Trading Strategies To Exploit Blog and News Sentiment

Wenbin Zhang and Steven Skiena
{wbzhang@cs.sunysb.edu} and {skiena@cs.sunysb.edu}
Department of Computer Science, Stony Brook University
Stony Brook, NY 11794-4400 USA

Abstract

We use quantitative media (blogs, and news as a comparison) data generated by a large-scale natural language processing (NLP) text analysis system to perform a comprehensive and comparative study on how a company's reported media frequency, sentiment polarity and subjectivity anticipates or reflects its stock trading volumes and financial returns. Our analysis provides concrete evidence that media data is highly informative, as previously suggested in the literature – but never studied on our scale of several large collections of blogs and news for over five years. Building on our findings, we give a sentiment-based market-neutral trading strategy which gives consistently favorable returns with low volatility over a five year period (2005-2009). Our results are significant in confirming the performance of general blog and news sentiment analysis methods over broad domains and sources. Moreover, several remarkable differences between news and blogs are also identified in this paper.

Introduction

The efficient market hypothesis asserts that financial markets are "informationally efficient", which means current stock prices already reflect all known information and all occurred facts. Moreover, prices in finance markets are unbiased and contain all the wisdom or future forecasts from investors. Therefore, investors cannot make excess profits from the market if their trading strategies are based on known information, because market prices are efficiently collecting and aggregating various information and keep changing without delay.

However, a large and growing literature documents that movements of financial indicators are not always consistent with the quantitative measures of firms' fundamentals (e.g. (Cutler, Poterba, and Summers 1989; Roll 1988; Tetlock, Saar-Tsechansky, and Macskassy 2007)). This mandates a rethinking of the fluctuation of stock prices to seek other evidence to explain it. Some encouraging results prove the conditional usage of the efficient market hypothesis. Particularly, (Chan 2003) shows that stock prices appear to drift after important corporate events for up to several months. This suggests that some of the drift is caused by the price's under-reaction to information. News data thus could provide a feasible and useful way to analyze financial markets.

Our primary goal is to study the relationship between stock market data and linguistic media data, both blogs and news, and to illustrate the extent to which they can contribute to the design of investment strategies. Our main contributions in this paper are:

- *Comparative Study of Blogs and News* – We conduct a thoughtful comparative study of four different linguistic sources, i.e., Twitter, Spinn3r RSS blogs, LiveJournal blogs, and Dailies news as a comparison. We compare their sentiments with corresponding stocks and evaluate the equity trading performance with using the four sources respectively. Our analysis also discovers many distinct properties between blogs and news. For example, news information could be incorporated into stock prices instantly (almost within 1 day) after release, while blog information like Twitter will be absorbed by stock market with a longer time period (around 2 to 3 days).
- *Large-Scale Analysis* – We give comprehensive results of analyzing stock market using roughly one terabyte of blog and news data and thousands of different companies. This scale of analysis has never been previously attempted in the literature, and enables us to identify short-term but statistically significant correlations between media volume / sentiment and financial returns / trading volumes.
- *Corpus Size Matters* – Previous work on sentiment-based financial analysis (e.g. (Tetlock, Saar-Tsechansky, and Macskassy 2007)) focus explicitly on national financial newspapers, namely the *Dow Jones News Service* and the *Wall Street Journal*. However, we demonstrate that a more significant, reliable sentiment signal comes from analyzing a full corpus of blogs and news than just reading the DJNS or WSJ.
- *Sentiment-oriented Equity Trading* – We propose a market-neutral stock trading strategy, based completely on sentiment data drawn from published blog or news sources. Through careful experimentation over five full years of news/price (2005-2009) data, we demonstrate that our strategy provides intriguing returns with low variance (ignoring both transaction costs and the timing resolution discussed in Section *Media Timing Issues*).

- *Validation of Sentiment Analysis Methods* – Perhaps another important contribution of our paper is the strongest validation to date of the accuracy of our media sentiment analysis methodology of *Lydia*. Proper validation is impossible in the absence of any agreed upon gold standard for entity-level sentiment analysis (Pang and Lee 2008). But our ability to extract a sufficiently reliable sentiment signal for successfully trading upon (regardless of timing resolution) provides rigorous evidence that our sentiment methods accurately reflect real changes in response to linguistic information.

This paper is organized as follows. First we review related work. We then describe the origin and characteristics of the media and financial data we work with. After that, we give a complete analysis of the correlation between major stock market variables and major media variables, which is the most important part of this paper. Finally, we propose and evaluate a market-neutral trading strategy based on media data. We conclude that financial prices are significantly correlated with quantitative media data and can be used to formulate interesting trading strategies.

Related Work

Previous work is divided between the finance and computer science academic communities. We first survey research from the financial realm.

Tetlock (Tetlock, Saar-Tsechansky, and Macskassy 2007) investigates whether the occurrence of negative words in firm-specific news articles can help us predict firms' cash flows and whether firms' stock market prices incorporate linguistic information efficiently. They claim that firms' stock prices under-react to the underlying negative information of news articles. More specifically, negative information in news articles are reflected in stock market prices with roughly one-day delay.

Chan (Chan 2003) examines monthly returns to a subset of stocks after public news about them is released and finds that investors react slowly to information, especially after bad news. Another important finding is that stocks tend to reverse in the subsequent month after extreme price movements unaccompanied by public news. In addition, these patterns are statistically significant. One limitation of this study uses coarse, monthly granularity. In our paper, we provide analysis of daily news and price movements.

Antweiler and Frank (Antweiler and Frank 2004) study more than 1.5 million messages from Yahoo! Finance and Raging Bull, which are the two most popular Internet Stock Message Boards. They employed Naive Bayes and Support Vector Machine classifiers to assess "bullishness" content of these stock messages. They show these message boards are quite informative, and further that bullishness is positively and significantly associated with returns. In terms of trading volume, the paper shows controversial opinions are associated with more trades.

From the Computer Science side, intense researches are delivered by text mining or machine learning communities. Their basic idea is to quantify linguistic information with text mining techniques, get the predefined set of features of the training data, and then build various models with classical statistical approaches or statistical learning algorithms.

A detailed survey of the text mining for market response to news can be found in (Mittermayer and Knolmayer 2006a). In particular, the 3-category model is widely used to label documents or words. The first category (positive sentiment) consists of news articles or words that make the associated financial variables increase to a certain degree in a certain time period, for example, a news event makes the price of the single stock "IBM" increase 0.5% in the following day. Similarly, the second category (negative sentiment) is defined accordingly. The third category consists of neutral news articles or words. Research which can characterized under this model includes (Fung, Yu, and Lam 2002; Mittermayer and Knolmayer 2006b; Thomas 2003; Wuthrich, Cho, and etc. 1998).

There has also been substantial interest in the opinion mining and NLP community on using financial text streams as a domain to test sentiment analysis methods, including (Chaovalit and Zhou 2005; Pang and Lee 2002). Broadly speaking, they apply information retrieval or machine learning techniques to classify text streams into some categories and hope to produce better classification accuracy than human being, and thus the underlying opinion could be discovered. Pang and Lee (Pang and Lee 2008) gave a detail review in this domain.

Stock and Media Data

Here we describe the stock and media data sources which is the basis for our analysis in this paper.

Stock Data

Our stock price and volume data is obtained from Thomson Datastream Services (Datastream), a comprehensive database with time series on more than two million instruments. Here we only consider the stocks listed in the New York Stock Exchange because those stocks have more intensive blogging coverage than stocks in other markets. We downloaded the data of all 3238 stocks within the period from 2005 to 2009, for their daily open, close, high, low prices, turnover volumes, and monthly market capitalizations.

Media Data

Company-related blog and news data was generated using the *Lydia* ((Lloyd, Kechagias, and Skiena 2005), *http://www.textmap.com*), a high-speed text processing system, which reduces large text streams to time series data on the frequency of sentiment of underlying media entities. In this paper, we compare four different collections of blog/news sources as follows.

1. *Dailies*, which includes the coverage of over 500 nationwide and local newspapers.
2. *Twitter*, which is a free social networking and microblogging service that enables its users to send and read messages known as tweets.

Depositories	2005	2006	2007	2008	2009
Dailies	√	√	√	√	√
Twitter					√
Spinn3r				√	√
LiveJournal		√	√	√	

Table 1: Time-range coverage of different depositories (sources) in our media database.

3. *Spinn3r RSS Feeds*, which is a collection of blogs worldwide.
4. *LiveJournal*, which includes all the blogs provided by LiveJournal.

Table 1 shows the corresponding time range for all the four media sources. Because Dailies depository has the biggest data volume among all above four sources, analysis based on Dailies should be more credible and we will regard it as the benchmark for blog analysis.

Media Timing Issues

Proper interpretation of our results requires careful attention to the timing of our news spidering (text retrieval) agents. For the Dailies news corpus we employ, the spidering program begins to download news at 11pm EST every day, a process which can take up to 12 hours. All of these articles are credited to the day when the spider program began running. Thus while the bulk of our news was certainly retrieved before the 9:30AM opening of the NYSE each day, we cannot guarantee that it is unpolluted by news reporting events after the market opening.

The three blog-type medias (Twitter, Spinn3r, and LiveJournal) are different, because accurate time stamp will be provided while blogs are published. The general consistency of our results across all four corpores (under different timing models) lends support to our conclusions, but the degree to which we follow rather than anticipate price movements cannot be conclusively answered by this study.

Lydia Sentiment Analysis

The *Lydia* sentiment data consists of time series of favorable (positive) and unfavorable (negative) words co-referenced with occurrences of each named entity (here denoting companies). With using a graphtheoretic approach described in (Godbole, Srinivasaiah, and Skiena 2007), the lexicon of almost 5,000 sentiment-laden words was constructed by expanding synonyms/antonyms from small sets of seed words associated with *Business*, *Crime*, *Health*, *Politics*, *Sports*, and *Media* domains. A *General* sentiment index aggregates the lexicons from all of these domains. Further details of the *Lydia* sentiment analysis methods and their validation are reported in (Bautin, Vijayarenu, and Skiena 2008; Godbole, Srinivasaiah, and Skiena 2007).

For financial market analysis, we were most interested in *General*, *Business*, and *Media* categories. After initial correlation analysis, we identified the *General* category is the most relevant one, and thus we only use the *General* sentiment in the subsequent analysis.

Let p and n denote the number of raw positive and negative references to a given entity, which occurs a total of N times in the corpus (including neutral references). Then we derive the following natural sentiment/subjectivity measures from these raw counts:

- polarity $= (p - n)/(p + n)$
- subjectivity $= (n + p)/N$
- pos_refs_per_ref $= p/N$
- neg_refs_per_ref $= n/N$
- senti_diffs_per_ref $= (p - n)/N$

These derived measures are not highly correlated with raw sentiment counts and they can provide additional information that raw data cannot. Therefore, with them we will be able to avoid multicollinearity during linear analysis.

Matching Stock / Media Entities

An important technical problem concerns matching the stock and news entity names. For example, the NYSE-listed “First Commonwealth Financial” is associated with three entities in our database: “First Commonwealth Financial”, “First Commonwealth Financial Corporation”, and “First Commonwealth Financial Corp.” We aggregate the time series of all the three media entities to define the media time series for this specific company. Our matching algorithm yields around 700 to 1300 matched stock/entity pairs for the four media sources we examined.

Intercorrelationship Among Media Sources

Here we investigate the relationship between any two of the four media corpuses we study. Our analysis shows data of one source is significantly correlated with data of any other sources. For example, some major observations between Spinn3r and Dailies are:

- *Reference Frequency* – The correlation between the monthly (yearly) normalized article counts of Spinn3r vs. Dailies is 0.4402 (0.6257). The corpuses share a higher correlation with frequency than with sentiment variables, reflecting both the greater variance in editorial outlook and the difficulty of detecting sentiment precisely using algorithmic methods.
- *Sentiment Polarity* – The monthly sentiment polarity correlation across the two corpuses is 0.3765, and is statistically significant. Our experimental results show sentiment differences per reference is a better measure than raw polarity because it eliminates some extreme values (+1, 0, or -1) of raw polarity, and thus we will use it to measure general polarity in the following sections.
- *Subjectivity* – The correlation coefficient of subjectivity between the two corpuses is 0.3127.

We can get consistent intercorrelation result for other source pairs. These substantial correlations indicates blogs and news share similar opinions to some extent over the same entities, and explains why we obtain qualitatively similar results for all corpuses.

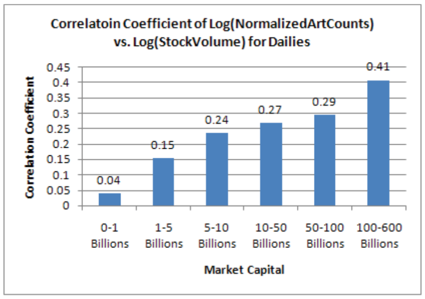

Figure 1: Logged Monthly Normalized Article Counts vs. Logged Stock Trading Volume for Dailies news, broken down by market capitalization.

Correlation of Media / Stock Data

In this section, we analyze the correlations between media and stock variables on a large scale. Here large scale means analysis for all "media vs. NYSE" matched name pairs.

Media Frequency vs. Trading Volume

The first problem we study is the relationship between media references and stock trading volume. Intuitively, more media coverage should lead to more trades.

To compensate for technical variations in spidering efficacy, we use normalized article counts instead of raw article counts to correct for fluctuations in the total volume of news spidered each day. In particular, we will always use logged normalized article counts as our standard measure of entity news frequency, which follows a Gaussian distribution.

Some significant observations are:

- *Strength of correlation* – For all four depositories, the correlation coefficient between logged normalized article counts and logged stock trading volume are more than 0.4.
- *Daily, monthly, and yearly analysis* – For Dailies, the correlation coefficient of logged normalized article counts versus logged normalized stock trading volume for daily, monthly, and yearly analysis are 0.2715, 0.4204, and 0.4747 respectively. Therefore, monthly analysis is a proper time scale for analysis.
- *Persistence over time* – How well does today's article counts correlate with yesterday's or tomorrow's trading volumes? For Dailies, a one day lag generates the highest correlation (0.74), but the correlation coefficients persist (between 0.64 and 0.68) for periods up to ten days in the future. This is due to the high auto-correlation for both article counts and stock trading volumes.
- *Influence of Market Sectors* – The correlation analysis could be further broken down by market sectors. We find that for sectors "Pharmaceuticals & Biotechnolog" and "Aerospace & Defense", intensive news references are more likely to cause more trades (corr > 0.7). By contrast, the trading volumes for sectors "Electronic & Electrical Equipments" and "Software & Computer Services" are less sensitive to media exposure (corr < 0.2).
- *Breakdown by Market Capitalization* – Figure 1 shows the breaking down analysis for different market capitalizations. This figure indicates that the correlation coefficients between article counts and stock trading volume become stronger and stronger with the increasing of market capitalization. For large enough companies, their news coverage reflects relative importance more than distinctive newsworthiness.

Although most of above provided data focus on Dailies news, the corresponding results are suitable for other three blog sources as well because of the significant intercorrelationship among difference sources.

Frequency vs. Capitalization

The second problem we studied is the relationship between firms' media references and their corresponding market capitalizations. Usually bigger firms receive more media coverage. Indeed, in our media database, the logged monthly normalized article count is also positively correlated with the logged market capitalizations with a correlation coefficient of 0.42, and it is statistically significant.

Media Polarity vs. Stock Returns

A more interesting question is the return of stocks. We believe the return of stocks are relevant to the public opinion of corresponding firms, say, how good or how bad people think about these firms. If people think a firm is good, more likely its stock price will raise, and vice versa. From the previous section, we know that "polarity" is a quantitative term to describe how good a firm is.

Variable Selections First we will figure out the right variables to study this problem. We consider three different performance measures for a given stock s: change of stock prices, stock returns ($R(s)$), and abnormal return. The abnormal return $R'(s)$ is calculated by

$$R'(s) = R(s) - R(NYSE)$$

In our correlation analysis we correlate each news variable from [polarity, change of polarity, percentage change of polarity], to each stock variable from [change of stock prices, stock return, stock abnormal return]. This gives six combination pairs for testing. Our experiments show (polarity, stock return) pair has the most significant correlations among all the combinations, so we only give the analysis results for polarity versus stock returns in the following parts. Actually, the (polarity, stock abnormal return) pair achieves very similar performance with (polarity, stock return).

Correlation Analysis with Shifting of Time Figure 2 examines how much today's polarity is correlated with stock returns on proximate days. For Dailies, we see that (1) the correlation coefficient of today's polarity versus previous return decrease gradually, and (2) for days 1 and later, all the correlation coefficients are almost zero, and all those correlations are not statistically significant.

This proves that today's news almost have no predictive power for the return of tomorrow or later days. We also notice that the return of day 0 has the best correlation with

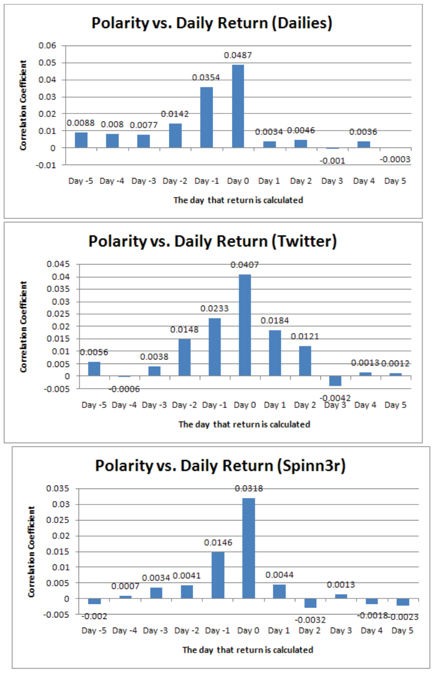

Figure 2: Correlation analysis of Polarity vs. Daily Return for Dailies, Twitter, and Spinn3r respectively. The correlation coefficients are calculated with time lags from -5 to 5 days. Please note: we do not show LiveJournal result here because its data volume is relative small and thus the correlation between polarity and stock return is not significant.

polarity. In fact, most of our daily news are published in the early morning each day, and thus it is reasonable to infer they have some predictive power for current day's return. In the other word, today's news has significant relationship with the current day's return, has some relationship with yesterday's return, but almost has no relationship with tomorrow's return.

The efficient market hypothesis states that the market reflects public information in the stock price within a very short time. Therefore, Dailies' polarity shown in Figure 2 illustrated this theory perfectly, i.e., the correlation between news polarity and stock returns disappear after 1 day.

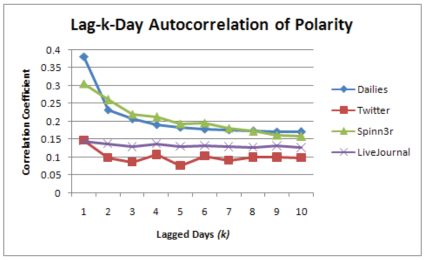

Figure 3: Polarity's Lag-*k*-Day autocorrelation for Dailies, Twitter, Spinn3r, and LiveJournal respectively.

Blogs vs. News Figure 2 also shows Spinn3r data are very similar to Dailies data. However, Twitter is somewhat different in that its polarity also has some corrrelaionship with tomorrow's or the day after tomorrow's return. That is, stock market incorporates Twitter sentiment slower than news, may need 2 or 3 days. In our opinion, there are two possible reasons (1) Our Twitter database only contains data for 0.5 year, and datasize is small and thus the result is less accurate, and (2) Twitter data has autocorrelationship between neighbor days and the sentiment of Twitter is more persistent.

To dig this problem further, we show polarity's Lag-*k*-Day autocorrelation for the four media sources in Figure 3. This figure indicates Dailies and Spinn3r have the similar autocorrelation levels of polarity, while Twitter and LiveJournal have the lower levels in terms of the strength of autocorrelations. One surprising fact is that all those three blog-type sources have more moderate slopes than Dailies, especially the curves of Twitter and LiveJournal are near flat. This is a very important difference between news and blogs, i.e., the sentiment conveyed by blog can last longer than news. This phenomenon is understandable because news has more significant recency effect.

Strengthening the Correlation The sentiment-return correlation can be improved by removing companies with the weakest detected polarity to focus on those showing significant sentiment signal. For example, we can try to remove $\alpha\%$ neutral data of polarity. Actualy, once 80% of the neutral sentiment are removed, the correlation coefficients will become as strong as 0.3∼0.6.

We can also break down the correlation analysis by market sector. Particularly in our Dailies source, the "Household Goods & Home Constructions","Life Insurance", and "Financial Services" sectors are most strongly affected by news sentiment (corr > 0.17); all of which are strongly associated with the subprime mortgage crisis. By contrast, returns from the staid "Fixed Line Telecommunications", "Industrial Transportation", and "Beverages" sectors have near zero correlation with news sentiment.

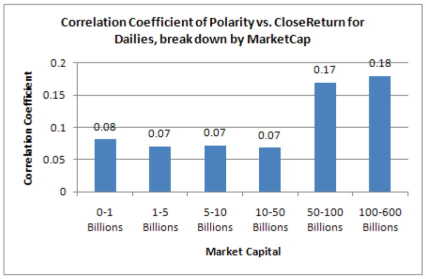

Figure 4: Polarity vs. Monthly Stock Close Return for Dailies. The analysis are broken down by the scale of market capitalizations.

From Figure 4, we can see that correlation coefficient for bigger firms is much stronger than those for smaller firms, especially, for firms who have more than 50-billion market capitalization. The result makes sense, because large firms generate more intensive media coverage and thus the collective information can better indicate the firms' situation.

Subjectivity vs. Trading Volume

Now we consider the relationship between blog/news subjectivity and stock trading volume. Subjectivity means the amount of sentiment references among total references. Within the sentiment analysis community, subjectivity is considered a more robust measure than polarity (Pang and Lee 2008).

In all blog and news sources, subjectivity is positively and significantly correlated with stock trading volume. This conclusion coincides with the result from Antweiler and Frank (Antweiler and Frank 2004) that controversial opinions are associated with more trades.

A Sentiment-Based Trading Agent

We have demonstrated significant correlations between media data and financial market indicators. In this section, we design a market-neutral trading agent to demonstrate the predictive power of news data. A *market-neutral* strategy seeks to profit from both increasing and decreasing prices in a single or numerous markets by taking matching long and short positions in different stocks.

We propose our market-neutral algorithm and backtest it using real market data from 2004 to 2009. The results suggest that blog/news sentiment analysis should be employed as an informative component of trading agents.

The Market-neutral Strategy

Our market-neutral strategy first ranks companies by their reported sentiment each day, then goes long (short) on equal amounts of positive (negative) sentiment stocks. The yearly or monthly returns generated by such a trading agent will be used for performance evaluation.

Our initial investment is M and backtesting period is from start date D_s to end date D_e. We identify four key tunable parameters in this strategy:

- n: The number of stocks selected from the top and bottom of the firm list (sorted by sentiment).
- s: The number of historical days used for sentiment calculation. If $s = 1$, only the current day's sentiment is considered.
- h: Holding days, which means how many days we will hold for the current portfolio.
- C_l and C_u: The lower bound and upper bound of firms' market capitalization. We only consider the stocks whose market capitalizations are in range $[C_l, C_u]$.

The four parameters impact our stock returns substantially, and the details will be given in the following sections. The specific algorithm is described in Algorithm 1. Because we always long stocks with the best sentiment, and short stocks with the worst sentiment, this algorithm is also called the *best-sentiment strategy*.

Algorithm 1 A sentiment-based market-neutral strategy

```
Require: $n$, $s$, $h$, $[C_l, C_u]$, $M$, $[D_s, D_e]$.
Get a list of matched pairs of NYSE stocks and firm entities in blogs/news.
For each matched pairs, get the stock price and firm polarity time series.
for each day $D_i$ from $D_s$ to $D_e$ do
    if $D_i$ is one of the first $h$ trading days in $[D_s, D_e]$ then
        Sort all stocks based on their polarity of day $D_i$, with filtering out stocks if their corresponding market capitalization is not in $[C_l, C_u]$.
        For each stock in top/bottom list, invest $M/(2nh)$ with open prices.
    else if $D_i$ is one of the final $h$ trading days in $[D_s, D_e]$ then
        Sell stocks bought at trading day $(D_i - h)$ with open prices.
    else
        Sort all stocks based on their polarity of day $D_i$, with filtering out stocks if their corresponding market capitalization is not in $[C_l, C_u]$.
        Sell stocks bought at trading day $(D_i - h)$ with open prices, get bank roll $M_{D_i}$.
        For each stock in top/bottom list, invest $M_{D_i}/(2n)$ with open prices.
    end if
end for
return Final bank roll $M'$, and yearly/monthly return $R$.
```

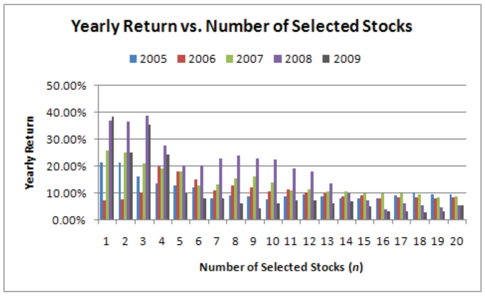

Figure 5: Yearly return vs. number of selected top and bottom stocks for Dailies depository. We tune n from 1 to 20, and fix parameters: $s = 1$, $h = 2$, $C_l = 10\ billions$, $C_u = 600\ billions$.

Performance Evaluation

In this section, we backtest our market-neutral strategy with real blog/news and stock data over a period from 2005 to 2009. There are four key parameters (n, s, h, C_l and C_u) that contribute to the final returns. We performed experiments to isolate one parameter while fixing the other three:

- *Diversification* – Figure 5 shows the impact of the number of selected stocks for Dailies. As we can see, with the increasing of the number of selected stocks, the yearly return decreases. The reason is straightforward – the stocks with the highest (lowest) sentiment should have the biggest price movement. If we select fewer stocks, the average sentiment of selected stocks will be higher, and the expected return will be higher as well.

  Blogs vs. News – Comparing with news result in Figure 5, Figure 6 shows the performance of blog-type sources, running the same experiments. A big different between news and blogs is that simulation with news uses current day's sentiment, while simulation with blogs uses yesterday's sentiment because current day's blogs are probably published after the open of stock market. Spinn3r has similar performance with Dailies, while Twitter and Live-

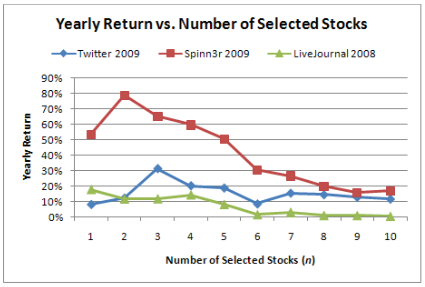

Figure 6: Yearly return vs. number of selected top and bottom stocks for blog-type depositories.

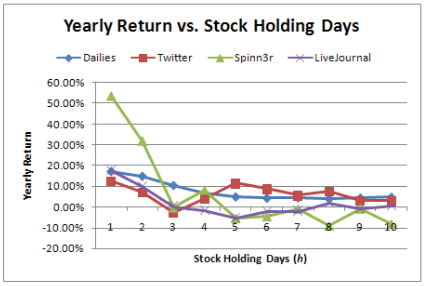

Figure 7: Yearly return vs. stock holding days. In this Figure, Dailies uses data of 2006.

Journal have much lower performance, basically because Twitter and LiveJournal do not have plenty of data volume. However, all the blog-type sources also show the same performance trend with that of news.

- *Sentiment Analysis Period* – For four of the five years studied (except 2008), yearly returns decrease with the increasing of sentiment analysis period s. This is consistent with the efficient market hypothesis, since longer periods dilute the freshness of the news.

- *Holding Period* – Another tunable parameter is length of time we hold the stock. For all the five years, longer holding time leads to lower returns. Again, the market will quickly reflect all the news information, and thus we will not benefit from extra holding days. Moreover, quickly redeeming the investment frees up capital to invest in more recently reported-on stocks.

  Blogs vs. News – Figure 7 give the comparison of Dailies news and blogs in terms of the influence of stock holding days. We notice that with news data, the performance monotonously and gradually decrease with the increase of holding days, but with blogs data, there are much more fluctuations in the performance curves.

- *Market Capitalization* – Our experiments showed an interesting influence of market capitalizations. Both large and small firms showed greater returns than medium-size firms. The return for small firms is enhanced because their price is more affected by news events/sentiment. For large firms, we more accurately measure sentiment due to the greater volume of media coverage.

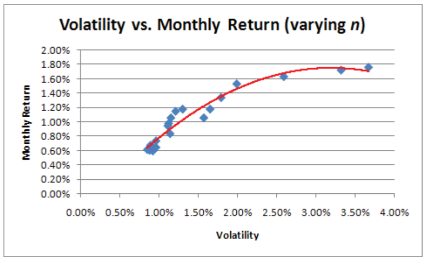

Figure 8: Portfolio Landscape of Monthly Return vs. Volatility analysis for experiments shown in Figure 5.

From these experiments, we conclude that our agent should hold small numbers of selected stocks, use short sentiment-calculation and stock holding periods, and avoid holding medium-sized firms.

To further validate the correctness of our blog/news sentiment analysis, we design other two strategies - *Worst-sentiment Strategy* and *Random-selection Strategy*, the former does the opposite of *Best-sentiment Strategy*, i.e., long bottom stocks and short top stocks, while the latter just randomly picks stocks to be long and short. Our simulation results show that *best-sentimen strategy* always returns positive gains, *worst-sentiment strategy* always produces negative returns, while *random-selection strategy* oscillates about zero return.

Returns and Volatility

Returns only capture part of investment performance. The degree of risk (volatility) taken on to achieve these returns determines to amount of leverage which can safely be employed to exploit the agent, and the overall desirability of a given portfolio in the risk-return horizon.

Here we analyze the monthly returns and volatility. The monthly return is taken to be the mean value of returns for all 60 months; the volatility is the standard deviation of these monthly returns.

Figure 8 demonstrates the tradeoff between risk and return, with a scatter plot of performance vs. volatility for strategies differing only in the number of stocks held in each period. Increased diversification reduces risk. The result is consistent with modern portfolio theory regarding risk and return. If we assume monthly return follows standard Gaussian distribution, the 95% confidence interval of monthly return could be constructed by two standard deviations from the mean, and according to this, investors could choose a diversification level to balance return and the risk he can afford. An interesting observation is that selecting only one stock does not always yield the highest returns, just as the result shown in Figures 5 and 6. Actually $n = 1$ is very risky, and usually $n = 2$ or 3 are much better. It n becomes larger and larger, it will make the investment less risky, but simultaneously it will reduce return.

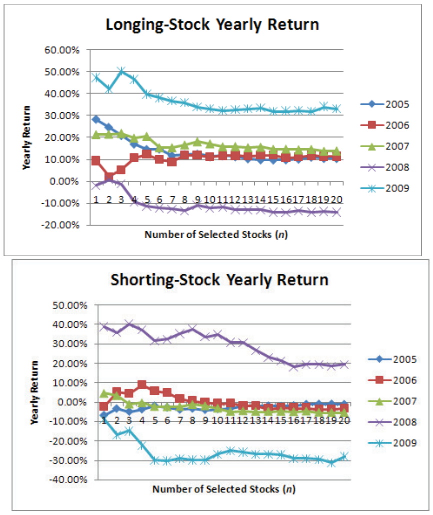

Figure 9: Returns for long vs. short in the experiments of Figure 5.

Long vs. Short

An interesting question in any market-neutral strategy is the relative contribution of returns going long vs. going short. To answer this, we partition the experiments in Figure 5 into long and short components in Figure 9.

Figure 9(a) shows long returns positive in 2005 to 2007, turning negative in 2008, but turning positive again in 2009. By contrast, Figure 9(b) shows near zero short returns in 2005 to 2007, but very high positive short returns in 2008, and very high negative short return in 2009. That is, most profits come from shorting stocks in 2008 and from longing stocks in 2009, which due to the collapse of the broad market in 2008 and recovery in 2009, and thus this result perfectly validates the market-neutrality of our strategy.

Conclusions

We have shown that raw or derived blog/news variables are significantly correlated with some indicators in stock markets, e.g., media references versus stock trading volume, media references versus market capitalization, media polarity versus stock returns, media subjectivity versus stock trading volume, and the opinions from one media source can reflect those from another media source. Based on blog/news sentiment data, we design a market-neural strategy, which is able to generate consistent returns for investors. There are several tunable parameters in this strategy, and thus investors need to carefully tune them to balance risk and return.

Our analysis also reveals many similar and distinct properties between news and blog sources. For example, both blog and news show similar sentiment vs. stock return correlationship, and we can get similar performance trends if we design trading strategies with these sources. However, opinions in blogs are more persistent and decay more gradually over time than news. In addition, comparing with blogs, news information takes a much shorter time period (within 1 day) to be incorporated into stock market completely.

References

Antweiler, W., and Frank, M. Z. 2004. Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance* 3:1259–1294.

Bautin, M.; Vijayarenu, L.; and Skiena, S. 2008. International sentiment analysis for news and blogs. In *Second Int. Conf. on Weblogs and Social Media (ICWSM 2008)*.

Chan, W. S. 2003. Stock price reaction to news and nonews: Drift and reversal after headlines. *Journal of Financial Economics* 70:223–260.

Chaovalit, P., and Zhou, L. 2005. Movie review mining: a comparison between supervised and unsupervised classification approaches. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*.

Cutler, D. M.; Poterba, J. M.; and Summers, L. H. 1989. What moves stock prices? *Journal of Portfolio Management* 15:4–12.

Datastream, T. http://www.datastream.com/.

Fung, G.; Yu, J.; and Lam, W. 2002. News sensitive stock trend prediction. *Proceedings 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining* 481–493.

Godbole, N.; Srinivasaiah, M.; and Skiena, S. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the First International Conference on Weblogs and Social Media*, 219–222.

Lloyd, L.; Kechagias, D.; and Skiena, S. 2005. Lydia: A system for large-scale news analysis. In *Proc. 12th String Processing and Information Retrieval (SPIRE 2005)*, volume LNCS 3772, 161–166.

Mittermayer, M., and Knolmayer, G. F. 2006a. Text mining system for market response to news: A survey. *Working Paper No 184*.

Mittermayer, M.-A., and Knolmayer, G. 2006b. Newscats: A news categorization and trading system. In *Proceedings of the International Conference in Data Mining (ICDM06)*.

Pang, B., and Lee, L. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 79–86.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* Vol. 2, No 1-2:1–135.

Roll, R. W. 1988. *R*-squared. *Journal of Finance* 541–566.

Tetlock, P. C.; Saar-Tsechansky, M.; and Macskassy, S. 2007. More than words: Quantifying language to measure firms' fundamentals. In *Proceedings of 9th Annual Texas Finance Festival*.

Thomas, J. 2003. News and trading rules. *Dissertation of Carnegie Mellon University, Pittsburgh*.

Wuthrich, B.; Cho, V.; and etc. 1998. Daily prediction of major stock indices from textual www data. In *Proceedings of 4th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*, 364–368.