Twitter Trader - Correlating Stocks with Twitter Activity

Curtis Ullerich, Daniel Stiner, Brandon Maxwell
Department of Computer Science and Engineering
Iowa State University Ames, Iowa, USA
{curtisu,stiner,bmaxwell}@iastate.edu

Abstract

Twitter is a popular source for data mining due to its massive scale and inclusivity of current trends. Similar studies into the correlation of overall Twitter sentiment (with respect to positivity) to the Dow Jones Industrial Average have been conducted before[1]. We propose using real-time Twitter sentiment ratings of specific companies to determine buy-sell-hold actions of small, short-term investments.

We will develop a system to aggregate tweets about a set of companies using a classifier. Each tweet about a company will be labeled with a sentiment positive, negative, neutral. We hope to use the classifications and confidences to provide a heuristic for potential investments by analyzing the sentiment of tweets about a company over time. We may augment this heuristic by incorporating recent trends in stock prices and/or Twitter sentiment.

Introduction

Goals

- Accurately classify the topicality of a tweet with respect to a particular company
- Accurately classify the sentiment of a tweet in the categories of positive, negative, neutral, and irrelevant
- Discover whether a correlation exists between sentiment of tweets about a particular company and the stock price of that company in the near future
- Make lucrative buy-sell-hold decisions using these correlations, if they exist

Anticipated Results

Based on prior research by J. Bollen, H. Mao, and X.-J. Zeng[1], we expect to find a small but exploitable correlation between the volume and positivity of tweets about a company, and its stock price one to three days in the future. This is based on a close examination Table 8 from Ruiz and Hristidis which shows not only a significant correlation of 0.12 between activity and stock price the same day for the DEG.-STD, but smaller correlations of 0.10 and 0.07 for one day and two days in the future respectively. At three days the correlation is -0.04 and is no longer significant.

We do not expect these correlations to exist for all companies. Given the tendency of Twitter users to be tech-friendly and young, the companies about which they tweet and their attitudes toward them will be varied. Additionally, the sheer volume of tweets about popular companies like Google and Microsoft have the potential to overshadow the number of tweets about Garmin (for instance) and other companies in our dataset. We do expect to find useful correlations for some profile of companies. The profitability of our decision process is easily verified using past data, though it will take time to accumulate enough data for this to be a solid recommendation. If the discovered correlation is not strong enough to predict future stock price, we will change our analysis to find correlation between Twitter activity and traded volume for the company stocks, a relationship that Ruiz and Hristidis[1] found to have a significantly higher correlation in their dataset than activity and stock price.

Related Work

System Breakdown

Corpus creation
Data aggregation and classification
Decision-making

Basic Approach

1) Develop classifiers for both sentiment and company-topicality
We do all machine learning with the Mallet[2] library developed by the University of Massachusetts. We are using an existing corpus from Sanders Analytics[3] to bootstrap our sentiment classifier. We created a very small corpus (500 total instances) of company-topical tweets by scraping the Twitter live stream based on keywords and verifying topicality by manual analysis. We trained a Mallet Classifier for each domain, specialized to accept Twitter's native JSON tweet format for classifying new instances. We have been collecting (hundreds of thousands of) tweets in a database to use as our corpus. We developed a basic web interface to efficiently human-verify the correct classification a large number of tweets as classified by our bootstrapped model. As we add more human-verified instances to the corpora, the accuracy of our models will improve over time.

A major component of this is corpus creation. Using Twitter's streaming API, we pull for tweets based on keyword. Using our trained models, we classify each tweets with the highest-confidence company and sentiment labels, including the confidences. We created a simple web interface for accessing the database in order to human-verify classifications for insertion into the database. This page displays a random classified tweet and asks the user if the tweet is correctly classified.

2) Develop web interface for tracking sentiment and volume of tweets

We plan to filter the Twitter live stream for all tweets about the following companies:

Amazon Boeing Cerner Coke Costco Dollar Tree eBay Ford Garmin Google HyVee Microsoft Monsanto Netflix Starbucks Travel Zoo Verizon Walmart Xerox

"About" here is as classified by our models. We will classify each tweet's sentiment and store the statistics for 6-hour intervals. We will provide an interface for viewing this sentiment data graphically over time, as compared to stock market prices and the volume of tweets about each company.

3) Develop decision-making system for buy-sell-holding stocks

Based on the aggregated data, we will determine the buy-sell-hold decision that maximizes the likelihood of profiting.

Application

Results

Acknowledgments

[1] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. Journal of Computational Science, abs/1010.3003, 2010

[2] Mallett, University of Massachussetts Department of Computer Science

[3] Sanders Analytics (http://www.sananalytics.com/lab/twitter-sentiment/)