Predicting Stock Returns and Risk from Financial Reports with Pretrained Language Models

1 Abstract

2 Introduction and Related Work

There are previous methods that use Machine learning to predict stock markets. [2] proposes a deep learning method for event-driven stock market prediction. First, events are extracted from news text, and represented as dense vectors, trained using a novel neural tensor network. Second, a deep convolutional neural network is used to model both short-term and long-term influences of events on stock price movements.

3 Experiments

There are similar papers in the literature [4]. Using BERT models in finance have also been used in the literature [1, 6]

We do not have many financial datasets for financial tasks.

We will compare our BERT based method with the following approaches:

- Baseline method will be volatility prediction based on GARCH similar to [4].
- SVM-based/Random Forest based volatility prediction as in paper [4].
- SVM-based/Random Forest based return prediction as in paper. Previous papers has not focused on return prediction.
- Original BERT trained model.
- Elmo-based trained model.

One evaluation will based Mean squared error (MSE) based.

Other evaluation type will be based on portfolio construction:

- Buckets on predicted returns.
- Stock portfolios based on Markovitz portfolios based on predicted returns and volatility and covariance between pairs.

We will focus on predicting the returns and volatilities for next quarter, year half and half respectively. Portfolios will be balanced monthly.

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

3.1 FiQA

The provided training dataset for WWW '18 [5] contains a total of 1,174 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.

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 [**?**].

The authors in [3] used a dataset of more than 900,000 news stories to test whether news can predict stock returns. They measured sentiment with a proprietary Thomson Reuters neural network and found that daily news predicts stock returns for only one to two days, confirming previous research. Weekly news, however, predicts stock returns for one quarter. Positive news stories increase stock returns quickly, but negative stories receive a long-delayed reaction. Much of the delayed response to news occurs around the subsequent earnings announcement.

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

We can predict volatility or whether volatility will go up or down of the following instruments:

- S&P 500
- 13-week Treasury Bills
- 10-year Treasury Notes

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

We mainly focus on predicting the change of the Standard & Poor's 500 stock (S&P 500) index, obtaining indices and stock price data from Yahoo Finance. Standard & Poor's 500 is a stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ.

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

Finally, the texts are stemmed using the Porter stemmer.

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.

Predicting Risk from Financial Reports with Regression is a related paper.

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.

In this paper, we use variants of BERT. 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. 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.

Our approach will be based on `https://github.com/microsoft/DeBERTa`

Sample datasets are:

- FIQA: Financial Opinion Mining and Question Answering `https://sites.google.com/view/fiqa/home`

- 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

Ticker price: Yahoo finance download https://towardsdatascience.com/downloading-historical-stock-prices-in-python-93f85f059c1f

This text regression problem has been discussed before.

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] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 2327–2333. AAAI Press, 2015.

[3] Steven L. Heston and Nitish Ranjan Sinha. News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73(3):67–83, 2017.

[4] Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 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, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[5] 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.

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