DeepFinRisk: 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. [?] 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.

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.

Financial risk, defined as the chance to deviate from return expectations, is most commonly measured with volatility. Due to its value for investment decision making, volatility prediction is probably among the most important tasks in finance and risk management. Although evidence exists that enriching purely financial models with natural language information can improve predictions of volatility, this task is still comparably underexplored. We introduce PRoFET, the first neural model for volatility prediction jointly exploiting both semantic language representations and a comprehensive set of financial features. As language data, we use transcripts from quarterly recurring events, so-called earnings calls; in these calls, the performance of publicly traded companies is summarized and prognosticated by their management. We show that our proposed architecture, which models verbal context with an attention mechanism, significantly outperforms the previous state-of-the-art and other strong baselines. Finally, we visualize this attention mechanism on the token-level, thus aiding interpretability and providing a use case of PRoFET as a tool for investment decision support [?].

3 Experiments

There are similar papers in the literature [?]. 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 [?].
- SVM-based/Random Forest based volatility prediction as in paper [?].
- 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 type of 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.

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/

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.