Proposal: Twitter sentiment classification and its application in trading

Xinye Xu (xinyexu)

Datasets:

1) Sentiment140 allows you to discover the sentiment of a brand, product, or topic on Twitter. (Twitter Dataset). Following is an example of the data file, which begins on April 7 to April 17 in minutes, containing 6 fields:

Polarity	ID of the tweet	Date	Query	User that tweeted	Text of the tweet
0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D

The polarity of the tweet (0 = negative, 2 = neutral, 4 = positive) is sentiment baseline. Analysis by Twittratr website using basic searching keywords method.

2) Stock Price: IBM price by minutes during same dates, scraping from Kibot.

Exploratory questions:

1) How to process text information into categories and label them with sentiments?
2) Did people more prefer to show their negative attitude in the Twitter?
3) Is there a correlated relationship between Twitter sentiment and the stock market?
4) How to utilize the sentiment data from Twitter to predict stock index movement and test its effectiveness?

Analysis/Visualization Methods:

1) Texts can be classified by natural language processing. More accurate sentiments classification methods: Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM).
2) Boxplots, Distribution density histogram, linear model with respect to time.
3) Correlation plots, Heat plot, Basic Linear Regression with Stock index returns.
4) Fama-French Factors model, Neural Network Regression, SVM, Boosted Regression Tree; Random Forest Regression. The moving window period cross validation can be used for time series data.