Project Proposal

**From:** Haijin Zhu, Xiaofei Shi, Yuin Fan, Ziyu Ma

**Date:** November 8th, 2020

**Project Description**

Sentimental analysis in stock trading is a very interesting topic for all of us. Always, we focus more on the numerical dataset in the Finance Area, and intuitively we connect Finance with numbers together. Therefore, in this project, we are more curious about the impact from those textual data and those questions related to, for example, how textual data reflect the perspectives from the firms, how investors would react when they see the publications of those textual files, as well as, eventually how the stock market will move. In this project, we would try our best to solve those questions and hope to get better outcomes.

**Main Questions**

The predictability of attitude of 10-K files on the stock returns and volatility.

**Hypothesis**

1. Sentiment index has a negative correlation with the monthly stock return.
2. Uncertainty words index has a positive correlation with the monthly stock volatility.

**Possible Data Requirements**

1. 10-K, 10-Q transcripts on Securities and Exchange Commission (SEC) EDGAR (2002 - 2020);
2. Conference call transcripts on Factiva (Dow Jones) (2002 - 2020);
3. Financial word dictionaries from Loughran and McDonald (2011);
4. Monthly and daily stock return for selected entities and S&P 500 (2002 - 2020);
5. Monthly and daily Treasury bill (2002 - 2020);

**Project Schedule**

1. Collect the 10-K, 10-Q, conference calls’ transcripts and pre-process those files by using web crawling from EDGAR and Factiva.
2. Remove all encoded images, tables, exhibits, HTML code, special symbols and other non-test items from the documents.
3. Create a bag of words, tokenize, and remove the stop words.
4. Calculate the monthly aggregated financial statement and conference call tone by using (positive words - negative words) / total words
5. Smooth the tone index using a four-month moving average weighted by the number of conference calls in each month to remove seasonality and idiosyncratic jumps.
6. Compute the monthly composite manager sentiment index.
7. Use machine learning models to predict the relationship between monthly composite manager sentiment index and the monthly excess return.
8. Use machine learning models to predict the relationship between uncertainty index and the monthly stock volatility.