Executive Summary

Can we predict future market trends based on news media sentiment? Market trends will be extracted using the YFinance API. News media sentiment will be extracted from the NYTimes using the NYTimes API (or possibly the Financial Times depending on availability). Sentiment analysis will be done using natural language processing. We will visualize the relationship between a metric of sentiment analysis vs stock price at time $t - t_{\text{article published}}$, for articles purportedly related to the stock (likely via word match). Following visualization, we will train an ML regression model to the data.

Project Goals

Stakeholders

We identified two primary groups of stakeholders. One group are financial analysts who might profit by predicting market trends. These stakeholders will only benefit if there is indeed a correlation between media sentiment and change in stock price.

The second group of stakeholders are the PR departments of companies whose stock is being tracked. These stakeholders will benefit if there is a correlation between change in stock price and subsequent publication of a media piece. In particular, these PR departments would benefit from an "advanced warning system" that alerts them to the increased likelihood of the publishing of a negative (or positive) article.

We will focus on the second group of stakeholders, the PR departments, because it is well-known how challenging it is to predict the market, and if such an approach was successful it would likely already be employed.

Key Performance Indicators

To discuss the KPIs we define the following variables. An article indexed by i is published at time t_i, has sentiment s_i, and polarity p_i. The total number of articles published within a given range of time, sentiment, and polarity $N(t, T, s, S, p, P) = \sum_i 1$, where $t_i \in [t, T]$, $s_i \in [s, S]$, and $p_i \in [p, P]$. The stock price at time t will be indicated by $y(t)$, and the change in stock price between two times t and $t + \tau$ will be indicated by $\Delta y(t, \tau) \equiv y(t + \tau) - y(t)$.

Defined as such, KPIs include...

- The covariance between p_i and $\Delta y(t_i, \tau)$, i.e. $\sigma(p_i, \Delta y(t_i, \tau))$, where τ is a time-delta hyper-parameter to optimize over to find the maximum absolute covariance. If this covariance is inconsistent with no covariance then it may be possible to predict $\Delta y(t, \tau)$ based on p_i.
- The same metric for s_i.

- The covariance between $\Delta y(t-\tau,\tau)$ and $N(t,T,s,S,p,P)$, i.e. $\sigma(\Delta y(t-\tau,\tau), N(t,T,s,S,p,P))$, where τ, T, s, S, p, and P are hyperparameters. If this covariance is unlikely to be consistent with zero, then it may be possible to predict the publication of articles following a change in stock.
- The extent to which predictions drawing on the above trends are better than random guesses (where the baseline model is poisson distribution fit to the data).

Modeling Approach

Here, we give modeling approaches for both predicting stock trends from the news, as well as, predicting article publications based on the market.

Models relevant to predicting stock trends from the news, and predicting article publications based on the market:

- **Planned:** Extract sentiment using TextBlob, and NLTK
- **Planned:** Extract market data using YFinance API.
- **Planned:** Extract news data from NYTimes API.
- **Planned:** Visualize histogram of stock relative to time of publication.
- **Adapted along the way:** Extract sentiment using FinBERT.

Modeling approaches for predicting stock trends from the news:

- **Planned:** Linear Regression
- **Planned:** Random Forest Regression
- **Planned:** Decision Tree Regression
- **Planned:** SVR
- **Planned:** Gradient Boosting Regressor
- **Adapted along the way:** Improved baseline (current stock)
- **Adapted along the way:** Modeling using difference in stock, not stock itself

Modeling approaches for predicting article publications based on the market:

- **Planned:** Mean baseline
- **Planned:** Poisson baseline
- **Planned:** Linear regression

- **Planned:** Neural net
- **Planned:** Recurrent neural net
- **Planned:** Try both classification and regression versions of the above when possible
- **Adapted along the way:** ARIMA
- **Adapted along the way:** No Stock info