FinSemantics: Predicting Financial Sentiment Through Transfer Learning

1 Abstract

Financial sector accumulates large amount of financial communication text. There is an increasing interest in the financial text mining tasks. Over the past few years, Natural Language Processing (NLP) based on deep learning has advanced rapidly. Significant progress has been made with deep learning, which shows promising results on financial text mining models. However, financial sentiment analysis is still a demanding task due to its specialized language and lack of labeled data in financial domain. General purpose machine learning methods are not as effective due to specialized language used in the financial context. To address this issue, we use pre-trained language models to analyze sentiment in financial text since these models can be further trained on domain specific corpora and they require fewer labeled examples to tackle NLP tasks in financial domain. We introduce financial variants of various pre-trained language models based on BERT and GPT2 trained on large-scale corpora to tackle NLP sentiment analysis tasks in financial domain. We also propose variants of models simultaneously trained on general corpora and financial domain corpora. Our results show that our models outperforms all current state-of-the-art model for two financial sentiment analysis datasets. We find that the proposed modified methods outperforms state-of-the-art machine learning methods even with a smaller training set and fine-tuning only a part of the models. Extensive experimental results demonstrate the effectiveness and robustness of especially xxx. The datasets, source-code, and pre-trained models are available online.

2 Introduction

Financial text data are utilized to predict and analyze future market trends in finance and economics [**?**]. Financial text mining plays an important role in financial technology, whether for official company announcements or analyst reports. The volume of financial text data continues to keep increasing since an unprecedented number of such texts are created daily. As a result, manual analysis of such texts and gaining actionable insights from them as a result of this analysis is a quite difficult task. Recent progress in machine learning have made financial text mining models in FinTech possible. Nevertheless, supervised training data construction is prohibitively expensive since it requires the use of expert knowledge in financial domain. Due to the small amount of labeled training data that can be used for financial text mining tasks, most financial text mining models cannot directly utilize recently-developed deep learning techniques [**?**].

In this paper, we focus on polarity analysis, which is classifying text as positive, negative or neutral, in a specific domain. It requires to address two challenges: 1) The most sophisticated classification methods that make use of neural nets require vast amounts of labeled data and labeling financial text snippets requires costly expertise. 2) The sentiment analysis models trained on general corpora are not suited to the task, because financial texts have a specialized language with unique vocabulary and have a tendency to use vague.

One solution is to use thoroughly and manually crafted financial sentiment lexicons since they incorporate existing financial semantics into textual analysis [6]. However, such methods are based on word counting, which fail to analyze deeper semantic meaning of provided text.

Language models have become a key step to achieve state-of-the art results in many different Natural Language Processing (NLP) tasks. Leveraging the huge amount of unlabeled texts nowadays available, they provide an efficient way to pre-train continuous word representations that can be fine-tuned for a downstream task, along with their contextualization at the sentence level. This has been widely demonstrated for English using contextualized representations [2, 8, 4, 9, 3, 13].

NLP transfer learning methods look like a promising solution to both of the challenges mentioned above. The core idea behind these models is that by training language models on very large corpora and then initializing down-stream models with the weights learned from the language modeling task, a much better performance can be achieved. The initialized layers can range from the single word embedding layer [8] to the whole model [4]. This approach should, in theory, be an answer to the scarcity of labeled data problem. Language models don't require any labels, since the task is predicting the next word. They can learn how to represent the semantic information. That leaves the fine-tuning on labeled data only the task of learning how to use this semantic information to predict the labels.

One particular component of the transfer learning methods is the ability to further pre-train the language models on domain specific unlabeled corpus. Thus, the model can learn the semantic relations in the text of the target domain, which is likely to have a different distribution than a general corpus. This approach is especially promising for a niche domain like finance, since the language and vocabulary used is dramatically different than a general one.

As a summary, we have the following main contributions in this paper:

- We introduce financial extensions of various BERT and GPT models for financial NLP tasks to transfer knowledge from financial domain corpora. We evaluate the proposed methods on two financial sentiment analysis datasets.
- We conduct number of experiments on several Financial PhraseBank and FiQA sentiment scoring benchmark datasets. We achieve the state-of-the-art results on both financial datasets.
- We compare the performance of various BERT and GPT models and understand why certain models can better explain financial text mining tasks. Our models are capable of efficiently capturing language knowledge and semantic information in large-scale pre-training corpora.
- We have investigated several aspects of the introduced models, including: effects of further pre-training on financial corpus, training strategies to prevent catastrophic forgetting and fine-tuning only a small subset of model layers for decreasing training time without a significant drop in performance.
- We implemented our BERT algorithms on Tensorflow and Hugginface frameworks. We make the source code and pre-trained models publicly available. With minimal task-specific architecture modifications.

2.1 Related Work

There are previous attempts to use BERT models on financial corpora [1, 5, 12].

Another work [11]. has employed high level semantic representations and methods of inductive transfer learning for NLP by using ULMFit [4].

There is other work on detecting semantic orientations [**?**].

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a

1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations [10]

The use of robo-readers to analyze news texts is an emerging technology trend in computational finance. In recent research, a substantial effort has been invested to develop sophisticated financial polarity-lexicons that can be used to investigate how financial sentiments relate to future company performance. However, based on experience from other fields, where sentiment analysis is commonly applied, it is well-known that the overall semantic orientation of a sentence may differ from the prior polarity of individual words.

3 Experiments

Using BERT models in finance have also been used in the literature [1, 12]. Other pretrained models have also been used [11].

We do not have many financial datasets for financial tasks.

3.1 FiQA

The provided training dataset for WWW '18 [7] contains a total of 1174 examples from news headlines and tweets. Each example contains the sentence and the sentence snippet associated with the target entity, aspect, and sentiment score. A sample FiQA entry is shown in 1. A Level 1 Aspect label takes on one of 4 possible labels (Corporate, Economy, Market or Stock), and our Level 2 Aspect label takes on one of 27 possible labels (Appointment, Risks, Dividend Policy, Financial, Legal, Volatility, Coverage, Price Action, etc.). The original dataset contained a small number of multilabel examples, however, we considered this number too few to train a meaningful multilabel classifier. Thus, we slightly stray from the original WWW '18 task for the purpose of this research. Finally, sentiment score takes on a continuous value between -1 and 1 – most negative to most positive.

We can use FiQA dataset for both classification purpose in terms of aspect levels, and regression purpose in terms of sentiment score.

A large collection of financial reports published annually by publicly-traded companies is employed to conduct our experiments; moreover, two analytical techniques – regression and ranking methods – are applied to conduct these analyses [**?**].

Sentence	easyJet expects resilient demand to withstand security fears.
Aspect Level 1	Corporate
Aspect Level 2	Risks
Sentiment Score	0.165
Target	easyJet

Table 1: An example entry from FiQA

Reuter's and Bloomberg dataset: Datasets are in Emre's email. `https://github.com/philipperemy/financial-news-dataset`

Reuter's news dataset: `https://github.com/duynht/financial-news-dataset`

NLTK's corpus `https://www.kaggle.com/boldy717/reutersnltk#__sid=js0`

Fed Meeting Notes `https://fraser.stlouisfed.org/title/federal-open-market-commit` `browse=2020s`

FOMC Statements Scraper https://github.com/souljourner/FOMC-Statements-Minutes-Scraper Some cleaned transcripts https://github.com/ali-wetrill/FOMCTranscriptAnalysis

Another source for datasets: `https://rstudio-pubs-static.s3.amazonaws.com/495650_c9c874694f164fb5948031801079157f.html#3_data`

`https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists`

Finally, the texts are stemmed using the Porter stemmer when needed.

10K dataset together with volatilities http://ifs.tuwien.ac.at/ admire/financialvolatility/

Recently, unsupervised pre-training of language models on large corpora has significantly improved the performance of many NLP tasks. The language models are pretained on generic corpora such as Wikipedia. However, sentiment analysis is a strongly domain dependent task. Financial sector has accumulated large scale of text of financial and business communications. Therefore, leveraging the success of unsupervised pretraining and large amount of financial text could potentially benefit wide range of financial applications.

Recent progress in pre-trained neural language models has significantly improved the performance of many natural language processing (NLP) tasks. In this paper we propose a new model architecture DeBERTa (Decoding-enhanced BERT with disentangled attention) that improves the BERT and RoBERTa models using two novel techniques. The first is the disentangled attention mechanism, where each word is represented using two vectors that encode its content and position, respectively, and the attention weights among words are computed using disentangled matrices on their contents and relative positions. Second, an enhanced mask decoder is used to replace the output softmax layer to predict the masked tokens for model pretraining. We show that these two techniques significantly improve the efficiency of model pre-training and performance of downstream tasks. Compared to RoBERTaLarge, a DeBERTa model trained on half of the training data performs consistently better on a wide range of NLP tasks, achieving improvements on MNLI.

Recent progress in pre-trained neural language models has significantly improved the performance of many natural language processing (NLP) tasks.

We show that these two techniques significantly improve the efficiency of model pre-training and performance of downstream tasks. In financial domain, the same is observed.

BERT and its variants have significantly enhanced word vector representation. Here, we will focus on specific BERT application on financial datasets.

3.2 Financial Phrasebank

http://www.cs.cmu.edu/ ark/10K/data/ Metadata var

10K downloads: https://pypi.org/project/sec-edgar-downloader/. Filing date is in each text file.

extract MDA and tokenize new files in Noah's website can be used to clean the dataset. CIK Ticker Mapping: https://www.sec.gov/include/ticker.txt

3.3 Corporate Reports 10-K & 10-Q

The most important text data in finance and business communication is corporate report. In the United States, the Securities Exchange Commission (SEC) mandates all publicly traded companies to file annual reports, known as Form 10-K, and quarterly reports, known as Form 10-Q. This document provides a comprehensive overview of the company's business and financial condition. Laws and regulations prohibit companies from making materially false or misleading statements in the 10-Ks. The Form 10-Ks and 10-Qs are publicly available and can be accesses from SEC website. We obtain 60,490 Form 10-Ks and 142,622 Form 10-Qs of Russell 3000 firms during 1994 and 2019 from SEC website. We only include sections that are textual components, such as Item 1 (Business) in 10-Ks, Item 1A (Risk Factors) in both 10- Ks and 10-Qs and Item 7 (Managements Discussion and Analysis) in 10-Ks.

References

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

[2] Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. NIPS'15, page 3079–3087, Cambridge, MA, USA, 2015. MIT Press.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Jeremy Howard and Sebastian Ruder. 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, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[5] Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and J. Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *IJCAI*, 2020.

[6] TIM LOUGHRAN and BILL MCDONALD. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

[7] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[8] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages

2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[9] A. Radford. Improving language understanding by generative pre-training. 2018.

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[11] Steve Yang, Jason Rosenfeld, and Jacques Makutonin. Financial aspect-based sentiment analysis using deep representations, 2018.

[12] Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *CoRR*, abs/2006.08097, 2020.

[13] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc., 2019.