,499
[TIME]	time unit smaller than a day	1 hour	53,049
[OTHER]	other numbers	G-8	8,491

Table 1: Categories of finance numbers

Models	Steps	MLM acc.	NSP acc.	Loss
FinBERT	150K	73.66%	96.62%	1.2604
FinBERT	300K	75.98%	97.37%	1.1257
NFinBERT	150K	76.48%	97.62%	1.1080
NFinBERT	300K	77.55%	98.25%	1.0416

Table 2: Results for pre-trained FinBERT and NFinBERT

linguistics,[4] resulting in 1,536 risky sentences and 896 non-risky sentences for binary classification.

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[5] 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,[6] resulting in 177,473 sentences in total.[7] 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.

[4] Dataset details can be found in (Lin et al., 2020).

[5] As 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.

[6] Note 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.

[7] Note 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.

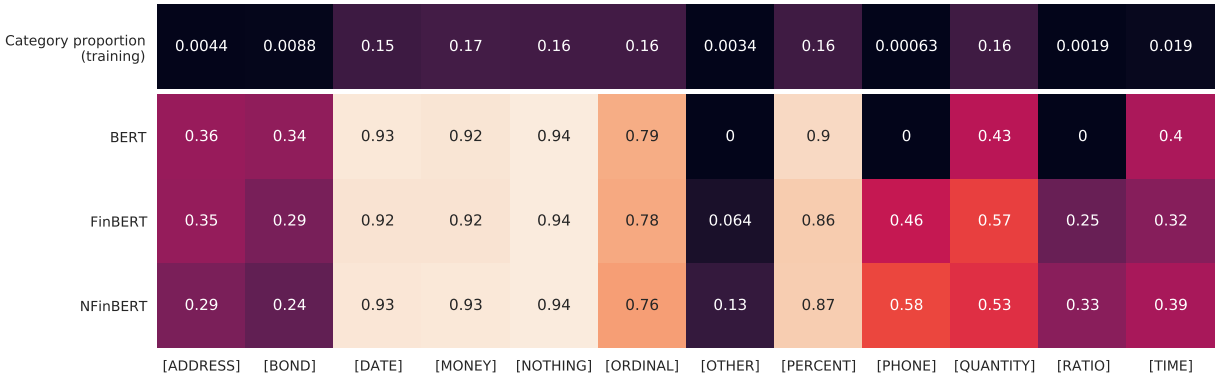

Figure 1: F1 score of each category

	Task 1	Task 2	
Models	Accuracy	Accuracy	Macro F1
BOW	86.83%	72.51%	36.91%
FastText (Joulin et al., 2016)	86.42%	62.44%	36.91%
CNN (Kim, 2014)	87.36%	72.51%	32.37%
BERT (Devlin et al., 2018)	88.61%	**91.51%**	50.49%
FinBERT	**88.75%**	91.34%	56.10%
NFinBERT	88.61%	91.19%	**57.67%**

Table 3: Performance of two fine-tuning tasks

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.

[8]The BERT-Base, Uncased pre-trained model was used in the experiments.

[9]The improvements compared to BERT are statistically significant at $p < 0.01$ with a paired t-test.

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

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Matthias MM Buehlmaier and Toni M Whited. 2018. Are financial constraints priced? Evidence from textual analysis. *The Review of Financial Studies*, 31(7):2693–2728.

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

Christopher D Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Vinicio DeSola, Kevin Hanna, and Pri Nonis. 2019. FinBERT: pre-trained model on SEC filings for financial natural language tasks. *Working paper*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Sheng-Chieh Lin, Wen-Yuh Su, Po-Chuan Chien, Ming-Feng Tsai, and Chuan-Ju Wang. 2020. Self-attentive sentimental sentence embedding for sentiment analysis. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1678–1682.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Jonathan SCHLER. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Symposium on Computational Approaches for Analyzing Weblogs, 2006*, pages 199–205.

Matheus Gomes Sousa, Kenzo Sakiyama, Lucas de Souza Rodrigues, Pedro Henrique Moraes, Eraldo Rezende Fernandes, and Edson Takashi Matsubara. 2019. Bert for stock market sentiment analysis. In *Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence*, pages 1597–1601.

Ming-Feng Tsai, Chuan-Ju Wang, and Po-Chuan Chien. 2016. Discovering finance keywords via continuous-space language models. *ACM Transactions on Management Information Systems*, 7(3):1–17.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, pages 5998–6008.

Wei Wang, Maofu Liu, Yukun Zhang, Junyi Xiang, and Ruibin Mao. 2019. Financial numeral classification model based on bert. In *Proceedings of the NII Conference on Testbeds and Community for Information Access Research*, pages 193–204. Springer.