# Evaluating Sentiment Analysis in the Context of Securities Trading

**Siavash Kazemian** and **Shunan Zhao** and **Gerald Penn**
Department of Computer Science
University of Toronto
{kazemian,szhao,gpenn}@cs.toronto.edu

## Abstract

There are numerous studies suggesting that published news stories have an important effect on the direction of the stock market, its volatility, the volume of trades, and the value of individual stocks mentioned in the news. There is even some published research suggesting that automated sentiment analysis of news documents, quarterly reports, blogs and/or twitter data can be productively used as part of a trading strategy. This paper presents just such a family of trading strategies, and then uses this application to re-examine some of the tacit assumptions behind how sentiment analyzers are generally evaluated, in spite of the contexts of their application. This discrepancy comes at a cost.

## 1 Introduction

The proliferation of opinion-rich text on the World Wide Web, which includes anything from product reviews to political blog posts, led to the growth of sentiment analysis as a research field more than a decade ago. The market need to quantify opinions expressed in social media and the blogosphere has provided a great opportunity for sentiment analysis technology to make an impact in many sectors, including the financial industry, in which interest in automatically detecting news sentiment in order to inform trading strategies extends back at least 10 years. In this case, sentiment takes on a slightly different meaning; positive sentiment is not the emotional and subjective use of laudatory language. Rather, a news article that contains positive sentiment is optimistic about the future financial prospects of a company.

Zhang and Skiena (2010) experimented with news sentiment to inform simple market neutral trading algorithms, and produced an impressive maximum yearly return of around 30% — even more when using sentiment from blogs and twitter data. They did so, however, without an appropriate baseline, making it very difficult to appreciate the significance of this number. Using a very standard, and in fact somewhat dated sentiment analyzer, we are regularly able to garner annualized returns over twice that percentage, and in a manner that highlights two of the better design decisions that Zhang and Skiena (2010) made, viz., (1) their decision to trade based upon numerical SVM scores rather than upon discrete positive or negative sentiment classes, and (2) their decision to go long (resp., short) in the $n$ best- (worst-) ranking securities rather than to treat all positive (negative) securities equally.

On the other hand, we trade based upon the raw SVM score itself, rather than its relative rank within a basket of other securities as Zhang and Skiena (2010) did, and we experimentally tune a threshold for that score that determines whether to go long, neutral or short. We sampled our stocks for both training and evaluation in two runs, one without *survivor bias*, the tendency for long positions in stocks that are publicly traded as of the date of the experiment to pay better using historical trading data than long positions in random stocks sampled on the trading days themselves. Most of the evaluations of sentiment-based trading either unwittingly adopt this bias, or do not need to address it because their returns are computed over very brief historical periods. We also provide appropriate trading baselines as well as Sharpe ratios (Sharpe, 1966) to attempt to quantify the relative risk inherent to our experimental strategies. As tacitly assumed by most of the work on this subject, our trading strategy is not portfolio-limited, and our returns are calculated on a percentage basis with theoretical, commission-free trades.



*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2094–2103,
Berlin, Germany, August 7-12, 2016. 

It is important to understand at the outset, however, that the purpose of this research was not to beat Zhang and Skiena's (2010) returns (although we have), nor merely to conduct the first properly controlled, sufficiently explicit, scientific test of the descriptive hypothesis that sentiment analysis is of benefit to securities trading (although, to our knowledge, we did). The main purpose of this study was in fact to reappraise the evaluation standards used by the sentiment analysis community. It is not at all uncommon within this community to evaluate a sentiment analyzer with a variety of classification accuracy or hypothesis testing scores such as F-measures, SARs, kappas or Krippendorf alphas derived from human-subject annotations — even when more extensional measures are available, such as actual market returns from historical data in the case of securities trading. With Hollywood films, another popular domain for automatic sentiment analysis, one might refer to box-office returns or the number of award nominations that a film receives rather than to its star-rankings on review websites where pile-on and confirmation biases are widely known to be rampant. Are the opinions of human judges, paid or unpaid, a sufficient proxy for the business cases that actually drive the demand for sentiment analyzers?

We regret to report that they do not seem to be. As a case study to demonstrate this point (Section 4.3), we exhibit one particular modification to our experimental financial sentiment analyzer that, when evaluated against an evaluation test set sampled from the same pool of human-subject annotations as the analyzer's training data, returns poorer performance, but when evaluated against actual market returns, yields better performance. This should worry any researcher who relies on classification accuracies, because the improvements that they report, whether due to better feature selection or different pattern recognition algorithms, may in fact not be improvements at all. Differences in the amount or degree of improvement might arguably be rescalable, but Section 4.3 shows that such intrinsic measures are not even accurate up to a determination of the delta's *sign*.

On the other hand, the results reported here should not be construed as an indictment of sentiment analysis as a technology or its potential application. In fact, one of our baselines alternatively attempts to train the same classifier directly on market returns, and the experimental approach handily beats that, too. It is important to train on human-annotated sentiments, but then it is equally important to tune, and eventually evaluate, on an empirically grounded task-specific measure, such as market returns. This paper thus presents, to our knowledge, the first real proof that sentiment is worth analyzing in this or any other domain.

A likely machine-learning explanation for this experimental result is that whenever two unbiased estimators are pitted against each other, they often result in an improved combined performance because each acts as a regularizer against the other. If true, this merely attests to the relative independence of task-based and human-annotated knowledge sources. A more HCI-oriented view, however, would argue that direct human-subject annotations are highly problematic unless the annotations have been elicited in manner that is *ecologically valid*. When human subjects are paid to annotate quarterly reports or business news, they are paid regardless of the quality of their annotations, the quality of their training, or even their degree of comprehension of what they are supposed to be doing. When human subjects post film reviews on web-sites, they are participating in a cultural activity in which the quality of the film under consideration is only one factor. These sources of annotation have not been properly controlled in previous experiments on sentiment analysis.

Regardless of the explanation, this is a lesson that applies to many more areas of NLP than just sentiment analysis, and to far more recent instances of sentiment analysis than the one that we based our experiments on here. Indeed, we chose sentiment analysis because this is an area that *can* set a higher standard; it has the right size for an NLP component to be embedded in real applications and to be evaluated properly. This is noteworthy because it is challenging to explain why recent publications in sentiment analysis research would so dramatically increase the value that they assign to sentence-level sentiment scoring algorithms based on syntactically compositional derivations of "good-for/ bad-for" annotation (Anand and Reschke, 2010; Deng et al., 2013), when statistical parsing itself has spent the last twenty-five years staggering through a linguistically induced delirium as it attempts to document any of its putative advances without recourse to clear empirical evidence that PTB-style syntactic derivations are a reliable approximation of seman-

tic content or structure.

We submit, in light of our experience with the present study, that the most crucial obstacle facing the state of the art in sentiment analysis is not a granularity problem, nor a pattern recognition problem, but an evaluation problem. Those evaluations must be task-specific to be reliable, and sentiment analysis, in spite of our careless use of the term in the NLP community, is not a task. Stock trading is a task — one of many in which a sentiment analyzer is a potentially useful component. This paper provides an example of how to test that utility.

## 2 Related Work in Financial Sentiment Analysis

Studies confirming the relationship between media and market performance date back to at least Niederhoffer (1971), who looked at NY Times headlines and determined that large market changes were more likely following world events than on random days. Conversely, Tetlock (2007) looked at media pessimism and concluded that high media pessimism predicts downward prices. Tetlock (2007) also developed a trading strategy, achieving modest annualized returns of 7.3%. Engle and Ng (1993) looked at the effects of news on volatility, showing that bad news introduces more volatility than good news. Chan (2003) claimed that prices are slow to reflect bad news and stocks with news exhibit momentum. Antweiler and Frank (2004) showed that there is a significant, but negative correlation between the number of messages on financial discussion boards about a stock and its returns, but that this trend is economically insignificant. Aside from Tetlock (2007), none of this work evaluated the effectiveness of an actual sentiment-based trading strategy.

There is, of course, a great deal of work on automated sentiment analysis itself; see Pang and Lee (2008) for a survey. More recent developments germane to our work include the use of information retrieval weighting schemes (Paltoglou and Thelwall, 2010), with which accuracies of up to 96.6% have models based upon Latent Dirichlet Allocation (LDA) (Lin and He, 2009).

There has also been some work that analyzes the sentiment of financial documents without actually using those results in trading strategies (Koppel and Shtrimberg, 2004; Ahmad et al., 2006; Fu et al., 2008; O'Hare et al., 2009; Devitt and Ahmad, 2007; Drury and Almeida, 2011). As to the relationship between sentiment and stock price, Das and Chen (2007) performed sentiment analysis on discussion board posts. Using this, they built a "sentiment index" that computed the time-varying sentiment of the 24 stocks in the Morgan Stanley High-Tech Index (MSH), and tracked how well their index followed the aggregate price of the MSH itself. Their sentiment analyzer was based upon a voting algorithm, although they also discussed a vector distance algorithm that performed better. Their baseline, the Rainbow algorithm, also came within 1 percentage point of their reported accuracy. This is one of the very few studies that has evaluated sentiment analysis itself (as opposed to a sentiment-based trading strategy) against market returns (versus gold-standard sentiment annotations). Das and Chen (2007) focused exclusively on discussion board messages and their evaluation was limited to the stocks on the MSH, whereas we focus on Reuters newswire and evaluate over a wide range of NYSE-listed stocks and market capitalization levels.

Butler and Keselj (2009) try to determine sentiment from corporate annual reports using both character n-gram profiles and readability scores. They also developed a sentiment-based trading strategy with high returns, but do not report how the strategy works or how they computed the returns, making the results difficult to compare to ours. Basing a trading strategy upon annual reports also calls into question the frequency with which the trading strategy could be exercised.

The work most similar to ours is Zhang and Skiena's (2010). They look at both financial blog posts and financial news, forming a market-neutral trading strategy whereby each day, companies are ranked by their reported sentiment. The strategy then goes long and short on equal numbers of positive- and negative-sentiment stocks, respectively. They conduct their trading evaluation over the period from 2005 to 2009, and report a yearly return of roughly 30% when using news data, and yearly returns of up to 80% when they use Twitter and blog data. Crucially, they trade based upon the ranked relative order of documents by sentiment rather than upon the documents' raw sentiment scores.

Zhang and Skiena (2010) compare their strategy to two baselines. The "Worst-sentiment" Strategy trades the opposite of their strategy: short

on positive-sentiment stocks and long on negative sentiment stocks. The "Random-selection" Strategy randomly picks stocks to go long and short on. As trading strategies, these baselines set a very low standard. Our evaluation uses standard trading benchmarks such as momentum trading and holding the S&P, as well as oracle trading strategies over the same holding periods.

## 3 Method and Materials

### 3.1 News Data

Our dataset combines two collections of *Reuters* news documents. The first was obtained for a roughly evenly weighted collection of 22 small-, mid- and large-cap companies, randomly sampled from the list of all companies traded on the NYSE as of $10^{th}$ March, 1997. The second was obtained for a collection of 20 companies randomly sampled from those companies that were publicly traded in March, 1997 and still listed on $10^{th}$ March, 2013. For both collections of companies, we collected every chronologically third Reuters news document about them from the period March, 1997 to March, 2013. The news articles prior to $10^{th}$ March, 2005 were used as training data, and the news articles on or after $10^{th}$ March, 2005 were reserved as testing data.[1] We split the dataset at a fixed date rather than randomly in order not to incorporate future news into the classifier through lexical choice.

In total, there were 1256 financial news documents. Each was labelled by two human annotators as being negative, positive, or neutral in sentiment. The annotators were instructed to gauge the author's belief about the company, rather than to make a personal assessment of the company's prospects. Only the 991 documents that were labelled twice as negative or positive were used for training and evaluation.

### 3.2 Sentiment Analysis Algorithm

For each selected document, we first filter out all punctuation characters and the most common 429 stop words. Because this is a document-level sentiment scoring task, not sentence-level, our sentiment analyzer is a support-vector machine with a linear kernel function implemented using SVM$^{light}$ (Joachims, 1999), using all of its default parameters.[2] We have experimented with raw term frequencies, binary term-presence features, and term frequencies weighted by the BM25 scheme, which had the most resilience in the study of information-retrieval weighting schemes for sentiment analysis by Paltoglou and Thelwall (2010). We performed 10 fold cross-validation on the training data, constructing our folds so that each contains an approximately equal number of negative and positive examples. This ensures that we do not accidentally bias a fold.

| Representation | Accuracy |
|---|---|
| bm25_freq | 81.143% |
| term_presence | 80.164% |
| bm25_freq_sw | 79.827% |
| freq_with_sw | 75.564% |
| freq | 79.276% |

Table 1: Average 10-fold cross validation accuracy of the sentiment classifier using different term-frequency weighting schemes. The same folds were used in all feature sets.

Pang et al. (2002) use word presence features with no stop list, instead excluding all words with frequencies of 3 or less. Pang et al. (2002) normalize their word presence feature vectors, rather than term weighting with an IR-based scheme like BM25, which also involves a normalization step. Pang et al. (2002) also use an SVM with a linear kernel on their features, but they train and compute sentiment values on film reviews rather than financial texts, and their human judges also classified the training films on a scale from 1 to 5, whereas ours used a scale that can be viewed as being from -1 to 1, with specific qualitative interpretations assigned to each number. Antweiler and Frank (2004) use SVMs with a polynomial kernel (of unstated degree) to train on word frequencies relative to a three-valued classification, but they only count frequencies for the 1000 words with the highest mutual information scores relative to the classification labels. Butler and Keselj (2009) also use an SVM trained upon a very different set of features, and with a polynomial kernel of degree

[1] An anonymous reviewer expressed concern about chronological bias in the training data relative to the test data because of this decision. While this may indeed influence our results, ecological validity requires us to situate all training data before some date, and all testing data after that date, because traders only have access to historical data before making a future trade.

[2] There has been one important piece of work (Tang et al., 2015) on neural computing architectures for document-level sentiment scoring (most neural computing architectures for sentiment scoring are sentence-level), but the performance of this type of architecture is not mature enough to replace SVMs just yet.

3.

As a sanity check, we measured our sentiment analyzer's accuracy on film reviews by training and evaluating on Pang and Lee's (2004) film review dataset, which contains 1000 positively and 1000 negatively labelled reviews. Pang and Lee conveniently labelled the folds that they used when they ran their experiments. Using these same folds, we obtain an average accuracy of 86.85%, which is comparable to Pang and Lee's 86.4% score for subjectivity extraction. The purpose of this comparison is simply to demonstrate that our implementation is a faithful rendering of Pang and Lee's (2004) algorithm.

Table 1 shows the performance of SVM with BM25 weighting on our Reuters evaluation set versus several baselines. All baselines are identical except for the term weighting schemes used, and whether stop words were removed. As can be observed, SVM-BM25 has the highest sentiment classification accuracy: 80.164% on average over the 10 folds. This compares favourably with previous reports of 70.3% average accuracy over 10 folds on financial news documents (Koppel and Shtrimberg, 2004). We will nevertheless adhere to normalized term presence for now, in order to stay close to Pang and Lee's (2004) implementation.

### 3.3 Trading Algorithm

Overall, our trading strategy is simple: go long when the classifier reports positive sentiment in a news article about a company, and short when the classifier reports negative sentiment.

We will embed the aforementioned sentiment analyzer into three different trading algorithms. In Section 4.1, we use the discrete polarity returned by the classifier to decide whether go long/abstain/short a stock. In Section 4.2.1 we instead use the distance of the current document from the classifier's decision boundary reported by the SVM. These distances do have meaningful interpretations apart from their internal use in assigning class labels. Platt (Platt, 1999) showed that they can be converted into posterior probabilities, for example, by fitting a sigmoid function onto them, but we will simply use the raw distances. In Section 4.2.2, we impose a safety zone onto the interpretation of these raw distance scores.

## 4 Experiments

In the experiments of this section, we will evaluate an entire trading strategy, which includes the sentiment analyzer and the particulars of the trading algorithm itself. The purpose of these experiments is to refine the trading strategy itself and so the sentiment analyzer will be held constant. In Section 4.3, we will hold the trading strategy constant, and instead vary the document representation features in the underlying sentiment analyzer.

In all three experiments, we compare the per-position returns of the following four standard strategies, where the number of days for which a position is held remains constant:

1. The momentum strategy computes the price of the stock $h$ days ago, where $h$ is the holding period. Then, it goes long for $h$ days if the previous price is lower than the current price. It goes short otherwise.
2. The S&P strategy simply goes long on the *S&P 500* for the holding period. This strategy completely ignores the stock in question and the news about it.
3. The oracle S&P strategy computes the value of the *S&P 500* index $h$ days into the future. If the future value is greater than the current day's value, then it goes long on the *S&P 500* index. Otherwise, it goes short.
4. The oracle strategy computes the value of the stock $h$ days into the future. If the future value is greater than the current day's value, then it goes long on the stock. Otherwise, it goes short.

The oracle and oracle S&P strategies are included as toplines to determine how close the experimental strategies come to ones with perfect knowledge of the future. "Market-trained" is the same as "experimental" at test time, but trains the sentiment analyzer on the market return of the stock in question for $h$ days following a training article's publication, rather than the article's annotation.

### 4.1 Experiment One: Utilizing Sentiment Labels

Given a news document for a publicly traded company, the trading agent first computes the sentiment class of the document. If the sentiment is positive, the agent goes long on the stock on the date the news is released; if negative, it goes short.

| Strategy | Period | Return | S. Ratio |
|---|---|---|---|
| Experimental | 30 days | -0.037% | -0.002 |
| | 5 days | 0.763% | 0.094 |
| | 3 days | 0.742% | 0.100 |
| | 1 day | 0.716% | 0.108 |
| Momentum | 30 days | 1.176% | 0.066 |
| | 5 days | 0.366% | 0.045 |
| | 3 days | 0.713% | 0.096 |
| | 1 day | 0.017% | -0.002 |
| S&P | 30 days | 0.318% | 0.059 |
| | 5 days | -0.038% | -0.016 |
| | 3 days | -0.035% | -0.017 |
| | 1 day | 0.046% | 0.036 |
| Oracle S&P | 30 days | 3.765% | 0.959 |
| | 5 days | 1.617% | 0.974 |
| | 3 days | 1.390% | 0.949 |
| | 1 day | 0.860% | 0.909 |
| Oracle | 30 days | 11.680% | 0.874 |
| | 5 days | 5.143% | 0.809 |
| | 3 days | 4.524% | 0.761 |
| | 1 day | 3.542% | 0.630 |
| Market-trained | 30 days | 0.286% | 0.016 |
| | 5 days | 0.447% | 0.054 |
| | 3 days | 0.358% | 0.048 |
| | 1 day | 0.533% | 0.080 |

Table 2: Returns and Sharpe ratios for the Experimental, baseline and topline trading strategies over 30, 5, 3, and 1 day(s) holding periods.

All trades are made based on the adjusted closing price on this date. We evaluate the performance of this strategy using four different holding periods: 30, 5, 3, and 1 day(s).

The returns and Sharpe ratios are presented in Table 2 for the four different holding periods and the five different trading strategies. The Sharpe ratio is a return-to-risk ratio, with a high value indicating good return for relatively low risk. The Sharpe ratio is calculated as: $S = \frac{E[R_a - R_b]}{\sqrt{\text{var}(R_a - R_b)}}$, where $R_a$ is the return of a single asset and $R_b$ is the risk-free return of a 10-year U.S. Treasury note.

The returns from this experimental trading system are fairly low, although they do beat the baselines. A one-way ANOVA test among the experimental, momentum and S&P strategies using the percent returns from the individual trades yields p values of 0.06493, 0.08162, 0.1792, and 0.4164, respectively, thus failing to reject the null hypothesis that the returns are not significantly higher.[3]

[3]An anonymous reviewer observed that Tetlock (2007) showed a statistically significant improvement from the use of sentiment, apparently contradicting this result. Tetlock's (2007) sentiment-based trading strategy used a safety zone (see Section 4.2.2), and was never compared to a realistic baseline or control strategy. Instead, Tetlock's (2007) significance test was conducted to demonstrate that his returns (positive in 12 of 15 calendar years of historical market data) were unlikely to have been generated by chance from a normal distribution centred at zero.

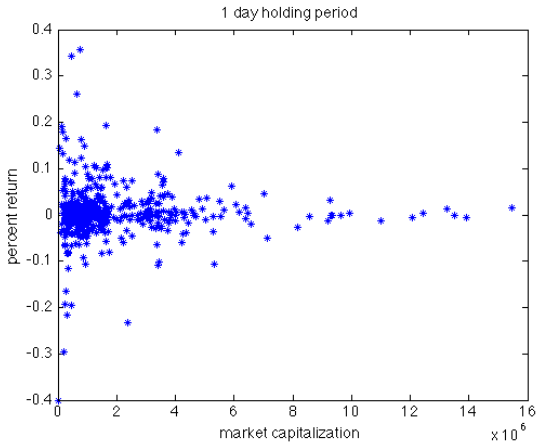


Figure 1: Percent returns for 1 day holding period versus market capitalization of the traded stocks.

Furthermore, the means and medians of all three trading strategies are approximately the same and centred around 0. The standard deviations of the experimental strategy and the momentum strategy are nearly identical, differing only in the thousandths digit. The standard deviations for the S&P strategy differ from the other two strategies due to the fact that the strategy buys and sells the entire S&P 500 index and not the individual stocks described in the news articles. There is, in fact, no convincing evidence that discrete sentiment class leads to an improved trading strategy from this or any other study with which we are familiar, based on their published details. One may note, however, that the returns from the experimental strategy have slightly higher Sharpe ratios than either of the baselines.

One may also note that using a sentiment analyzer mostly beats training directly on market data. This vindicates using sentiment annotation as an information source.

Figure 1 shows the market capitalizations of each individual trade's companies plotted against their percent return with a 1 day holding period. The correlation between the two variables is not significant. Returns for the other holding periods are similarly dispersed.

The importance of having good baselines is demonstrated by the fact that when we annualize our returns for the 3-day holding period, we get 70.086%. This number appears very high, but the annualized return from the momentum strategy is

70.066%[4], which is not significantly lower.

Figure 2 shows the percent change in share value plotted against the raw SVM score for the different holding periods. We can see a weak correlation between the two. For the 30 days, 5 days, 3 days, and 1 day holding periods, the correlations are 0.017, 0.16, 0.16, and 0.16, respectively. The line of best fit is shown. This prompts our next experiment.

### 4.2 Utilizing SVM scores

#### 4.2.1 Experiment Two: Variable Single Threshold

Before, we labelled documents as positive (negative) when the score was above (below) 0, because 0 was the decision boundary. But 0 might not be the best threshold, $\theta$, for high returns. To determine $\theta$, we divided the evaluation dataset, i.e. the dataset with news articles dated on or after March 10, 2005, into two folds having an equal number of documents with positive and negative sentiment. We used the first fold to determine $\theta$ and traded using the data from the second fold and $\theta$. For every news article, if the SVM score for that article is above (below) $\theta$, then we go long (short) on the appropriate stock on the day the article was released. A separate theta was determined for each holding period. We varied $\theta$ from $-1$ to 1 in increments of 0.1.

Using this method, we were able to obtain significantly higher returns. In order of 30, 5, 3, and 1 day holding periods, the returns were 0.057%, 1.107%, 1.238%, and 0.745% ($p < 0.001$ in every case). This is a large improvement over the previous returns, as they are average per-position figures.[5]

#### 4.2.2 Experiment Three: Safety Zones

For every news item classified, SVM outputs a score. For a binary SVM with a linear kernel function $f$, given some feature vector $\mathbf{x}$, $f(\mathbf{x})$ can be viewed as the signed distance of $\mathbf{x}$ from the decision boundary (Boser et al., 1992). It is then possibly justified to interpret raw SVM scores as degrees to which an article is positive or negative.

As in the previous section, we separate the evaluation set into the same two folds, only now we use two thresholds, $\theta \geq \zeta$. We will go long when the SVM score is above $\theta$, abstain when the SVM score is between $\theta$ and $\zeta$, and go short when the SVM score is below $\zeta$. This is a strict generalization of the above experiment, in which $\zeta = \theta$.

For convenience, we will assume in this section that $\zeta = -\theta$, leaving us again with one parameter to estimate. We again vary $\theta$ from 0 to 1 in increments of 0.1. Figure 3 shows the returns as a function of $\theta$ for each holding period on the development dataset. If we increased the upper bound on $\theta$ to be greater than 1, then there would be too few trading examples (less than 10) to reliably calculate the Sharpe ratio. Using this method with $\theta = 1$, we were able to obtain even higher returns: 3.843%, 1.851%, 1.691, and 2.251% for the 30, 5, 3, and 1 day holding periods, versus 0.057%, 1.107%, 1.238%, and 0.745% in the second task-based experiment.

| Representation | Accuracy | 30 days | 5 days | 3 days | 1 day |
|---|---|---|---|---|---|
| term_presence | 80.164% | 3.843% | 1.851% | 1.691% | 2.251% |
| bm25_freq | 81.143% | 1.110% | 1.770% | 1.781% | 0.814% |
| bm25_freq_dnc | 62.094% | 3.458% | 2.834% | 2.813% | 2.586% |
| bm25_freq_sw | 79.827% | 0.390% | 1.685% | 1.581% | 1.250% |
| freq | 79.276% | 1.596% | 1.221% | 1.344% | 1.330% |
| freq_with_sw | 75.564% | 1.752% | 0.638% | 1.056% | 2.205% |

Table 3: Sentiment classification accuracy (average 10-fold cross-validation) and trade returns of different feature sets and term frequency weighting schemes in Exp. 3. The same folds were used for the different representations. The non-annualized returns are presented in columns 3-6.

### 4.3 Experiment Four: Feature Selection

In our final experiment, let us now hold the trading strategy fixed (at the third one, with safety zones) and turn to the underlying sentiment analyzer. With a good trading strategy in place, it is clearly possible to vary some aspect of the sentiment analyzer in order to determine its best setting in this context. We will measure both market return and classifier accuracy to determine whether they agree. Is the latter a suitable proxy for the former? Indeed, we may hope that classifier accuracy will be more portable to other possible tasks, but then it must at least correlate well with task-based performance.

In addition to evaluating those feature sets attempted in Section 3.2, we now hypothesize that the passive voice may be useful to emphasize in our representations, as the existential passive can be used to evade responsibility. So we add to the

[4]The momentum strategy has a different number of possible trades in any actual calendar year because it is a function of the holding period.

[5]Training directly on market data, by comparison, yields -0.258%, -0.282%, -0.036% and -0.388%, respectively.

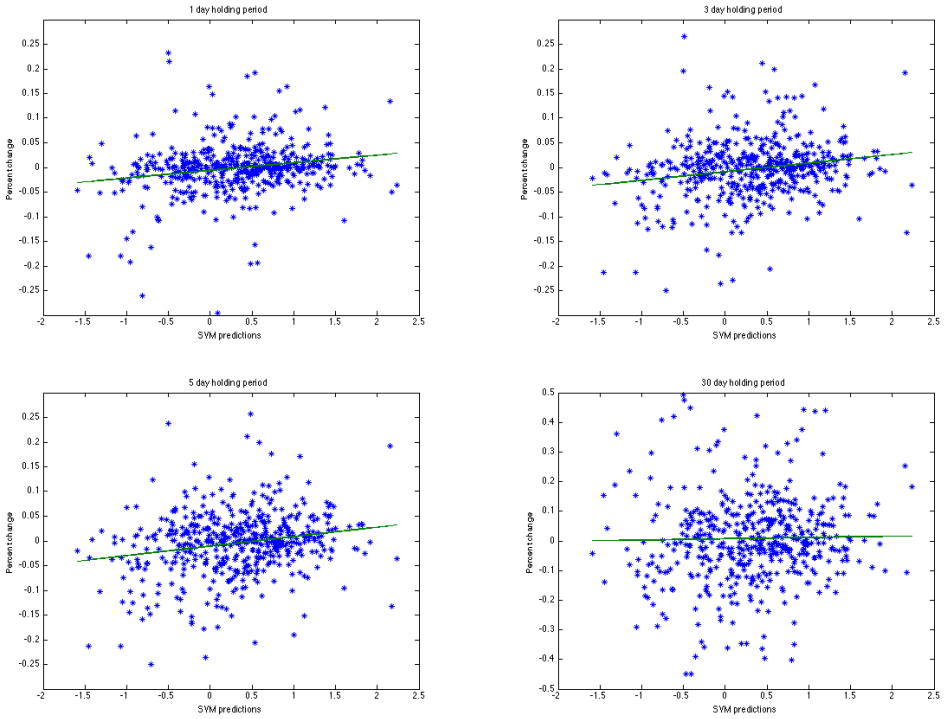


Figure 2: Percent change of trade returns plotted against SVM values for the 1, 3, 5, and 30 day holding periods in Exp. 1. Graphs are cropped to zoom in.

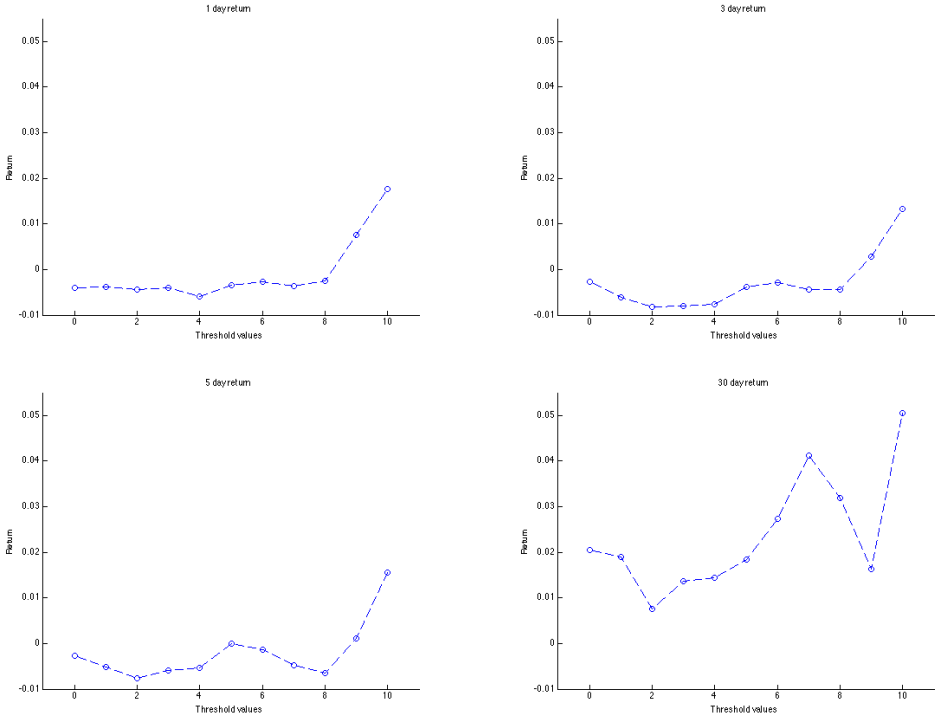


Figure 3: Returns for different thresholds on the development data for 30, 5, 3, and 1 day holding periods in Exp. 2 with safety zone.

BM25 weighted vector the counts of word tokens ending in "n" or "d" as well as the total count of every conjugated form of the copular verb: "be", "is", "am", "are", "were", "was", and "been". These three features are superficial indicators of the passive voice. Clearly, we could have used a part-of-speech tagger to detect the passive voice more reliably, but we are more interested here in how well our task-based evaluation will correspond to a more customary classifier-accuracy evaluation, rather than finding the world's best indicators of the passive voice.

Table 3 presents returns obtained from these 6 feature sets. The feature set with BM25-weighted term frequencies plus the number of copulars and tokens ending in "n", "d" (bm25_freq_dnc) yields higher returns than any other representation attempted on the 5, 3, and 1 day holding periods, and the second-highest on the 30 days holding period. But it has the worst classification accuracy by far: a full 18 percentage points below term presence. This is a very compelling illustration of how misleading an intrinsic evaluation can be.

## 5 Conclusion

In this paper, we examined sentiment analysis applied to stock trading strategies. We built a binary sentiment classifier that achieves high accuracy when tested on movie data and financial news data from *Reuters*. In four task-based experiments, we evaluated the usefulness of sentiment analysis to simple trading strategies. Although high annual returns are achieved simply by utilizing sentiment labels while trading, they can be increased by incorporating the output of the SVM's decision function. But classification accuracy alone is not an accurate predictor of task-based performance. This calls into question the suitability of intrinsic sentiment classification accuracy, particularly (as here) when the relative cost of a task-based evaluation may be comparably low. We have also determined that training on human-annotated sentiment does in fact perform better than training on market returns themselves. So sentiment analysis is an important component, but it must be tuned against task data.

Our price data only included adjusted opening and closing prices and most of our news data contain only the date of the article, with no specific time. This limits our ability to test much shorter-term trading strategies.

Deriving sentiment labels for supervised training is an important topic for future study, as is inferring the sentiment of published news from stock price fluctuations instead of the reverse. We should also study how "sentiment" is defined in the financial world. This study has used a rather general definition of news sentiment, and a more precise definition may improve trading performance.

## Acknowledgments

This research was supported by the Canadian Network Centre of Excellence in Graphics, Animation and New Media (GRAND).

## References

Khurshid Ahmad, David Cheng, and Yousif Almas. 2006. Multi-lingual sentiment analysis of financial news streams. In *Proceedings of the 1st International Conference on Grid in Finance*.

Pranav Anand and Kevin Reschke. 2010. Verb classes as evaluativity functor classes. In *Interdisciplinary Workshop on Verbs: The Identification and Representation of Verb Features (Verb 2010)*.

Werner Antweiler and Murray Z Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3):1259–1294.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, New York, NY, USA. ACM.

Matthew Butler and Vlado Keselj. 2009. Financial forecasting using character n-gram analysis and readability scores of annual reports. In *Proceedings of Canadian AI'2009*, Kelowna, BC, Canada, May.

Wesley S. Chan. 2003. Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260.

Sanjiv R. Das and Mike Y. Chen. 2007. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388.

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125. Association for Computational Linguistics.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the ACL*.

Brett Drury and J. J. Almeida. 2011. Identification of fine grained feature based event and sentiment phrases from business news stories. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 27:1–27:7, New York, NY, USA. ACM.

Robert F. Engle and Victor K. Ng. 1993. Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5):1749–1778.

Tak-Chung Fu, Ka ki Lee, Donahue C. M. Sze, Fu-Lai Chung, Chak man Ng, and Chak man Ng. 2008. Discovering the correlation between stock time series and financial news. In *Web Intelligence*, pages 880–883.

Thorsten Joachims. 1999. Making large-scale svm learning practical. advances in kernel methods-support vector learning, b. schölkopf and c. burges and a. smola.

Moshe Koppel and Itai Shtrimberg. 2004. Good news or bad news? let the market decide. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 86–88. Press.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.

Victor Niederhoffer. 1971. The analysis of world events and stock prices. *Journal of Business*, pages 193–219.

Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F. Smeaton. 2009. Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion measurement*.

Georgios Paltoglou and Mike Thelwall. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the ACL*, pages 1386–1395. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

William F Sharpe. 1966. Mutual fund performance. *The Journal of business*, 39(1):119–138.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of EMNLP*, pages 1422–1432.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168.

Wenbin Zhang and Steven Skiena. 2010. Trading strategies to exploit blog and news sentiment. In *The 4th International AAAI Conference on Weblogs and Social Media*.