Title: Predicting VIX and US Treasury Yield Changes after a Federal Open Market Committee Statement Release

Team Members: Bryan Beller and Jooho Yu

Project Overview:

       The Federal Open Market Committee (FOMC) is the body of the Federal Reserve that controls monetary policy. They have a meeting about once every six weeks where the Committee's members discuss the US economic outlook and necessary policy decisions. After each meeting a statement is released which provides an overview of the meeting to the public. This statement often causes an increase in volatility in the US equities and treasury markets. We would like to use natural language processing to create a model that predicts VIX performance on a day where the FOMC statement is released. We are primarily interested direction (classification) of the VIX move on that day and over the next few days. We will also explore the short-term effects of FOMC language on US Treasury yields.

Literature Survey:
- *Predicting Market-Volatility from Federal Reserve Board Meeting Minutes NLP for Finance*
  - http://www.cs.cmu.edu/~zollmann/publications/frb-meetings-nlp.pdf
  - This paper uses a modified bag-of-words method to predict "short-term" (3-14 days) and "long-term" volatility after the FOMC statement release.
  - Results are not promising, but they did not attempt to predict the 1-day implied volatility change (VIX) and instead focused on the realized volatility over longer periods.
- *Text Mining Systems for Market Response to News: A Survey*
  - https://pdfs.semanticscholar.org/55a3/d05383a496eb8c19c2ae811d3d64c672166c.pdf
  - This paper compares 7 different prototypes using NLP to predict financial price trends or volatility. Although the prototypes had varied success, there were a lot of similarities between them. They almost all used a 3-category model or Naive Bayes, and used bag of words or tuples combined with AND as their features.
  - The paper recognizes the limitations of using NLP on financial data as NLP is bad with numerical information out of context. They also provided some recommendations to give more weight to relevant words by forcing words into the final set if they appeared in any of the training documents.
- *The Effects of the Content of FOMC Communications on US Treasury Rates*
  - http://www.aclweb.org/anthology/D16-1226
  - The authors used the FOMC statements from 1999 to 2016 to try and predict the corresponding interest rate. This paper provided many key insights as they suggested useful preprocessing steps in addition to looking at the latest macroeconomic data which could be useful in teasing out the baked-in expectation.

- - This paper used the Max-Entropy Discrimination Latent Dirichlet Allocation model build for classification. While the results were very good, they did look at a period of time were rates were mostly 0, so their model could have had good success guessing 0.
  - *On the Importance of Text Analysis for Stock Price Prediction*
    - https://nlp.stanford.edu/pubs/lrec2014-stock.pdf
    - The goal of this study was to create a model that incorporates text analysis of a stock's 8-K with other fundamental/technical features to predict if the stock will go up, down, or stay flat in the near term (relative to SPY). They trained the model using a random forest classifier.
    - They found that their predictive models that incorporated text analysis performed significantly better than the models that did not.
    - Potentially helpful for our preprocessing: "For the linguistic features we used unigram features, first lemmatizing all unigrams and then incorporating a model of negation by marking as a negative every word appearing between a linguistic negation and a clause-level punctuation mark. We then removed any features that occurred fewer than 10 times throughout the training data and used PMI for feature selection to retain 2319 linguistic features."
  - *An Analysis of Verbs in Financial News Articles and their Impact on Stock Price*
    - *verbs_link*
    - The goal of this study was to build a model to predict future stock price based on the verbs used in articles about S&P 500 stocks released during the trading day.
    - Although I'm not convinced in the validity and robustness of his work there are a few interesting things here:
      - He uses a SVR (support vector regression) model to predict what the stock price should be after the article's release
      - He only looks at verbs and treats them as binary features (1 if the verb is in the article and 0 if not).

Plan of Action:

By mid-quarter presentation:
- Continue to do research on what has been done before
  - Learn more about standard preprocessing techniques of economic statements
- Gather data
  - Scrape the web for all FOMC statements for as far back as we can go (1994)
  - Download all necessary VIX and Treasury yield data and match data with statement release dates
    - Open to close VIX and Treasury yield data for day of statement release
    - Close to close data for the following five days
- Preprocess the documents
- Create our buckets for labeling and bucket the FOMC statements

- We want to have 3 buckets: VIX up, VIX down, and neutral (and similar plan for Treasury yields). We plan to create thresholds for "up" and "down" that keep our three buckets close to equal sizes.

By final presentation:
- Train the models
  - Determine our features: most likely we will create a bag-of-words and/or bigrams from the statements
  - We plan on exploring several models for classification including Naive Bayes, Logistic regression, and neural networks
  - We can tune all necessary parameters and evaluate prediction accuracy using cross validation
  - Make revisions to our models or assumptions as necessary
- Evaluate the models
  - See evaluation description below. In short, we will be comparing the prediction accuracy of our models to random guessing
- Determine what future work is needed
  - Is it possible to use this work to actually create a trading strategy?

Project Concerns:

Limited Amount of Data
- FOMC Statements were first released in 1994 and only were released regularly starting in 1999, we will only have roughly 240 statements to test and train our data. Also, the VIX calculation was changed slightly in 2003 and 2014, which may limit our dataset.
- *Current Solution* - While we wish we had more data, we believe 240 is enough to still tease out some interesting conclusions. We could also try using multiple epochs to train our models with more data. For VIX since we are only looking at daily % change it should not be an issue that the calculation changed.

How much information in the FOMC statement was already baked in?
- Due to the desired transparency of the Federal Reserve, they try their best to telegraph their moves beforehand. If the FOMC statement doesn't reveal any new information, then it may not have any market impact even if the language used was extremely positive or negative. The surprise (vs. what is already priced in) of the market is measurable but the baseline is much harder to account for and this may cause some inconsistencies in our model.
- *Current Solution* - Anyone who uses NLP to predict movement of financial instruments will run into this issue to some degree. This is very challenging to account for and we currently have to live with this limitation.

Lack of Access to Time Specific Data
- The FOMC statement is released at 1 p.m. Ideally, we have access to financial data from the moment of the release to effectively gauge its impact and increase its real world applicability. Unfortunately, we do not have access to this data.

- *Current Solution* - We decided to look at open prices for VIX and Treasury Yields from the days the FOMC statement was released and compare it to close prices from same day (1 day) or 4 days after (5 day) to calculate the percentage moves. While this is not perfect, we believe this is a better solution than looking at close-close data because open-close data will not be contaminated by after-hours activity.

Consistent Results after FOMC statement
- Intuitively, it seems like VIX would tend to go down after the FOMC statement as some uncertainty about the Fed's actions is revealed. We could end up with a dataset that has mostly VIX down labels.
- *Current Solution* - After looking at normal VIX market moves and VIX moves after the FOMC statement, we will look to assign partitions for VIX moves to get roughly a ⅓ split amongst our VIX up, VIX down, and VIX unchanged labels.

Evaluation:
After training our models, we will test them with data that our models have not seen and see what proportion of the test set they correctly classify. Because we will have 3 classes that were created to be equally likely to be chosen (⅓ of the time), we will see if our model can do better than random guessing. We will run statistical tests to see if the predictions are statistically significant vs. random guessing. To increase the scope of the project, we will also run evaluations on Treasury yields, as FOMC statements have a major and direct impact on these instruments. In addition to looking at 1-day data (open-close), we will also look at 5 day data to attempt to capture any effect that may take longer than 1 day.

Dividing workload: We plan on dividing the workload evenly and having frequent discussions and meetings. We both live near each other so we hope to meet in person once a week (more or less depending on schedules) and have been sharing our ideas via email. We will share our coding progress via github so that we can both manage the progress of the project.