Harvard Extension School, S-117

Guanwen Cheng, Chunyi Li

Final Project Outline

**Project Topic**

Investigate impacts from the wording of Federal Reserve (Fed) meetings to financial market.

**Background**

Financial Markets react to many kinds of information, such as GDP growth and unemployment rate. Traditionally, such quantitative factors can be easily included in trading or investment models, which provide market practitioners signals. However, along with the rise of Deep Learning, qualitative factors play more and more important roles in the investment world. One main category among qualitative factors is language analysis based on Natural Language Processing (NLP). The wording in policies from the government, especially regulators, such as Federal Reserve, is a perfect example.

**Target Problem**

The Federal Reserve Open Market Committee (FOMC) has eight regular meetings per year to determine U.S. monetary policy. Fed policy is based on the current economy environment and the forecasted situations. Meanwhile, the policy itself also impacts the future economy. The endogenous relationship between policy and reality makes their cause-effect relationship complicated. After each FOMC meeting, Fed publishes press conference minutes, statements, as well as scripts in text, as a result, the market regularly observes significant volatility around FOMC meetings. However, because of the complexity of real-world financial markets, it is not straightforward to quantify the exact impact from FOMC meetings on markets.

**Project Object**

The objective of this project is to apply NLP on those texts published by FOMC to attempt to quantify the communications in a systematic manner and find the relationship with Treasury yield curve.

**Dataset**

The dataset mainly includes two parts. One is historical wordings from FOMC meetings. It is available from Fed’s website. We need to preprocess the documents in an organized way, so as to convert them as appropriate inputs to models. Another part is the Treasury yield curve. It is easy to obtain from financial data vendors, such as Bloomberg. It is clean time-series numerical data.

**Exploratory Analyses**

Exploratory analyses include total numbers of files, paragraphs, sentences and words. Top words and Word Cloud.

**Modeling**

First, we will implement a dictionary-based model to generate sentiment scores and investigate the relationship between the scores and the changes of 10-year treasury rate in following periods, i.g. one month. The dictionary-based model will also serve as the baseline model.

Second, we will apply topic modeling techniques to find different topics in the text, calculate sentiment scores based on these topics, and investigate relationships between topic-based sentiment scores and the changes of 10-year treasury rate in following periods.

Finally, we will apply GloVe and BERT embedding to have better word representations and to capture the context. Since the number of time-dimensional observations is too small to use complex models, we will use low-dimensional sentence representations based on the word representations. RNN and LSTM can have difficulty dealing with very long input sequences as these models tend to give dis-proportionally higher weights to input vectors near the end of a sequence; however, a lengthy FOMC record can have as many as 471 sentences. We will use a CNN layer that summarizes sentence embeddings in a document as a “document vector and use it in a classification model to see if we could have better predictions.

**Deployment**

Use fast.ai for deployment

**Limitation**

To be identified along with project development.

**Impact and Future Works**

Based on the result and limitation, we will justify if economic value created from this model has meaningful business impact. We will also explore that, if the model is deployed, what are the major components in ROI analysis, such as feasibility and technical debt.

**Appendix:**

FOMC Website

<https://www.federalreserve.gov/monetarypolicy/fomccalendars.html>

the website contains materials for each meeting but the contents are not always the same. There are unscheduled meetings and conference calls in addition to regular meetings. Some texts are in html while others are only in pdf files. Different websites for historical data and the page structure varies. Most of the work can be done by BeautifulSoup for HTML text. Use textract to extract PDF and re module for searching by regular expression.

Loughran and McDonald Sentiment Word Lists

<https://sraf.nd.edu/textual-analysis/resources/>

Financial Sentiment Dictionaries, which are sentiment dictionaries for general and financial sentiment analysis. It’s used to look up sentiment score for each word.

Treasury Yields

[https://www.treasury.gov/resource-center/data-chart-center/interest rates/Pages/TextView.aspx?data=yieldAll](https://www.treasury.gov/resource-center/data-chart-center/interest%20rates/Pages/TextView.aspx?data=yieldAll)

Historical daily U.S. Treasury yields were collected from the US treasury website and 10-year treasury rate changes are used.