Automated Market Sentiment Analysis of Twitter for Options Trading

Rowan Chakoumakos, Stephen Trusheim, Vikas Yendluri
{rowanc, trusheim, vikasuy}@stanford.edu

Abstract
We implemented predictive classifiers that combine economic analysis of stocks with features based on natural-language processing of Twitter comments related to each stock in a specified portfolio to enable options-straddling stock trading strategies. Identical SVM models built on this combination set of features showed an average improvement of 35.4 percentage-points over economic analysis features alone, enough to make an options-straddling strategy profitable for some stocks in the portfolio. These results show that information gleaned from Twitter comments can be predictive of major stock price movements for individual stocks.

Introduction
Stock trading is an open opportunity for applications of machine-learning algorithms. Of particular interest is a trading strategy called "options straddling," which allows a trader to profit by betting that a stock will have a large price movement, regardless of the direction of change.

Most public trading strategies use economic models of individual stocks to predict days when a stock will have a "major price difference" (MPD) event (i.e., when the opening price of a stock changes by >= 2% of its previous opening value). We propose that these standard economic models could be improved by automated sentiment analysis of Twitter.

This approach builds upon work by Bollen et al. [1], which showed the general sentiment reflected in Twitter posts was predictive of the the daily closing direction of the Dow Jones Industrial Average (i.e., the entire market) with 86.5% accuracy. We expand this work to focus on a specific small portfolio of stocks; we further expand it to become more directly applicable to an options-straddling strategy by predicting when MPD events will occur for all stocks in our portfolio, instead of predicting their future closing direction.

Methodology

Choice of stocks

We chose a portfolio of 10 stocks from the NASDAQ stock exchange to test our classification. Based on previous work [11], which suggests that quantitative analysis of public sentiment is most useful with a large corpus of Tweets, we selected consumer technology manufacturers and consumer retailers - the types of companies that we thought would be frequently discussed on Twitter: ATVI, ADBE, AAPL, DELL, ERTS, GRMN, MAT, NFLX, RIMM, and URBN.

Data: Training and Testing

Twitter data were acquired through a corpus provided by Yang & Leskovec [10], which contained 20-30% of tweets daily from June-December 2009. Standard stock data (open/close/high/low/volume) were acquired from Yahoo! Finance [9] for every trading-day in the period January-December 2009. All models were trained from June 1 – October 29 2009 (107 trading-days) and tested from October 30 – December 31 2009 (41 trading-days). 30.3% of stock-days in the training period had MPD events, and 26.6% of stock-days in the testing period had MPD events.

Metrics of Success

Because an options-straddle trading strategy necessitates trading only on days when there will be a MPD event, accuracy is not a meaningful statistic — we only need high precision when predicting MPD events. For that reason, our primary metric of success was high precision and reasonable recall when predicting MPD events. Informal analysis of profitability of trading strategies suggests that approximately 55% precision when predicting MPD events could lead to a 1% expected return per option straddle.

Development Methodology

We created classifiers in three phases, based on the framework suggested by Pham, Chien & Lim [2]:

1. Development of a baseline classifier that uses only economic data to predict future MPD events;

2. Development of classifiers based on analysis of Twitter data; and

3. Development of combination classifiers that use both sources of data.

Development and Results

1. Baseline Classifier: Replicating Bollen et al.

As a simple sanity check, we first wanted to replicate the baseline model ("I0") found in Bollen et al. A regression SVM fit using the SMO method with a RBF kernel and a complexity parameter of 0.0001 resulted in 76.7% accuracy in predicting the future close directions of AAPL stock based on the past three days of close prices. These results are within a 10% margin of error of the results found in Bollen, which used a self-organizing fuzzy neural network.

This result suggested the applicability of non-linear SVMs to the problem of predicting stock events, a heartening first step.

2. Baseline Classifier: MPD events

Based on the success of the replicated Bollen baseline classifier, we constructed a classifier specifically to predict MPD events. We constructed 149 features based on standard technical analysis methods [8] in an attempt to provide as much information as possible based on the simple data available to us.

We cross-validated multiple models and model parameters on the dataset created by this baseline classifier, and found that logistic SVM models built by the SMO algorithm using an RBF kernel, with gamma=0.01 and c=2 resulted in the highest precision over the testing dataset. We used Weka 3.7.4 to quickly build and test models [5]. We found that back-propagation neural networks did not result in good fits in general, with high bias on the training dataset. Based on the success with SVM modeling, the rest of the models we built and tested used these choices of parameters.

Table 1. Information Gain of Baseline Classifier Features

Feature	I. Gain
deviate-eavg5-amt*	0.0281
1d-derivative-mavg5**	0.0272
deviate-mavg10-amt***	0.0265
1d-derivative-mavg20	0.0255
1d-derivative-mavg26	0.0254
deviate-eavg10-amt	0.0247
1d-derivative-mavg10	0.0247
deviate-mavg5-amt	0.0243
1d-derivative-mavg3	0.0220

* normalized amount of deviation above the 5-day exponential moving average of opening price. ** normalized moving average of the one-day derivative of the opening price. *** normalized amount of deviation above the 10-day opening price moving average.

Table 2. Twitter Dataset Statistics (Total Tweets and Avg. Per Day)

Dataset	Total	Avg/Day
AAPL-TERMS	3994569	42,495.4
ATVI-TERMS	500169	5,320.95
ERTS-TERMS	473542	5,037.68
ADBE-TERMS	284719	3,028.93
DELL-TERMS	225345	2,397.29
DELL-TICKER	118603	1,261.73
NFLX-TERMS	63176	672.09
MAT-TERMS	42204	448.98
RIMM-TERMS	40960	435.74
MAT-TICKER	35038	372.74
GRMN-TERMS	23710	252.23
URBN-TERMS	9709	103.29
AAPL-TICKER	1884	20.04
RIMM-TICKER	564	6.00
NFLX-TICKER	189	2.01
URBN-TICKER	129	1.37
ATVI-TICKER	93	0.99
GRMN-TICKER	80	0.85
ADBE-TICKER	71	0.76
ERTS-TICKER	70	0.74

3. Baseline Classifier: Analysis of Best Features and Results

Table 1 shows the results of information-gain analysis on the top 10 constructed features. The results show that none of the features are particularly predictive of MPD events, and those that are primarily rely upon long-term trends (such as a stock's price opening above its 10-day average). We hypothesize that these results show that a baseline classifier is good at predicting the sorts of long-term gains that might not be reflected in Twitter sentiment, which could form the basis for an excellent supporting classifier.

Table 5 shows precision/recall for the best classifiers over our portfolio. The average precision in predicting MPD events is 0.297 (max 0.500), much too low to create an effective trading strategy alone.

4. Preparation of Twitter dataset

We wanted to consider only those tweets which were related to a stock in our development of NLP-based classifiers for the data. We started by only considering those tweets that contained the stock ticker symbol of the company (the "TICKER" datasets), but found that most of the stocks in our portfolio had < 100 tweets/day using that criterion (Table 2).

We therefore expanded our criteria to seven, in the "TERMS" datasets: stock ticker symbol, company name, names of major products, informal names of major products, CEO/ President names, informal company

names, and names/hashtags of major conferences that the company puts on. Our terms list only included terms known before June 2009 to avoid information leaks. Using these criteria, all stocks had over 100 tweets/day (Table 2)

We then tokenized each tweet using Potts' sentiment and twitter aware tokenizer since it is better suited than other tokenizers for the type of content that appears on Twitter [3].

5. Development of Sentiment Classifiers

Because Bollen et. al showed success with sentiment analysis, we attempted to replicate their method of classifying sentiment using GPOMS analysis of Twitter posts. Unfortunately, GPOMS is not publicly available, so we attempted to re-create it based on available literature.

6. Sentiment Classifiers: POMS-Expanded

As with GPOMS, we built upon the Profile of Mood States (POMS) psychological rating scale to measure 6 mood dimensions: Calm, Alert, Sure, Vital, Kind, & Happy. The original lexicon for POMS uses 65 adjectives to measure mood [6]. We expanded the lexicon to 626 terms using Wordnet by mapping each adjective to parent synsets that reflect the sentiment of the adjective [7]. All words in the synset families were stemmed and included in the POMS-Expanded set.

For each stock in our portfolio, we performed a daily measure of the 6 mood dimensions by tallying the occurrence of each of the 626 terms in the tweets related to the stock, grouping the tallies of each term into six counts corresponding to the POMS mood dimensions, and normalizing with respect to the volume of tweets for the day. We then used these normalized counts as features to predict whether the next day would have an MPD event.

7. Sentiment Classifiers: Results

Our classifier using these mood dimensions did not have good results: precision and recall were almost always zero (Table 3). While Bollen indicates that measures of calmness on Twitter are predictive of the closing direction of the DJIA, information-gain analysis and model validation showed that neither calmness nor the 5 other mood dimensions were predictive of MPD events of any stock in our portfolio.

Table 3. POMS-Expanded Classifier Results ("TERMS" datasets)

Dataset	Precision	Recall
ATVI	0	0
ERTS	0	0
AAPL	0	0
ADBE	0	0
DELL	0	0
NFLX	1	0.077
MAT	0	0
RIMM	0	0
GRMN	0	0
URBN	0	0

We hypothesized that the public mood was less indicative of individual stock performance than it is of the market as a whole; for that reason, we sought a more general method that could utilize a larger lexicon to predict MPD events.

8. Lexical Analysis: Mulitinomial Distribution over Words

As a first attempt, we simply counted the frequencies of all tokens in the previous day of tweets in the 'ticker' dataset. We theorized that directly related tweets would have the best predictive power. Our portfolio showed unfortunately low average results of 0.211 precision and 0.238 recall.

Data further indicated that the classifier had high variance, as testing on the training set resulted in an average of 0.9956 precision and 0.5492 recall. We therefore elected to add more training data, increasing the corpus to the "terms" datasets. This increase in data resulted in a gain of 8.9 percentage points in precision and 12 percentage points in recall (Figure 1). Even with this increase, however, classifiers showed high variance and we had no more data that we could easily add. Additionally noting that we had far more features (up to 290 thousand) than instances, we elected to determine if feature selection could help reduce variance.

9. Lexical Analysis: Feature Selection

We first used information gain feature selection available from Weka to select the top twenty features using Ranker search method. This resulted in a slight decline in both average precision and average recall, 0.2076 and 0.183 respectively. The precision when testing on the training set also decreased, to a average precision of 0.8468 and average recall of 0.6272. Thus, we had a smaller variance gap, but still experienced poor performance.

Our second attempt at performing feature selection was to use "token frequency–inverse document frequency" (TF.IDF)

instead of raw frequency in our normalized counts [4]. We pruned tokens that were in the bottom 10% and top 10% to eliminate both common words and rarely seen words that may result in over-fitting. This method resulted in higher average precision of 0.362 and lower average recall of 0.134. This precision, though greatly improved, was not satisfying. Thus, we turned our attention to applying domain knowledge of the stock market.

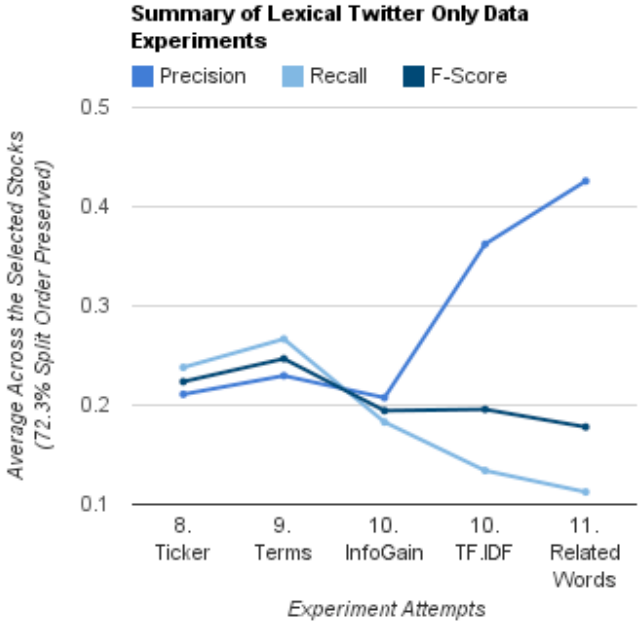
10. Lexical Analysis: Domain-specific words

We analyzed our Twitter corpus to determine words that may contain some domain knowledge of stock-market trading and could provide predictive knowledge on whether a stock would be having a major price movement. We determined seven categories of words and manually generated example words in each category (Table 4). We then counted the frequency of these words occurring in the previous day. This method performed poorly when classified by our tuned SVM. However, when we ran this experiment with a Multilayer Perceptron configured using the default parameters in Weka 3.7.4, data showed average precision of .426 and a recall of .113 when averaged across of all our stocks (Figure 1).

Table 4: Example Domain Words

Classification	Example Words
Buy Signal	buy, call, bought
Stock Event	analyst, earnings
Sell Signal	sell, sold, short
Trading Activity	trade
Investing Activity	invest
Short-Term Signal	day, tomorrow
Long-Term Signal	week, month

Figure 1



11. Lexical Analysis: Results

We were able to increase our precision up to .426 (Figure 1). Even though the f-score decreased to a little under .2, the key part of our trading strategy to ensure profitability is to maximize precision. We selected the final domain-specific words method to combine with our baseline classifier.

12. Combination Classifiers

After testing and validating both approaches, we combined the features generated in our baseline economic information classifier (step 3) and in our domain-specific words lexical analysis (step 10) and compared the results to identical SVM models built on each dataset individually. In this final step, we compared models at c=0.01 to reduce variance.

13. Combination Classifier Results

Table 5 shows results from the combination classifiers on each stock. Analysis of these results shows that combination classifiers improve the precision of MPD prediction by an average of 35.3 percentage points, and improve some stocks up to 78% precision. Average precision over our portfolio is 0.402, insufficient to make a profitable trading strategy, but two stocks — ATVI and ERTS — show sufficient precision to trade (0.603 and 0.778, respectively). ATVI and ERTS, additionally, had a recall of .567 and .438 which would trigger 11 trading events over the 41 trading-days in the test set.

Figure 2

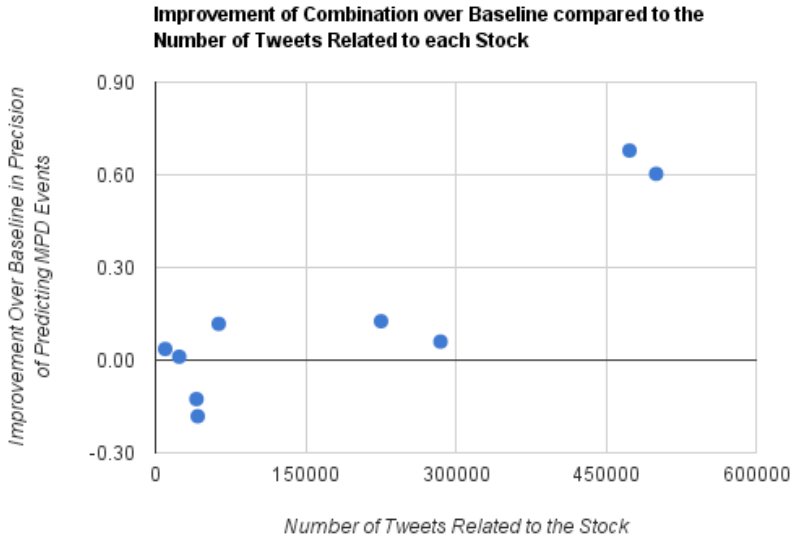


Table 5. Precision and Recall of Developed Classifiers

Test Set	Baseline Economic Classifier			Stock Words Classifier			Combination Classifier		
Train Set	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
ATVI-TERMS	**0.250**	**0.182**	**0.211**	**0.000**	**0.000**	**0.000**	**0.630**	**0.567**	**0.597**
	0.902	0.860		0.000	0.000		0.952	0.952	
ERTS-TERMS	**0.250**	**0.667**	**0.364**	**0.000**	**0.000**	**0.000**	**0.778**	**0.438**	**0.560**
	0.854	0.729		0.000	0.000		0.874	0.970	
ADBE-TERMS	**0.235**	**0.667**	**0.348**	**0.000**	**0.000**	**0.000**	**0.250**	**0.667**	**0.364**
	0.938	0.857		0.000	0.000		0.935	0.829	
DELL-TERMS	**0.333**	**0.583**	**0.424**	**0.000**	**0.000**	**0.000**	**0.381**	**0.667**	**0.485**
	0.870	0.784		0.000	0.000		0.860	0.725	
NFLX-TERMS	**0.353**	**0.500**	**0.414**	**0.000**	**0.000**	**0.000**	**0.400**	**0.500**	**0.444**
	0.837	0.706		0.000	0.000		0.860	0.725	
MAT-TERMS	**0.182**	**0.500**	**0.267**	**0.000**	**0.000**	**0.000**	**0.154**	**0.500**	**0.235**
	0.962	0.862		0.000	0.000		0.963	0.897	
RIMM-TERMS	**0.500**	**0.400**	**0.444**	**0.000**	**0.000**	**0.000**	**0.444**	**0.400**	**0.421**
	0.833	0.690		0.613	0.967		0.837	0.707	
GRMN-TERMS	**0.360**	**0.563**	**0.439**	**0.250**	**0.063**	**0.101**	**0.364**	**0.500**	**0.421**
	0.808	0.909		0.412	0.117		0.816	0.700	
URBN-TERMS	**0.214**	**0.250**	**0.231**	**0.333**	**0.231**	**0.273**	**0.222**	**0.154**	**0.182**
	0.813	0.867		0.560	0.240		0.769	0.690	
Average	**0.297**	**0.479**	**0.349**	**0.065**	**0.033**	**0.041**	**0.403**	**0.488**	**0.412**
	0.869	0.807		0.176	0.147		0.874	0.799	

This result may be due to the fact that ATVI and ERTS had the most related tweets in the dataset period. Figure 2 shows a weak correlation between the total number of tweets observed in the timeframe and the percent improvement in the combination classifier, with R^2=0.799

Conclusion

Our results suggest that Twitter data are predictive of major price difference events (MPD events): SVM classifiers using a combination of economic data and features gleaned from Twitter feeds predict MPD events with higher precision than classifiers based on economic data alone. Over a portfolio of 10 consumer-technology companies, combination classifiers improved the precision of MPD prediction by an average of 35.32 percentage points.

While the average precision is 0.402, too low to create a profitable options-straddle trading strategy over the entire portfolio, the two stocks with the highest Twitter comment volume show precisions of 0.630 and 0.778; this suggest that having more twitter data related to the stock will improve predictive performance of MPD events on the stock.

Acknowledgements

We would like to thank Mihai Surdeanu for his support and Andrew Ng for the opportunity to construct this project.

Works Cited

[1] Johan Bollen, Huina Maoa, & Xiaojun Zeng (2011); "Twitter mood predicts the stock market." Journal of Computational Science, Volume 2, Issue 1.

[2] Pham, Hung and Chien, Andrew and Lim, Youngwhan. "A Framework for Stock Prediction." 2009. http://cs229.stanford.edu/proj2009/PhamChienLim.pdf.s

[3] Potts, Christopher. "Sentiment Symposium Tutorial." 2011. Sentiment Analysis Symposium. http://sentiment.christopherpotts.net/

[4] Manning, Christopher D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass. [u.a.: MIT, 2005. Print.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Experimentplorations, Volume 11, Issue 1.

[6] McNair, Douglas and Maurice, Lorr and Droppelman, Leo. *Profile of mood states*. Educational and Industrial Testing Service, San Diego, CA. 1971. Original source unattainable, list retrieved from http://www2.ul.ie/pdf/526781607.pdf

[7] Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>

[8] "Technical Indicators and Overlays - ChartSchool - StockCharts.com."*StockCharts.com - Simply the Web's Best Financial Charts*. Web. 16 Dec. 2011. <http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators]>.

[9] "Yahoo Stock Search Data Set | Infochimps." *Infochimps | Smart Data for Apps and Analytics | Big Data Solutions | Infochimps*. Web. 16 Dec. 2011. <*http://www.infochimps.com/datasets/yahoo-stock-search*>.

[10] Yang J., Leskovec J. Temporal Variation in Online Media. ACM International Conference on Web Search and Data Mining (WSDM '11), 2011.

[11] Zhang, Wenbin, AND Skiena, Steven. "Trading Strategies to Exploit Blog and News Sentiment" International AAAI Conference on Weblogs and Social Media (2010): n. pag. Web. 16 Dec. 2011