Influence as Predictor in Machine Learning Systems
Appinions Project Proposal

Team Members:
Yi Li (yl2326)
John Chen (cc899)

Statement of Problem and Task:

The topic/influencer graph represents rich information about a topic. We would like to see whether this information can be used predictively. Out task is to determine if the information from the topic/influencer graph is statistically significant in predictions. Our primary topic we will explore are stock prices, but we will also look into other topics such as market share of products, etc.

Data Sources and Resources:

We plan on using the Bloomberg terminals to get our stock data from Bloomberg. We plan on using the Appinions platform and API to get data for our topic/influence graph. We will research and talk with Appinions client services in getting additional data for other topics.

General Approaches:

We will conduct several different experiments to evaluate the relationship between the influence data and target function throughout the whole project. We have chosen stock price to be our first and primary target function since it's more accessible and meaningful. We would use different approaches throughout the development of the project. First, we will use stock data in previous several days to predict the stock price movement (a binary classification) and closing price on the next day. Our training data includes the market data of the top five companies of three major sectors. And our training and test data would be the stock movement and closing price on the next day. Second, we will add the parsed influence data into the predictor and see if the prediction accuracy improves. Several baseline machine learning techniques, such as K-means, SVM and ID3 decision tree would be used in this process. The result would be evaluated via several statistics test including McNemar's test, z-test or t-test. Third, we would experiment some advanced machine learning technique that fit more with the data to see if any improvement can be made. If time permits, we would also apply the same machine learning algorithms on some other target functions, such as product sales and market growth rate.

Milestones:

The project is divided into five different phases and would incrementally explore the relationship between the influence data and the target function we try to predict.

Phase I: Data Collection and Parsing

In this phase we will collect the stock data and influence data needed for the following steps and parse them into a format that facilitates machine learning algorithms. We plan to collect the stock price of the top five companies in each of three major sectors from Bloomberg since we suspect they are most prone to top influencers' opinions. At the same time, we would collect and parse the influence data from the Appinions API.

Phase II: Stock Price Prediction without Influence Data

This is the baseline approach will be compared later with the Influencer prediction. The stock prices in the previous several days will be used to predict the price movement and closing price on the next day.

Phase III: Stock Price Prediction with Influence Data

At this step, the influence data would be added into the feature set to predict the price movement and closing price the next day. The general machine learning skills such as KNN, K-means, SVM and decision trees would be applied. The result would be compared and evaluated using McNeymar's test, z-test or t-test. We would have a basic idea of whether the influencer data would be significant with respect to the stock prices by then.

Phase IV: Stock Price Prediction with Advanced Algorithm

We would explore some more advanced machine learning skills to see if any improvement can be made on the prediction. We would choose the algorithms that fit more with the time series as the feature set since the stock data are all related to a point in time. We would also evaluate the result using the same tests as in phase III.

Phase V: Predictions on Other Target Functions

Once finished the predictions with stock prices, we would explore the relationship between influencer data and other target functions. The potential targets are product sales and market share growth across companies. But this step depends on whether we would be able to find a good data resource of the relevant data.

Schedule:

2/4
- Complete the final draft of proposal
- Speak with Professor Cardie to complete our MEng forms

2/9
- Milestone 1: Gathering Stock Data is completed
- K-Means and ID3 Decision Tree implementations are complete

2/16
- SVM implementation is complete, begin making test runs
- Get Influence Data, Begin Parsing for Learning Usage

2/23 (switch)
- Run our baseline algorithms through the stock data
- Continue Parsing Influence Data; Complete a Rough Sketch of Data

3/2
- Run all baseline algorithms with the stock-only based predictors
- Parse all Influence Data, Begin Testing with Influence Data

3/9
- Complete 2/3 Baseline Algorithms Tests with the Influence Data

3/16
- Run all baseline algorithms with the all predictors

3/23
- Re run all baseline algorithms, become computing statistical models

3/30
- Finish computing Statistical Models
- Look at new target function and design predictors
- Begin acquiring data for the new target function

4/7
- Complete acquiring data
- Begin running target function on old algorithms, modifying them if necessary
- Look at implementing several advanced ML algorithms

4/14
- Complete running target function on old algorithms

- Complete the advanced ML algorithms

4/21
- Finish running the advanced ML algorithm
- Begin analysis, continue testing and tuning if necessary
- Begin making statistical models

4/28
- Continue testing/tuning
- Complete statistical models
- Begin final report

5/5
- Complete testing/tuning
- Finish Final Report

4) Evaluation Methodologies

We plan to evaluate our results using the accuracy of our predictions are our main indicator. We also plan to run McNemar’s, paired t-test, and z-tests to evaluate the statistical significance of our results.