QUANTITATIVE TRADING

QUANTITATIVE INVESTMENT AND TRADING IDEAS, RESEARCH, AND ANALYSIS.

THURSDAY, SEPTEMBER 07, 2017

StockTwits Sentiment Analysis

By Colton Smith
===

Exploring alternative datasets to augment financial trading models is currently the hot trend among the quantitative community. With so much social media data out there, its place in financial models has become a popular research discussion. Surely the stock market's performance influences the reactions from the public but if the converse is true, that social media sentiment can be used to predict movements in the stock market, then this would be a very valuable dataset for a variety of financial firms and institutions.

When I began this project as a consultant for QTS Capital Management, I did an extensive literature review of the social media sentiment providers and academic research. The main approach is to take the social media firehose, filter it down by source credibility, apply natural language processing (NLP), and create a variety of metrics that capture sentiment, volume, dispersion, etc. The best results have come from using Twitter or StockTwits as the source. A feature of StockTwits that distinguishes it from Twitter is that in late 2012 the option to label your tweet as bullish or bearish was added. If these labels accurately capture sentiment and are used frequently enough, then it would be possible to avoid using NLP. Most tweets are not labeled as seen in Figure 1 below, but the percentage is increasing.

Year	Bearish	Bullish	None
2011	1.3%	3.9%	94.8%
2012	1.2%	3.4%	95.4%
2013	2.9%	10.1%	87.0%
2014	3.6%	16.4%	80.0%
2015	4.2%	16.6%	79.2%
2016	4.0%	17.8%	78.2%

Figure 1: Percentage of Labeled StockTwits Tweets by Year

This blog post will compare the use of just the labeled tweets versus the use of all tweets with NLP. To begin, I did some basic data analysis to better understand the nature of the data. In Figure 2 below, the number of labeled tweets per hour is shown. As expected there are spikes around market open and close.

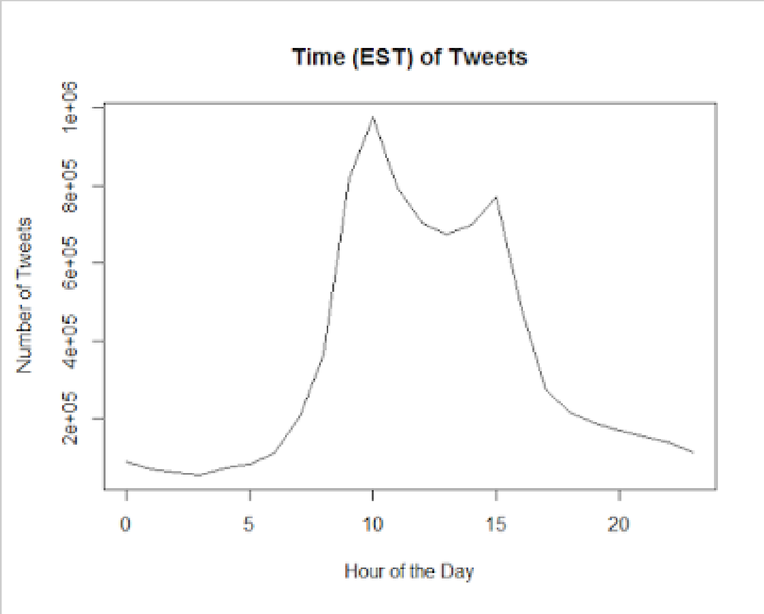

Figure 2: Number of Tweets Per Hour of the Day

The overall market sentiment can be estimated by aggregating the number of bullish and bearish labeled tweets each day. Based on the previous literature, I expected a significant bullish bias. This is confirmed in Figure 3 below with the daily mean percetage of bullish tweets being 79%.

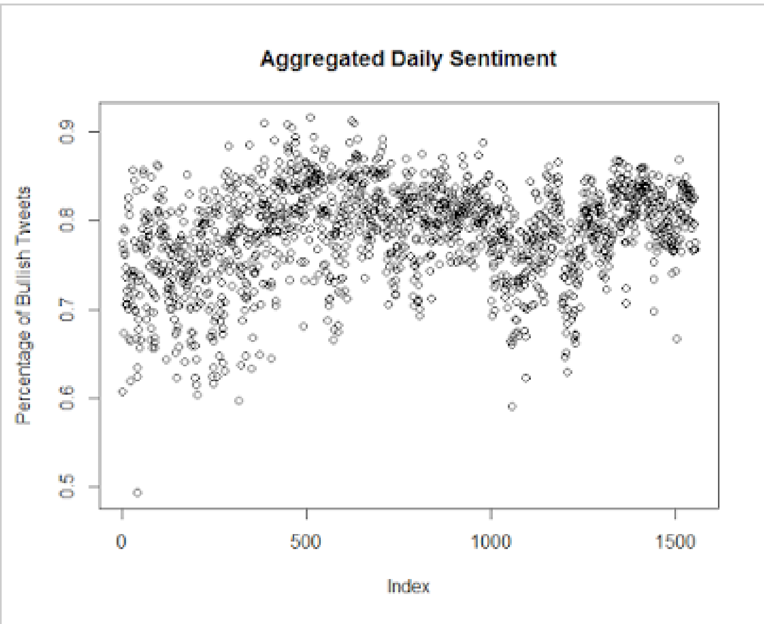

Figure 3: Percentage of Bullish Tweets Each Day

When writing a StockTwits tweet, users can tag multiple symbols so it is possible that the sentiment label could apply to more than one symbol. Tagging more than one symbol would likely indicate less specific sentiment and predictive potential so I hoped to find that most tweets only tag a single symbol. Looking at Figure 4 below, over 90% of the tweets tag a single symbol and a very small percentage tag 5+.

MY TRADING WORKSHOPS

Algorithmic Options Strategies

PARTNER CENTER

SUBSCRIBE TO MY BLOG

Enter your Email

Subscribe me!

TWITTER

Tweets by @chanep

RECOMMENDED BOOKS

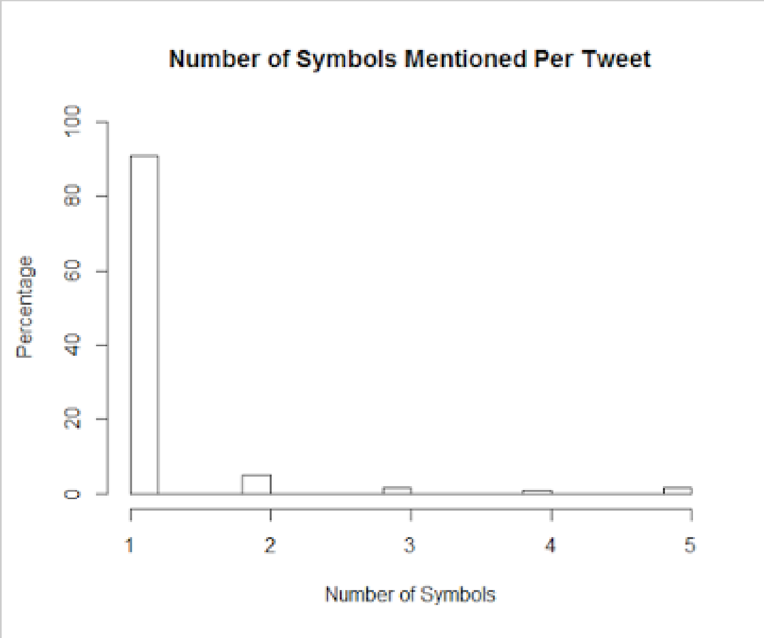

Figure 4: Relative Frequency Histogram of the Number of Symbols Mentioned Per Tweet

The time period of data used in my analysis is from 2012-11-01 to 2016-12-31. In Figure 5 below, the top symbols, industries, and sectors by total labeled tweet count are shown. By far the most tweeted about industries were biotechnology and ETFs. This makes sense because of how volatile these industries are which hopefully means that they would be the best to trade based on social media sentiment data.

Top 20 Symbols	
AAPL	406267
SPY	227364
BBRY	159017
FB	126701
MNKD	118718
MGT	118350
TWTR	116536
TSLA	114589
SUNE	110513
PLUG	102493
GPRO	93285
GEVO	82891
VRX	82856
NFLX	77602
AMD	73807
GBSN	73155
KNDI	69334
ES_F	61443
BABA	59537
DRYS	57304

Top 10 Industries	
Biotechnology	834560
Exchange Traded Fund	601421
Personal Computers	406653
Internet Information Providers	402848
Auto Manufacturers - Major	198966
Diversified Communication Services	177222
Drug Manufacturers - Other	172978
Semiconductor - Specialized	149869
Semiconductor - Integrated Circuits	142247
Diversified Electronics	140418

Top 5 Sectors	
Technology	1932150
Healthcare	1429825
Financial	816114
Services	756177
Basic Materials	590188

Figure 5: Top Symbols, Industries, and Sectors by Total Tweet Count

Now I needed to determine how I would create the sentiment score to best encompass the predictive potential of the data. Though there are obstacles to trading an open to close strategy including slippage, liquidity, and transaction costs, analyzing how well the sentiment score immediately before market open predicts open to close returns is a valuable sanity check to see if it would be useful in a larger factor model. The sentiment score for each day was calculated using the tweets from the previous market day's open until the current day's open:

S-Score = (#Bullish-#Bearish)/(#Bullish+#Bearish)

This S-Score then needs to be normalized to detect the significance of a specific day's sentiment with respect to the symbol's historic sentiment trend. To do this, a rolling z-score is applied to the series. By changing the length of the lookback window the sensitivity can be adjusted. Additionally, since the data is quite sparse, days without any tweets for a symbol are given an S-Score of 0. At the market open each day, symbols with an S-Score above the positive threshold are entered long and symbols with an S-Score below the negative threshold are entered short. Equal dollar weight is applied to the long and short legs. These positions are assumed to be liquidated at the day's market close. The first test is on the universe of equities with previous day closing prices > $5. With a relatively small long-short portfolio of ~250 stocks, its performance can be seen in Figure 6 below (click on chart to enlarge).

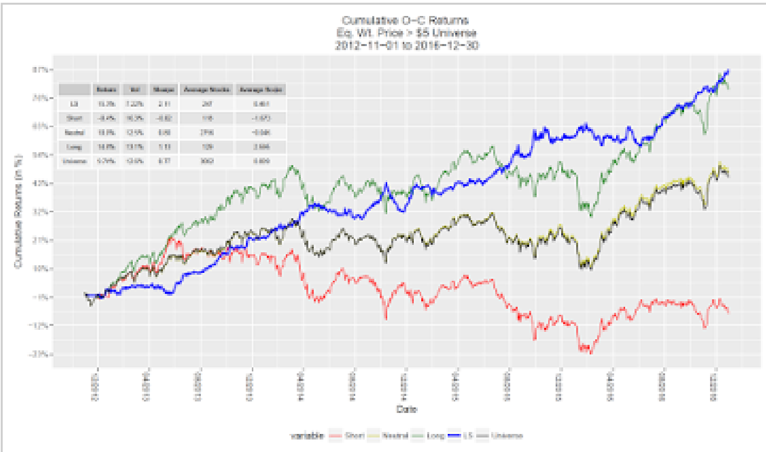

Figure 6: Price > $5 Universe Open to Close Cumulative Returns

The thresholds were cherry-picked to show the potential of a 2.11 Sharpe Ratio but the results vary depending on the thresholds used. This sensitivity is likely due to the lack of tweet volume on most symbols. Also, the long and short thresholds are not equal in an attempt to maintain roughly equal number of stocks in each leg. The neutral basket contains all of the stocks in the universe that do not have an S-Score extreme enough to generate a long or short signal. Using the same thresholds as above, the test was ran on a liquidity universe which is defined as the top quartile of 50-day Average Dollar Volume stocks. As seen in Figure 7 below, the Sharpe drops to a 1.24 but is still very encouraging.

LINKS

QTS Managed Accounts
RSS Site Feed
Quantitative Research & Trading
Quantocracy
Quant News
Insider Monkey
FactorWave
Eran Raviv (Quant Analyst at large pension fund)
Advertise With Us
High frequency historical data
MATLAB automated trading course

BLOG ARCHIVE

▶ 2018 (1)
▼ 2017 (5)
▶ November (1)
▼ September (1)
StockTwits Sentiment Analysis
▶ July (1)
▶ May (1)
▶ March (1)
▶ 2016 (4)

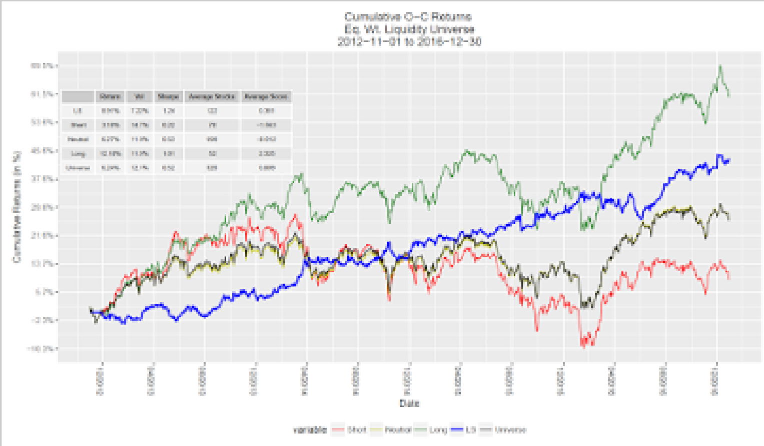

Figure 7: Liquidity Universe Open to Close Cumulative Returns

The sensitivity of these results needs to be further inspected by performing analysis on separate train and test sets but I was very pleased with the returns that could be potentially generated from just labeled StockTwits data.

In July, I began working for Social Market Analytics, the leading social media sentiment provider. Here at SMA, we run all the StockTwits tweets through our proprietary NLP engine to determine their sentiment scores. Using sentiment data from 9:10 EST which looks at an exponentially weighted sentiment aggregation over the last 24 hours, the open to close simulation can be ran on the price > $5 universe. Each stock is separated into its respective quintile based on its S-Score in relation to the universe's percentiles that day. A long-short portfolio is constructed in a similar fashion as previously with long positions in the top quintile stocks and short positions in the bottom quintile stocks. In Figure 8 below you can see that the results are much better than when only using sentiment labeled data.

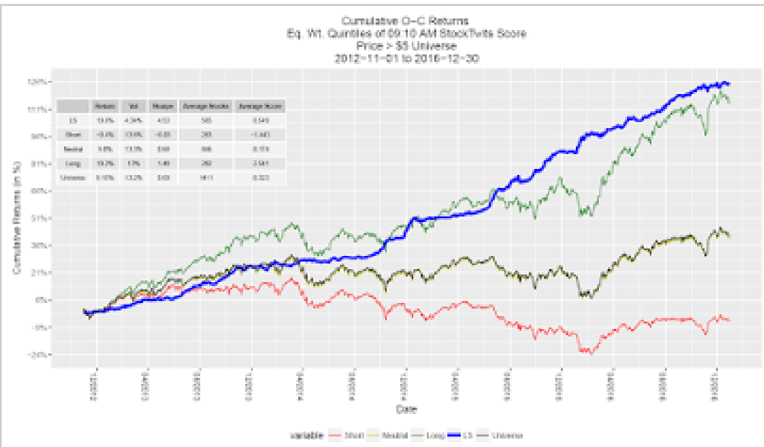

Figure 8: SMA Open to Close Cumulative Returns Using StockTwits Data

The predictive power is there as the long-short boasts an impressive 4.5 Sharpe ratio. Due to having more data, the results are much less sensitive to long-short portfolio construction. To avoid the high turnover of an open-to-close strategy, we have been exploring possible long-term strategies. Deutsche Bank's Quantitative Research Team recently released a paper about strategies that solely use our SMA data which includes a longer-term strategy. Additionally, I've recently

▶ 2015 (7)
▶ 2014 (8)
▶ 2013 (11)
▶ 2012 (12)
▶ 2011 (15)
▶ 2010 (17)
▶ 2009 (32)
▶ 2008 (28)
▶ 2007 (50)
▶ 2006 (24)

developed a strong weekly rebalance strategy that attempts to capture weekly sentiment momentum.

Though it is just the beginning, my dive into social media sentiment data and its application in finance over the course of my time consulting for QTS has been very insightful. It is arguable that by just using the labeled StockTwits tweets, we may be able to generate predictive signals but by including all the tweets for sentiment analysis, a much stronger signal is found. If you have questions please contact me at coltonsmith321@gmail.com.

Colton Smith is a recent graduate of the University of Washington where he majored in Industrial and Systems Engineering and minored in Applied Math. He now lives in Chicago and works for Social Market Analytics. He has a passion for data science and is excited about his developing quantitative finance career. LinkedIn: https://www.linkedin.com/in/coltonfsmith/

===

Upcoming Workshops by Dr. Ernie Chan

September 11-15: City of London workshops

These intense 8-16 hours workshops cover Algorithmic Options Strategies, Quantitative Momentum Strategies, and Intraday Trading and Market Microstructure. Typical class size is under 10. They may qualify for CFA Institute continuing education credits.

November 18 and December 2: Cryptocurrency Trading with Python

I will be moderating this online workshop for Nick Kirk, a noted cryptocurrency trader and fund manager, who taught this widely acclaimed course here and at CQF in London.

POSTED BY ERNIE CHAN AT 7:28 AM

4 COMMENTS:

stephen ashley said...

This comment has been removed by the author.
FRIDAY, SEPTEMBER 8, 2017 AT 11:43:00 PM EDT

Michael Harris said...

This is very interesting but two things are missing from the analysis in my opinion: out-of-sample tests and corrections for data snooping. Obviously, if one tortures the data long enough, a high Sharpe will be obtained but the statistical significance will be very low. In addition, I am not sure what commission rate was used for the analysis. Another other problem is that there may not be ability to short some names due to lack of inventory.

SUNDAY, SEPTEMBER 24, 2017 AT 3:39:00 PM EDT

Ken said...

Yep, very interesting article. Any idea if there is any social market datas on Agricultural commodities ?
I am really surprised that it's working so easily. I would have imagined that being contrarian to social market datas could have been good. Average investors has a very poor track record of stock picking. How the average investor could have a very poor track record and a nice predictive power in their twitt ?
Any idea ?

MONDAY, SEPTEMBER 25, 2017 AT 8:48:00 AM EDT

Colton Smith said...

Michael - Yes, the analysis using only the labeled StockTwits tweets lacks out-of-sample testing and is heavily data snooped. Using NLP on all of the tweets is much more robust as seen in the portion of the post using SMA data, which is entirely out-of-sample. The data used is exactly what our clients were seeing at that point in time; nothing is ever recalculated. To exemplify its statistical significance we like to look at the quintile/decile views (https://socialmarketanalytics.wordpress.com/2017/03/10/decile-spreads-for-twitter-stocktwits/). As explained in the post, the open to close simulation is ran merely as a sanity check because of implementation obstacles including transaction costs and hard-to-borrow stocks like you mention. Some of the best, implementable results come from incorporating sentiment data into a multi-factor model. If you're interested, please email me and I can send you third-party and internal research on our data.

Ken - Yes, we cover US Equities, ETFs, Futures, FX, and most recently cryptos. If you take the entire firehose from StockTwits or Twitter there would be a lot of extra garbage. Before applying any NLP, we filter the firehose on our investor universe which ensures we are only evaluating tweets from reputable sources and maximizing the predictive power of our data. If you have any more questions, please email me.

MONDAY, SEPTEMBER 25, 2017 AT 10:55:00 AM EDT

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)