Alternative Data Sources for Measuring Market Sentiment and Events (Using R)

Joe Rothermich, CFA
Natural Selection Financial, Inc.

R/Finance 2011
Apr 29, 2011
Chicago, IL
**The Hemline Index**

**Wikipedia:** “The theory suggests that hemlines on women's dresses rise along with stock prices. In good economies, we get such results as miniskirts (as seen in the 1960s), or in poor economic times, as shown by the 1929 Wall Street Crash, hems can drop almost overnight.”

The Danceability Index

**Theory:** Dancing without moderation suggests a peak of irrational exuberance. When dancing reaches a maximal level, the market may be overvalued.
Data Sources

- last.fm: Weekly Historical Top 5 Songs in New York City
- The Echo Nest: Danceability Score of Top Songs
Analysis — Danceability Index

Danceability of Top Songs

Echo Nest Danceability Score (5 week moving average)

Analysis — S&P 500 Returns
Analysis — Returns vs. Danceability

Returns by Danceability Intervals

Mean Danceability Score

(0.4,0.45] (0.45,0.5] (0.5,0.55] (0.55,0.6] (0.6,0.65] (0.65,0.7]
PART 2: OTHER SOURCES
Data Trends

• Big
• Open
• Structured
• Accessible
R Packages

- XML
- RCurl
- twitteR
- RGoogleTrends
- Infochimps
- tm
- etc.
Sentiment Analysis

- Insider trading
- News
- Social media

Context and volume are important

\[
\sqrt{\heartsuit} = ? \quad \cos \heartsuit = ?
\]

\[
\frac{d}{dx} \heartsuit = ? \quad \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \heartsuit = ?
\]

\[
F\{\heartsuit\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{it\heartsuit} dt = ?
\]

My normal approach is useless here.

http://xkcd.com/55/
Twitter Mood

Soon to be a hedge fund...

Twitter mood predicts the stock market.
Johan Bollew\textsuperscript{a}, Huina Mao\textsuperscript{b,c}, Xiao-Jun Zeng\textsuperscript{c}.
\textsuperscript{a} authors made equal contributions.

Abstract—Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e., can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJI) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinioFinder that measures positive vs. negative mood and Google-Profile of Mood State (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2006. A Granger causality analysis and a Self-Organizing Feature Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinioFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 87.6\% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 5\%.

Index Terms—stock market prediction — twitter — mood analysis.

I. INTRODUCTION

STOCK market prediction has attracted much attention from academia as well as business. But can the stock market really be predicted? Early research on stock market prediction [10], [11] was based on random walk theory and the Efficient Market Hypothesis (EMH) [9]. According to the EMH, stock market prices are largely driven by new information, i.e., news, rather than present and past prices. Since news is unpredictable, stock market prices will follow a random walk pattern and cannot be predicted with more than 50\% accuracy [11].

There are two problems with EMH. First, numerous studies show that stock market prices do not follow a random walk and can indeed to some degree be predicted [12], [13], [14], thereby calling into question EMH's basic assumptions. Second, recent research suggests that news may be unpredictable but that very early indicators can be extracted from online social media (blogs, Twitter feeds, etc.) to predict changes in various economic and commercial indicators. This may conceivably also be the case for the stock market. For example, [17] shows how online chat activity predicts book sales, [18] users assessments of blog sentiment to predict movie sales, and [15] predicts future product sales using a Probabilistic Latent Semantic Analysis (PLSA) model to extract indicators of sentiment from blogs. In addition, Google search queries have been shown to provide early indicators of disease infection rates and consumer spending [14], [19] investigates the relations between breaking financial news and stock price changes. Most recently [20] provides a groundbreaking demonstration of how public sentiment related to movies, as expressed on Twitter, can actually predict box office receipts.

Although news most certainly influences stock market prices, public mood states or sentiment may play an equally important role. We know from psychological research that emotions, in addition to information, play a significant role in human decision-making [21], [22], [23]. Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood [24]. It is therefore reasonable to assume that the public mood and sentiment can drive stock market values as much as news. This is supported by recent research by [25], who extract an indicator of public anxiety from LiveJournal posts and investigate whether its variations can predict S&P500 values.

However, if it is our goal to study how public mood influences the stock markets, we need reliable, scalable and early assessments of the public mood at a time-scale and resolution appropriate for practical stock market prediction. Large surveys of public mood over representative samples of the population are generally expensive and time-consuming to conduct, cf. Gallup's opinion polls and various consumer and well-being indices. Some have therefore proposed indirect assessment of public mood or sentiment from the results of soccer games [26] and from weather conditions [27]. The accuracy of these methods is however limited by the low degree to which the chosen indicators are expected to be correlated with public mood.

Over the past 5 years significant progress has been made in sentiment tracking techniques that extract indicators of public mood directly from social media content such as blog content [28], [29], [30], [31] and in particular large-scale Twitter feeds [22]. Although each so-called \textit{tweets}, i.e. an individual user post, is limited to only 140 characters, the aggregate of millions of tweets submitted to Twitter at any given time may provide an accurate representation of public mood and sentiment. This has led to the development of real-time sentiment-tracking indicators such as [17] and "Pulse of Nation" [32].

In this paper we investigate whether public sentiment, as expressed in large-scale collections of daily Twitter posts, can be used to predict the stock market. We use two tools to measure variations in the public mood from tweets submitted...
...and it’s been getting a lot of attention.
Buzz, News Volume, and Headlines

Can signal a change in trends
Search for tweets with the phrase “Stock Market”, then plot frequency of tweets per hour:

```
library(twitteR)

tweets <- searchTwitter("Stock Market", n=1500)

times <- sapply(tweets, function(x) format(x@created, "%b %d %H:00"))

users <- sapply(tweets, function(x) x@screenName)

times <- times[!duplicated(users)]  # removing duplicate users to avoid spammers and news

counts <- table(times)

bp <- barplot(counts, main="Counts of 'Stock Market' Tweets by Hour", col="lightblue", border=NA, ylim=c(0,200))

lines(spline(counts ~ bp), lwd=3, lty="dashed", col="darkblue")
```
Trend of “Stock Market” tweets

Counts of 'Stock Market' Tweets by Hour

Apr 09 06:00  Apr 09 07:00  Apr 09 08:00  Apr 09 09:00  Apr 09 10:00  Apr 09 11:00  Apr 09 12:00
The Hathaway Effect*

“How Anne gives Warren Buffett a Rise”

Oct. 3, 2008 - Rachel Getting Married opens: 
BRK.A up .44%

Jan. 5, 2009 - Bride Wars opens: 
BRK.A up 2.61%

Feb. 8, 2010 - Valentine's Day opens: 
BRK.A up 1.01%

March 5, 2010 - Alice in Wonderland opens: 
BRK.A up .74%

Nov. 24, 2010 - Love and Other Drugs opens: 
BRK.A up 1.62%

Nov. 29, 2010 - Anne announced as co-host of the Oscars: 
BRK.A up .25%

*I don’t really believe this is true

http://www.huffingtonpost.com/dan-mirvish/the-hathaway-effect-how-a_b_830041.html
Search Volume with Google Trends

Search volume seems to be a close proximity to news volume
library(RGoogleTrends)
library(quantmod)

ans = getGTrends("Anne Hathaway")
trend <- xts( ans$Week$anne.hathaway, order.by=as.Date(ans$Week$Week, "%b %d %Y"))
brk <- getSymbols("BRK-A", auto.assign=F, from = "2004-01-01" )
x <- na.locf(merge(trend, Cl(brk)))[index(trend)]

#plot returns and search volume
par(mfrow=c(2,1))
plot(x$trend, main="Google Trends: Anne Hathaway", col="blue")
plot(x$BRK.A.Close, main="Berkshire Hathaway Share Price", col="red", cex=.7)

#evaluate returns by search volume
x$Return <- Delt(x$BRK.A.Close)
breaks <- cut(x$trend, seq(0,10, 1))
boxplot(as.numeric(x$Return) ~ breaks, ylab="Weekly Return", xlab="Search Volume", col="lightblue", border="darkblue" )
abline(h=0, col="blue")
Returns Analysis

Doesn’t hold up with search volume

(at least on a weekly basis, and you wouldn’t expect short term spikes to last)
Popular Data Sets (Browse All Data sets)

- Word List - 350,000+ Simple English Words (with Definitions, Excel format)
- Word List - 100,000+ official crossword words (Excel readable)
- GeoNames.org Postal Code files - US Zip Code Geolocations
- Crime Rates by State, 2004 and 2005, and by Type, 2005 (Cleaned up version)
- National Center for Educational Statistics (NCES): Tables and Figures
- Retrosheet: Transactions in Major League Baseball (Trade, Signing, Draft, etc.)
- Daily 1970-2010 Open, Close, Hi, Low and Volume (NYSE exchange)
- Social Security-Beneficiaries, Annual Payments, and Average Monthly Benefit, and by State and Other
- 90,000+ Documented UFO Sightings With Text Descriptions And Metadata

Popular Tags (Browse All Tags)

- Government
- Social
- Music
- Locations
- Economics
- Chemistry
- Zipcode
- Pollution
- Health
- Law
- Sports
- Statistics
- Survey
- Language
- Spending
- Income
- Word
- Age
- Football
- Death
- Demographics
- Character
- Geonames
- Longitude
- Latitude
- Maps
- Science
- Census
- National
- Housing
- Employment
- Corpora
- Population
- Commodities
- Twitter
- Size-large

• Word frequencies from the British National Corpus
• Twitter influence
• Wikipedia abstracts

• Yahoo stock prices
• Wikipedia abstracts
• etc.
Thank you

joe.rothermich@gmail.com