Statement of Work

Harvard AC295/CS115

EDUARDO PEYNETTI, JESSICA WIJAYA, ROHIT BERI, STUART NEILSON

Background

Recent advances in natural language processing and machine learning, together with increasing computing power, have made available tools and methods to exploit and model large and unstructured sources of textual data for a variety of tasks, such as text classification, sentiment analysis, text summarization, among many others, in increasingly sophisticated ways.

At the same time, the recent explosion and availability of large amounts of financial market data and financial news provides a rich playground in which to test these methodologies and models. Financial markets derive a significant part of their moves from news releases. Financial statements, economic releases, technological advances, analyst reviews, political and global developments all play active roles in the dynamics of prices.

Assigning a sentiment score to a given document by estimating weights based on a pre-specified sentiment dictionary is a well-studied methodology (see, **Loughran 2016** for a historical summary). However, the use of state-of-the-art natural language processing models for this particular task isn't as well developed in the literature.

Problem Statement

We explore the use of state-of-the-art Natural Language Processing models to assign a sentiment score to financial documents and attempt to predict returns based on this information. We intend to explore sentiment on a universe of 500+ stocks, labeled by industry and market capitalization, and to build a long-short portfolio based on the sentiment scores from the models.

Following **Ke 2019**, we first label news articles by stocks mentioned in the article, and then assign a "soft" sentiment score by looking at the stock's daily return direction. With this information, we intend to train different models to obtain a sentiment score for each news article.

As a baseline, we use the Loughran/McDonald sentiment dictionary, a classical benchmark for financial sentiment, as well as a sentiment score obtained from a BERT model trained with Rotten Tomatoes reviews, available through HuggingFace.

We intend to train:

- The pretrained BERT sentiment model mentioned above, finetuned to financial data and our defined sentiment scores
- A pretrained BERT model, where we train sentiment with financial data and our defined sentiment scores from scratch.

We would like to measure the effectiveness of our model by estimating out-of-sample returns on a long-short portfolio of stocks generated by these sentiment scores. We would also like to explore what kinds of words appear to be relevant in classifying news sentiment given by our trained models, compared to typically used dictionaries like Loughran/McDonald.

We'd also like to look at a strategy creating using an ensemble of all these models. We believe that each model might capture different facets of sentiment, and we may be able to obtain better out-of-sample returns by using all of them together.

Resources

We require:

- A multi-year database of financial news, labeled by stocks mentioned
    - We are still in the process of choosing the source of financial news
- A multi-year database of stock end-of-day prices, as well as industry and market cap labels.
- Pre-trained BERT models, available through the HuggingFace library
- Tools for entity labelling and tokenization, available through the Spacy library
- Multi-stock market analysis tools that can handle market calendars, stock splits, dividends, de-listings, etc., such as those provided the Zipline library.

High-Level Project Stages

- Obtain raw financial news database
- Parse and tag news and price databases
- Create baseline models
- Train BERT models
- Create Visualization Models
- Market and sentiment analysis
- Create Cloud Deployment Architecture

References

- Loughran, Tim, and Bill Mcdonald, 2016, Textual Analysis in Accounting and Finance: A Survey, Journal of Accounting Research 54, 1187–1230
- Zheng Tracy Ke Bryan T. Kelly Dacheng Xiu, 2019, Predicting Results with Text Data, National Bureau of Economic Research, Working Paper 26186