Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards

Werner Antweiler and Murray Z. Frank*

First version: March 2001

This version: August 21, 2001

Abstract

Literally millions of messages have been posted on internet stock message boards. Financial press reports claim that these postings can move markets. We studied message posting on Yahoo! Finance and Raging Bull for the firms that were in the Dow Jones Industrial Average and the Dow Jones Internet Commerce Index during the year 2000. Using computional linguistics methods we measure the bullishness of the messages. Significant predictive content was found between message posting and trading volume, between message posting and volatility, between the degree of bullishness of the messages and trading volume. These results were obtained even after taking news stories from the Wall Street Journal into account. *(JEL: G12, G14)*

Keywords: Volatility, Trading Volume, Internet Message Boards

*Both authors are at the Faculty of Commerce and Business Administration, University of British Columbia, 2053 Main Mall, Vancouver BC V6T 1Z2, Canada. Corresponding author is Murray Frank. Phone: 604-822-8480. Fax: 604-822-8477. E-mail: murray.frank@commerce.ubc.ca. We would like to thank Richard Arnott, Alan Kraus, Jacob Sagi and the seminar audience at UBC for helpful comments. The second author thanks the B.I. Ghert Family Foundation for financial support. All errors are ours.

> "Internet message boards have come of age. [...] even investment pros are watching the message boards closely and profiting from it. With "posts" running in the millions, Internet message boards have become an essential part of the savvy investor's arsenal. [...] Internet messages really do move markets, for better or worse." (Weiss, 2000.)

1 Introduction

Many people are devoting a considerable amount of time and effort to create and to read the messages posted on internet stock message boards. News stories report that the message boards are having a significant impact. The Securities Exchange Commission has prosecuted people for internet messages. All this attention to internet stock messages caused us to wonder whether these messages actually contain financially relevant information.

This paper studies the information content of postings on the Yahoo! Finance and Raging Bull stock message boards during the year 2000. These are two of the largest and most prominent message boards. The sample of stocks are the 45 firms that together made up the Dow Jones Industrial Average (DJX), and the Dow Jones Internet Commerce Index (ECM). These firms were fairly large and well known.

Our focus is on the issue of whether internet messages really do move markets. We study the informativeness of both the observed message board activity levels, and the contents of the messages. A key issue is whether the markets tend to move first, or whether the message boards tend to move first. To the extent that the message boards move first they can be helpful in forecasting.

What patterns did we expect to find? There are plausible arguments in both directions. Many messages discuss recent events. Therefore it seems reasonable to expect that when unusual activity takes place in the market, people will post messages about what just happened. This is similar in spirit to ordinary news reports of unusual stock market activity. For this reason we expected to find significant effects from the stock market to the message boards.

Many messages assert that a particular stock is either a good buy or a bad buy. If the people posting messages are better informed than the marginal trader, there is the possibility of predictive content for returns. Assertions of predictive content have been made in the business press. Finance economists are likely to view such claims with skepticism.

Perhaps more interesting is a broader hypothesis. People think before they trade. Many people like to discuss their ideas. Suppose that they also post messages while they think about trading. This opens up the possibility that posting activity might lead trading volume. It is well known that volume and volatility often move together. This also opens the possibility that the messages could help forecast volatility.

Due to the nature of the data, we employ methods from computational linguistics. The reason is that our message board data contains more than 1.5 million text messages – far too many to interpret manually. We therefore use computer algorithms to interpret the messages. The oldest algorithm used to interpret text is called Naive Bayes. Another algorithm called Support Vector Machine has become very popular for use in many classification problems, including text classification. Both algorithms are used to code the individual messages as bullish, bearish, or neither. We then aggregate the codings into indices that measure the bullishness of each stock message board during each time period.

As might have been anticipated, the messages do not have any special forecasting ability for excess stock returns. The messages are remarkably bullish despite poor market performance for most of the firms in our sample. As shown in Figure (6), during the year 2000 the DJX declined by about 20% while the ECM declined by more than 70%. Yet, as can be seen in Figure (7), right through the year the messages were more bullish regarding the average ECM firm than about the average DJX firm.

There is more to the messages than just the content of the text. The simple fact that a message has been posted about a particular firm at a particular time could itself be informative. Sometimes many messages are posted. At other times few messages are posted. It turns out that more messages are posted during periods when market trading is also particularly active. When above average numbers of messages are posted, we also find that subsequent trading volume tends to be high. As conjectured, we also find that an above average number of messages forecasts high volatility.

A skeptic might hypothesize that the internet messages would merely rehash what was reported that morning in the Wall Street Journal or similar media. This raises the important issue of whether the message boards play an independent role. Given the large number of sources of real world information, this is a difficult challenge to meet fully. We have collected all articles published in the Wall Street Journal about the firms in our sample during the year 2000. We have checked the days surrounding the news stories, both for unusual market activity and for unusual message board activity.

The message boards have predictive content even controlling for the presence of articles in the Wall Street Journal. The message boards do not merely rehash the morning's news. The day before a story is published proves to be a more important date than the day of the publication. Presumably this reflects the fact that many stories are reported on the internet and on news wires the day before the published version. There is some evidence of unusual stock market behavior two days before Wall Street Journal news stories. However there is no evidence that this is reflected on the message boards.

We are familiar with a small number of closely related previous studies. Bagnoli, Beneish and Watts (1999) compared the First Call analyst earning forecasts to unofficial "whispers." The whispers were collected from a number of sources including Internet web pages and news stories that reported the whisper forecasts. The analysts from First Call tended to underestimate corporate earnings announcements, while the "whispers" tended to be more accurate. Dewally (2000) collected stock recommendations from two newsgroups (misc.invest.stocks and alt.invest.penny-stocks). He found that there was not much predictive content in the forecasts on these newsgroups. The recommended stocks typically had strong prior performance.

Wysocki (1999) measured the cumulative message postings on Yahoo! Finance to July 1, 1998. He studied the cross sectional differences between firms in order to determine which firms had a large number of messages posted. In contrast to our paper, he did not attempt to assess the content of the messages. The firms with high posting activity were characterized by: high market valuation relative to fundamentals; high short seller activity; high trading volume; and high analyst following but low institutional holdings.

Tumarkin and Whitelaw (2001) study messages posted on Raging Bull for a sample of firms in the internet service sector. Using data at daily frequency they found not much of a link from the stock message boards to daily stock returns or trading volumes. In order to assess the opinions of the messages they use the "disclosure" feature of Raging Bull. The current paper differs from Tumarkin and Whitelaw (2001) in many ways. We include Dow Jones Industrial average firms rather than restricting attention to a sample of internet firms. We use intra-day data in order to allow for the role of day trading. We use computer algorithms to measure opinions directly rather than relying on the subsample of messages that provide self reported disclosure. As a result we are also able to compare results from the more popular Yahoo! Finance to those on Raging Bull.

The rest of the paper is organized as follows. Section (2) explains the reasoning behind a number of the decisions we made in designing the study. Section (3) discusses the messages and how we extracted information from the texts. In section (4) we describe a number of the basic features of the data. Section (5) provide the results of tests for temporal sequencing between the message boards and the financial market features. We conclude in section (6).

2 Defining the Scope of the Study

The first issue we face is the decision of which firms to study. We are interested in the information content, but not in market manipulation. There are many reports of attempted, and sometimes successful, market manipulation using the stock message boards. In one highly publicized case a 15 year old is reported to have made a fair bit of money using a rather traditional "pump and dump" approach. He apparently focused on small stocks using Yahoo! Finance.[1] Penny stocks are reported to be susceptible to such behavior. While market manipulation is an important topic, it is not our focus in this paper. We are interested in the simpler issue of information revelation for heavily traded securities.

Popular discussions of financial markets during the period studied, often distinguished "old economy" and "new economy" firms. It seemed plausible that these firms might have been affected differently by the messages. To represent the "old economy," we use the 30 firms that are in the Dow Jones Industrial Average (DJX). To represent the "new economy" we use the 15 firms that are included in the Dow Jones Internet Index (ECM). This gives us a sample of 45 large firms. To represent movements in the overall market we used the exchange traded fund that mimics the S&P 500 (SPY).

The second issue is the time period to study. News stories suggest that the importance of

[1] "After he had picked and bought his stock, he would write a single message about it and stick it up in as many places on Yahoo Finance as he could between 5 and 8 in the morning, when he left home for school. There were no explicit rules on Yahoo Finance, but there were constraints. The first was that Yahoo limited the number of messages he could post using one e-mail address. He would click onto Yahoo and open an account with one of his four AOL screen names; a few minutes later, Yahoo, mysteriously, would tell him that his messages could no longer be delivered. Eventually, he figured out that they must have some limit that they weren't telling people about. He got around it by grabbing another of his four AOL screen names and creating another Yahoo account. By rotating his four AOL screen names, he found he could get his message onto maybe 200 Yahoo message boards before school. He also found that when he went to do it the next time, with a different stock, Yahoo would no longer accept messages from his AOL screen names. So he was forced to create four more screen names and start over again. Yahoo never told him he shouldn't do this. "The account would be just, like, deleted," he said. "Yahoo never had a policy; it's just what I figured out." The S.E.C. accused Jonathan of trying to seem like more than one person when he promoted his stocks, but when you see how and why he did what he did, that is clearly false. (For instance, he ignored the feature on Yahoo that enables users to employ up to seven different "fictitious names" for each e-mail address.) It's more true to say that he was trying to simulate an appearance on CNBC." ("Jonathan Lebed: Stock Manipulator, S.E.C. Nemesis – and 15" By Michael Lewis, *New York Times*, February 25, 2001.)

the message boards is growing. So we decided to study the year 2000.

The third issue is the time frequency to use. Day traders have attracted considerable recent attention. News reports suggest that day traders are particularly involved with the internet stock message boards. It seems likely that their responses to postings might be observed within the trading day. Therefore we decided to use high frequency data provided by the TAQ database rather than relying on the daily data from CRSP.

The fourth issue is the choice of message boards. There are many message boards on the internet. We use data from Yahoo! Finance because it is reported to have the highest volume of posting activity. In order to ensure that our results are not unduly limited we also use data from Raging Bull, which is another of the popular message boards. Since different message boards may attract different people, there is no guarantee that the results will be the same on both sets of message boards. In fact the results did turn out to be quite similar.

Our data set includes more than 1.5 million messages about the 45 companies in our sample. Because this makes manual classification impractical, we turned to the use of computer algorithms to classify the messages.

3 Message Board Data and Classification

Messages were downloaded from the Yahoo! Finance (YF) and Raging Bull (RB) message boards using specialized software written by the authors. Messages were stored in a simple plain-text database format, one file per day per company. Each message is uniquely identified by the bulletin board code (YF or RB), the company's ticker symbol, and the message board sequence number. The file contents were then summarized in an index file which also lists the date and time of posting, the message's length in words, and the screen name of the originator of the message.

To understand the nature of the postings it is helpful to look at examples. Figure (1) provides two fairly typical examples of messages in the database format. Each message is dated and timed to the minute, has a title, and has a text. The text very often contains a predicted price change and at least some explanation for the prediction. Most of the explanations are fairly short. Table (1) shows that the number of words in a message is most frequently between 20 and 50. Relatively few messages have more than about 200 words.[2] It is fairly rare

[2]Many of the messages are rude and/or off-topic, and such messages are often long. On the other hand, a

for a message to have more than 500 words. More than 40% of the messages are posted by people who post only a single message.[3] However, there are some people who are very active and account for more than 50 messages each.

We use two methods for classifying the messages: Naive Bayes (NB) and Support Vector (SV). We started by using a training data set of 1,000 messages that we classified manually. We then filtered our entire sample of 1,559,621 messages through the classification software to obtain buy, hold, or sell signals for each message. These were in turn aggregated into time periods of 15-minutes, 1 hour and 1 day.

While stock prices maybe set efficiently, there is no economic force that would cause messages posted on internet stock message boards to be efficient. Accordingly, we did not expect the messages to provide unbiased forecasts. There are at least two aspects, a clientele issue and an institutional issue.

A simple categorization of stock market participants includes market makers, institutional investors, day traders, and ordinary private investors. It seems unlikely that either market makers or institutional investors would do much posting on internet stock message boards. Ordinary private investors are a very large and very diverse group, and so we expected the message boards to reflect activity from this group. Potentially most important however are the day traders. While their absolute number may be fairly small relative to ordinary investors, press reports suggest that many of them are extremely active on the stock message boards. Putting these groups together implies that we did not expect the message boards to reflect the views of a random sample of all traders.

There is an institutional aspect that may also be important. People who hold either a long position or a short position are likely to be particularly interested in a given stock. They need to decide whether to enhance their exposure, or to unwind it. People who hold a zero position may be less likely to pay attention to a particular message board. Current institutions make it much easier for a small trader to hold a long position than a short position. Thus we expected to find a bullish tone on average.

For these reasons we approach the data knowing that while market prices reflect all market participants, what we observe on the boards probably will not. The message boards probably

reasonable proportion, while perhaps speculative, at least provided some discussion and a discernible prediction for the firm in question.

[3]We only observe the chosen screen name rather than the author's actual name. Therefore if one author posts messages using more than one screen name we will count these as if they were separate authors.

reflect a particular segment of the market. As will be shown, this perspective does seem to help account for several features of the data.

3.1 Naive Bayes Message Coding

The Naive Bayes algorithm is the oldest of the algorithms used to classify documents. Lewis (1998) provides a perspective of the history of the algorithm. It continues to attract interest and further refinement. Recent related papers include McCallum and Nigam (1998) and McCallum, Nigam, Rennie and Seymore (2000). Beyond the fact that Naive Bayes has a long and rather successful track record in text classification, there is another reasons for our attention to this approach. Using Bayes rule in this manner has a natural attraction given the wide use of Bayesian methods in financial econometrics.

For Naive Bayes text classification we have employed the *Rainbow* package developed by McCallum (1996).[4] The key assumption underlying the NB classification method is that occurrences of words are independent of each other. The assumption of independence among words is the reason that the algorithm is referred to as "naive". Even though this is a highly unrealistic assumption, NB performs rather well in practice.[5]

In the context of text classification, Naive Bayes can be understood most easily as a straightforward mechanism of updating odds ratios. Consider a stream of words W_i that are found either in a message of type T or its anti-type $\tilde{T}$. Let m be the number of occurrences of this word in type T, and let $\tilde{m}$ be the number of occurrences in anti-type $\tilde{T}$. Further let n and $\tilde{n}$ denote the total number of words in classes T and $\tilde{T}$, respectively. For words found in messages from the training set we observe the conditional probabilities $P(W_i|T) = m_i/n_i$ and $P(W_i|\tilde{T}) = \tilde{m}_i/\tilde{n}_i$. Now consider Bayes' rule, updating our prior $P(T|W_{i-1})$ to posterior $P(T|W_i)$ when we observe word W_i and thus $P(W_i|T)$ and $P(W_i|\tilde{T})$:

$$P(T|W_i) = \frac{P(T|W_{i-1})P(W_i|T)}{P(T|W_{i-1})P(W_i|T) + (1 - P(T|W_{i-1}))P(W_i|\tilde{T})}. \tag{1}$$

[4]This software can be downloaded freely for academic purposes from the web at http://www.cs.cmu.edu/-˜mccallum/bow/.

[5]This approach is an example of a "bag of words" approach to text classification. This approach makes no direct use of the gramatical structure. As an empirical matter it has been found that a surprisingly small amount is gained at substantial cost by attempting to exploit gramatical structure in the algorithms. For a helpful discussion of the various approaches to analyzing text see Manning and Schutze (1999).

That is easily rewritten in odds-ratios form as

$$\frac{P(T|W_i)}{1 - P(T|W_i)} = \frac{P(T|W_{i-1})}{1 - P(T|W_{i-1})} \cdot \frac{P(W_i|T)}{P(W_i|\tilde{T})} \tag{2}$$

with $P(T|W_0) \equiv P(T)$. Classifying a document thus amounts to multiplying odds ratios when processing the document word by word.

For reasons of computational accuracy, it is however common practice to add up logs of odds ratios

$$P(T|W_N) = P(T) \exp\left[\sum_{i=1}^{N} \log\left(\frac{P(W_i|T)}{P(W_i|\tilde{T})}\right)\right] \tag{3}$$

where N is the number of words in a given document. Adding logs of odds ratios avoids the problem of computational "underflow" or "overflow" errors, which could easily arise when odds ratios are multiplied directly in a long message. The prior $P(T)$ is based on the document frequencies for each of our three classes. The problem with equation (1) is that either $P(W_i|T)$ or its anti-class counterpart may be zero. In this case, a method known as *Laplace smoothing* is applied to replace the zero-value with estimates $\mathcal{E}(P(W_i|T)) = (1 + m_i)/(1 + n_i)$, and likewise for $P(W_i|\tilde{T})$.

Usage of the Rainbow software package proceeded in three steps. First we split the 1,000 messages into buy, sell, and hold messages stored in individual directories. In the second step we ran the rainbow utility to process the messages in the training data set using the `--method=naivebayes` and `--prune-vocab-by-infogain=1000` options. The latter restricts the number of words in the vocabulary to the top 1,000 words as ranked by the average mutual information with the class variable.[6] After training is complete, rainbow is put into server mode and individual messages k containing words W^k are sent from a client program to the server for evaluation. The server returns three probabilities $P(T_c|W_n^k)$ for each of the three categories T_c (buy=+1, hold=0, and sell=–1), and we choose the classification with the highest probability. Our maximum likelihood classification rule is thus

$$x_k = \arg\max_{T_c} P(T_c|W_n^k) \tag{4}$$

At this stage each individual message has been classified. Before turning to the aggregation step, we discuss the Support Vector Machine coding.

[6] See Cover and Thomas (1991) for details.

3.2 Support Vector Message Coding

The Support Vector Machine (SVM) approach stems from the work on statistical learning and computer classification problems. This method has been applied successfully to a broad range of classification problems. It has been successfully used for text classification by academic scholars such as Joachims (1999), as well as by scholars at Microsoft Research, see Dumais, Platt, Heckerman, and Sahami (1998). Given the reports of successful use of this method in text classification, we employ SVM as a second classification method. Using two different methods for text classification also helps to ensure that our results are "robust" with respect to the choice of classification method. For SVM text classification we use software by Joachims (1998).[7]

Similar to Naive Bayes, but unlike other machine-learning techniques such as neural networks, SVM can be represented and analyzed algebraically. While a complete discussion of this method lies beyond the scope of this paper, we point out the basic idea behind SVM.

SVM transforms texts into "feature vectors," where each feature corresponds to a word in the text to be classified, along with an attribute that describes the word's importance.[8] It is customary to reduce the feature space to avoid "overfitting" of the data. We only include words that occur in at least 1,000 of our 1.5 million messages. The words are ranked by what is known as a *minimum information criterion*, and we choose only the top 1,000 words from this list to form our feature space. The SVM algorithm is then optimizing an objective function to calculate the hyperplane that optimally separates the feature space (words) into a class and its anti-class.[9]

SVM classification delivers values $r_k(B)$ and $r_k(S)$ for each message k and signals buy (B) and sell (S), where positive and negative numbers indicate whether or not the message

[7]This software is freely available on the internet for academic purposes and can be downloaded from http://ais.gmd.de/~thorsten/svm_light/.

[8]In the context of our study, the attribute that corresponds to each word i is the *normalized inverse document frequency* $\mathrm{IDF}_i = \log(M/m_i)\left[\sum_{j=1}^{N}\log(M/m_j)\right]^{-1/2}$, where M is the number of messages to be classified, m_i is the number of documents in which word i occurs, and N is the dimension of the word space.

[9]In detail, we split our training sample into buy and sell messages and create appropriate dictionary files in which words are coded as numbers. Thse two dictionary files are used to train the classification system along with a data file which describes the words attributes. We have used the default parameters of Joachim's SVM system except for the choice of kernel function, where we have opted for a polynomial kernel function instead of a linear kernel function.

belongs in a given class. To aggregate these two numbers we apply the decision criterion

$$x_k = \begin{cases} \text{B} & \text{if } r_k(B) > r_k(S) \geq 0 \\ \text{S} & \text{if } r_k(S) > r_k(B) \geq 0 \\ \text{H} & \text{otherwise} \end{cases} . \tag{5}$$

We then set $B = +1$, $S = -1$, and $H = 0$ in order to construct an index of the message board's bullishness for a particular stock. As we will show below, this coding is also particularly helpful in deriving a meaningful and easy-to-calculate "agreement index" which captures the degree to which message authors concur in their bullish or bearish sentiment.

Table (2) shows the in-sample classification accuracy of the Naive Bayes and Support Vector methods. Our training data set of 1,000 messages contains 25.2% buy signals, 5.5% sell signals, and 69.3% hold signals. The in-sample accuracy is very high with both classification methods. The sell category is the smallest, with the Naive Bayes method slightly under-representing this category within sample. Out-of-sample Naive Bayes found a slightly smaller number of sell messages than did the Support Vector method.

3.3 Aggregation of the Coded Messages

For each of the two classification methods we aggregate the message classifications x_i in order to obtain a "bullishness signal" θ_t for each of our time intervals t. A bullishness signal is defined as

$$\theta_t \equiv \frac{\sum_{i \in \mathcal{D}(t)} w_i x_i}{\sum_{i \in \mathcal{D}(t)} w_i} \tag{6}$$

where $\mathcal{D}(t)$ denotes the document space for time period t.

In addition to equal weighting, we consider two unequal weighting schemes. Because longer messages might be more important than short ones, we have tried weighting the messages by their length (L_i). Some people post many very similar messages. Because repeated messages by the same author may have diminishing impact, we have also tried weighting each message by the inverse of the total number of messages posted the the author of the particular message (A_i). A bullishness signal θ is then defined as the weighted mean of the buy (+1) and sell (–1) signals x_i of each message.

We have also considered using weights based on the citation frequency of individual messages. This information is available for messages on the Yahoo! Finance boards, where messages may contain a single citation of a previous message. Less than 40% of messages had one or more citations. We do not employ such a weighting system in our analysis because citation weights are only determined ex-post. The number of citations that a particular message will receive is unknown at the time it is posted. This would cause serious problems for our time sequencing analysis. Another concern is that citations may not be genuine citations of an earlier message. This happens because some message board contributors find it more convenient to use an e-mail system's "reply" function instead of the "new message" function. As a result the number of citations can be a distorted measure.

The treatment of periods during which no messages are posted is a potential source of concern. If no new messages are posted does this mean that everyone agrees with the last message? Or does it simply mean that nobody has anything that they wish to say? In the reported results we assume that an absence of postings is a zero.[10]

Potentially interesting is the extent of disagreement among the messages posted. Disagreement can induce trading in some settings.[11] To measure agreement we proceeded as follows. First use the evaluation function $\delta(\cdot)$ which returns 1 when the expression in parenthesis is true and zero otherwise. Then define $S_t \equiv \sum_{i\in\mathcal{D}(t)} w_i\delta(x_i = 1) \geq 0$, $B_t \equiv \sum_{i\in\mathcal{D}(t)} w_i\delta(x_i = -1) \geq 0$ and then observe that equation (6) simplifies to $\theta_t = (B_t - S_t)/(B_t + S_t)$. By construction this implies $-1 \leq \theta_t \leq +1$. The variance of x_i during time interval t corresponding to (6) can then be calculated as

$$\sigma_t^2 \equiv \frac{\sum_{i\in\mathcal{D}(t)} w_i(x_i - \theta_t)^2}{\sum_{i\in\mathcal{D}(t)} w_i} = \frac{\sum_i w_i x_i^2}{\sum_i w_i} - \theta_t^2 = 1 - \theta_t^2 \tag{7}$$

In the last simplification step observe that $\forall i : x_i^2 = 1$, because x_i is either +1 or –1. This permits us to measure the prevailing level of agreement among message board contributors as an "agreement index"

$$\eta_t \equiv 1 - \sqrt{1 - \theta_t^2} \tag{8}$$

If disagreement produces trading then η_t should be negatively correlated with measures of

[10] Another possibility is to maintain the previous value instead of using a value of zero during a period without messages. We did some experiements with such a procedure. These indices are biased in the bullish direction. The empirical results obtained using this approach are considerably noisier.

[11] Harris and Raviv (1993) provide a model of such trading. The hypothesis that disagreement induces trading is often invoked in theory such as Allen and Gale (1999) and Daniel, Hirshleifer and Subrahmanyam (2001). In a standard rational expectations model however, there can be "no trade" theorems such as in the analysis of Milgrom and Stokey (1982). Related empirical evidence is provided by Kandel and Pearson (1995) and Bessembinder, Chan and Seguin (1996).

trading volume.

4 Basic Features of the Data

4.1 Messages

More messages are posted on Yahoo! Finance than on Raging Bull. Firms that are listed on NASDAQ and included in the ECM index generate more messages than do firms listed on the NYSE and included in the DJX. Intel and Microsoft are the only firms in our sample that are listed on NASDAQ and included in DJX. Many messages are posted about both of these firms.

Figure (2) shows the weekly level of message posting over the full year 2000. News reports in the earlier part of the year 2000 suggested that posting activity was increasing at a dramatic rate. News reports in the later parts of the year 2000 suggested that posting activity was falling off dramatically. Neither of these match what we observed for our sample of firms. There was some decline in activity during the late spring and the summer months. Otherwise message posting activity was reasonably stable over the year for our sample of firms.

Within the trading week there is a very strong weekend effect. Many fewer messages are posted during weekends.

4.2 Financial Data

Financial data are from the TAQ database for the 45 stocks and for the exchange traded fund that serves as our proxy for the market (SPY). We extracted and then aggregated the following information: a bid-ask midpoint at the end of each 15-minute time interval, the corresponding bid-ask spread at that point, the volume-weighted average trading price for that 15-minute interval, the corresponding volume-weighted volatility, the number of shares traded, and the number of transactions in each of three transaction value categories (below $100,000, between $100,000 and under $1,000,000, and $1,000,000 or above).

In scanning through the TAQ trades file we ignore information from exchanges other than NASDAQ or NYSE, we ignore trades or quotes with sequence numbers that are "out of sequence," we ignore trades which are marked as irregular, and we ignore opening and closing quotes. In addition, we filter out spreads that are negative or in excess of 40% of the bid-ask midpoint, and we filter out trades with prices in excess of 100% of the bid-ask midpoint for the relevant time period.

Our measure of the stock market index is the exchange traded fund that mimics the S&P 500 (ticker symbol SPY). This is a market factor that traders can buy and sell easily with low transaction costs. It also has the further advantage that we can observe its market price directly at the same frequency as the rest of the financial data.

As can be seen in Figure (3) there was a decline in the volume of stock trading over the year for our sample of firms. Comparison of Figure (2) with Figure (3) suggests that trading volume is often elevated during the same weeks that message posting is elevated.

4.3 Wall Street Journal

Using Lexis-Nexis we collected all articles about the firms that appeared in the Wall Street Journal during the year 2000. As shown in Table (3) there are no news stories about most firms on most days, while the occasional firm has more than a single news story on a given day. The main firm that had multiple news stories on the same date is Microsoft, which was involved in a controversial antitrust trial during the year. Table (3) also shows that while the message board postings are very heavy for the internet firms, the Wall Street Journal provides much more coverage for the typical Dow Jones Industrial Average firm. In other words there is a difference in emphasis between these two sources.

The Wall Street Journal is usually taken to be the statement of record and it is published each business day morning. Investors, however, commonly obtain information from news wire sources directly or from a range of sources that post news on the internet – including the web page of the Wall Street Journal. This kind of information commonly becomes available during the day before it is published in the Wall Street Journal. Therefore it is important to also examine the days surrounding the day that a story appears in the Wall Street Journal. We consider two days prior and one day subsequent to a news story day.

4.4 Time

The NYSE and NASDAQ are only open from 9:30 AM to 4:00 PM Eastern time. As is well known, the behavior at the market open and at the close of trading have somewhat different properties than during the rest of the day. In particular, as shown in Figure (4), trading volume tends to be lower during the middle of the day. As explained by Bacidore and Lipson (2001) the opening and closing procedures on the NYSE are different from trading during most of the day. The opening auction can take up to a half hour.

In addition to the different trading institutions there are also potentially different trader motivations to consider. Many small traders think about their portfolios during the evening when they are home from work. They may call their brokers or place automated trades through discount brokers before they start work in the morning. This will result in many trades at the market open. Near the end of the day institutional traders may wish to close out a position in order to avoid overnight risk. Managers of mutual funds may wish to prepare their portfolio for the end of the day valuation. These arguments are consistent with what is observed in Figure (4).

One might have imagined that a high proportion of the messages would be posted after dinner. However, this is not the dominant pattern. Figure (5) shows that message posting is concentrated during working hours, while the stock markets are open. This is suggestive of day trader activity, but it could also reflect people posting messages from their jobs.

Both the messages and the stock data are time stamped by the minute. This suggests using one minute as the time period. However, with this time period definition there will be no messages posted for most periods, and in many cases relatively few stocks traded. This large number of zeros can create a misleading impression. If one increases the amount of time in each time interval, several effects take place. First, there are fewer empty cells. Second, there are many fewer time periods and hence reduced "sample size." Third, with longer time periods some of the information that is included in a given period may become more stale since it is less recent. Fourth, the numbers within each time period are based on more evidence.

There is no clear best balance among these factors. Some traders really do follow the market on a minute by minute basis. Other investors may check the market daily, or even less frequently. With a market composed of a mixture of such people it is quite possible that different relationships could be revealed on different time scales. In the market microstructure literature many papers, such as Hasbrouck (1999) and Hasbrouck and Seppi (2001), use a 15 minute time period. Of course other papers define time units at different levels. With these considerations in mind we carried out tests using 15 minute, 1 hour, and 1 day time periods. The results from 1 hour do not add much further information, and so we report the results based on 15 minute and 1 day time period definitions.

There is another time problem to be considered. The markets are only open for part of the day, while the stock message boards are open 24 hours a day. Thus the two classes of data have different natural calendars. We have employed three strategies for dealing with this in our empirical analysis. First, much of our analysis focuses only on the time during which the markets are open. Second, we use time-of-day dummy variables throughout the analysis in an effort to control for deterministic time patterns. Third, we have carried out a number of additional tests to see if the messages posted while the markets were shut predict the market opening excess stock returns. We found no evidence that the opening excess returns are forecastable using the message board information, and so we do not report these test results in detail.

Figure (6) depicts the market performance of our versions of the DJX and ECM indices.[12] The DJX declined by about 20% while the ECM declined by more than 70%. The decline in the ECM reflects what is often described as the end of the internet bubble. Two of the firms in the ECM ceased trading during October due to mergers.

Figure (7) depicts movement of the bullishness signals for the DJX and ECM indices. Weekly average values are shown because the daily, and higher frequency, values ae much more volatile. The bullishness signals for the ECM held up remarkably strongly. Gradual decline was observed during the year.

A popular aphorism during the so-called internet bubble was "buy on the dips." There were many reports in the business press that at least some investors believed that following a "buy on the dips" strategy would be key to building long term wealth. Of course, this saying does not specify a particular definition of a dip, nor does it specify how much to buy. Nonetheless it may help explain the degree to which bullishness held up in the face of declining markets.

Comparison between Figures (6) and (7) is interesting. One might imagine that stock price declines would be matched by declines in bullishness. There were 3 very sharp declines in ECM, one in February, one from late March to mid-April and one from early September to mid-October. There is no simple tight connection between the declining stock prices and the associated bullishness indices that is apparent in these Figures. There does seem to have been a decline in bullishness during the fall relative to the earlier part of the year.

[12]We weighted the indices by market capitalization and standardized them at 100 at the start of the year 2000.

4.5 Descriptive Statistics

Table (3) reports a number of descriptive statistics for the companies in our sample. Comparing Yahoo! Finance to Raging Bull we find that more messages are posted for all firms, and the messages are often more bullish. The average message is longer on Raging Bull. In comparison to the ECM firms, the DJX firms have lower losses, lower volatility, lower activity levels on the stock message boards, and less bullishness expressed on the stock message boards.

The coverage of our sample of firms in the Wall Street Journal is not all that highly correlated with message posting. The ECM firms generate a great deal of message posting, but a much lower level of coverage by the Wall Street Journal.

Tables (4)-(6) provide a number of descriptive statistics. Price can be measured as either the average price at which shares actually traded during a time interval, or it can be measured as the midpoint of the bid-ask spread. Jones, Kaul and Lipson (1994) argue that the midpoint of the bid-ask spread is preferable because it avoids the bid-ask bounce. We follow their recommendation. Empirically the two are very highly correlated. The average price (midpoint of the bid and ask) at which our sample of stocks traded during the period under study was \$58.25. The highest price observed in the data was \$496.03 and the lowest price observed was \$0.17. This seems a remarkable range for firms that are all large enough to be included in major stock indices.

Table (5) provides a number of alternative time aggregations for the Yahoo! Finance message boards. 'No lag' refers to data constructed strictly within a 15 minute time interval. A '1 hour lag' means that the data is aggregated over the previous hour. Similarly '4 hour' and '1 day' lags report longer time aggregations. Of course the number of words and the number of messages increase as longer periods of time are included. On average 3.365 messages are posted every hour for each stock in our sample over the entire year on Yahoo! Finance. This is a fair bit higher than the 0.58 messages per hour on Raging Bull.

The bullishness measures increase as the length of time included in the measure increases. This effect is simply an artifact of the manner in which the indices are constructed. When no messages are posted in a period, neither a buy nor a sell is recorded for that period. When only a single message is recorded in a period, and that message is neither a buy nor a sell, then a value of zero is recorded. As the length of the time period increases less weight is given to such "accidental holds." Since there are more buy messages than sells, as more messages are added the mean tends to rise.[13]

In order to compute elasticities and to control for scaling, a number of the key variables are log transformed. In particular variables related to trading volume and to number of messages posted are calculated as $log(1 + x)$ in order to avoid taking the log of zero when x is zero.

4.6 Correlations

The correlation patterns in the data provide evidence of relevant information. Tables (7) and (8) report the correlations that are significantly different from zero at a 99% confidence level. Correlations for Yahoo! Finance and Raging Bull are reported separately, and they prove to be quite similar.

The correlations reported at the top of Table (7) are among the financial measures. These generally reproduce results that are already well-known from studies of earlier time periods. Volume is positively correlated with volatility, and negatively correlated with the spread. There is greater trading volume near the start of the trading day, and near the end of the trading day. In the middle of the day volume is lower. Despite the greater trading volume, the spread is wider at the start of the trading day.

Two aspects of the financial correlations seemed noteworthy. First, there is a positive return during the first 15 minutes of the day followed by a negative return over the subsequent 45 minutes. Second, the spread is negatively correlated with small trades, but positively correlated with medium and large trades. It seems easy to imagine that market makers worry more about the adverse selection issue associated with larger trades. For helpful discussion of adverse selection effects in market microstructure, see Campbell, Lo and MacKinlay (1997) and Madhavan (2000).

The correlations between the stock market features and the message board features are large. In many cases these correlations are larger than heavily studied correlations between different stock market features. For instance, the heavily studied correlation between trading volume and volatility is 0.063 in our sample. The correlation between trading volume and the number of messages is 0.322, while the correlation between the number of messages and volatility is 0.132. In other words, the magnitudes of the relationships between stock market

[13] As already mentioned, we experimented with constructing bullishness indicators in which the missing values were replaced with lagged values. These versions of the indices do not have the time lengthening artifact shown in Tables (5) and (6). However the indices constructed in this manner were much noisier and did not perform as well empirically.

attributes and message board measures are not trivial.

Trading volume is positively correlated the number of messages posted. Small-sized trades are more closely correlated with posting activity than are large-sized trades. Bullishness is also much more strongly correlated with small trades than with large ones. These results are consistent with the idea that the posting activity reflects day traders and not institutional investors. However, the correlations are not zero with respect to the larger trades. So there must be other aspects at work as well.

There are a number of distinct approaches to measuring volatility. We have tried a number of methods, but for simplicity we report results from the use of a particularly simple method. We calculate the standard deviation of actual trading prices within each 15 minute time interval. We take the measured standard deviation as our definition of volatility. This approach is similar in spirit to Schwert (1990) and French, Schwert and Stambaugh (1987).[14]

Volatility is positively correlated with message posting activity and, perhaps surprisingly, with the measures of bullishness. This effect is quite robust. Even more surprising is the fact that volatility is positively correlated with agreement among the messages. In light of the fact that some financial theories imply that disagreement induces trading, we had expected to find the reverse correlation. The interpretation of this finding is discussed in footnote (17).

Stock returns are almost uncorrelated with the number of messages, the bullishness of the postings, and agreement amongst the posted messages. This reflects the well known difficulty predicting stock returns. Some minor negative correlations are observed. This is likely special to the period studied. During the year 2000 the average returns were quite negative as shown in Table (3). Thus we are not too surprised to pick up an occasional negative correlation. A natural conjecture is that had the stock market been booming, then an occasional positive correlation would have been found.

Many significant correlations between the stock market measures and the message board activity measures are found. The similarity between the patterns found on Yahoo! Finance and Raging Bull is marked and encouraging. Table (7) and (8) show that there is information

[14]In some results not reported here we have experimented with the estimation of EGARCH and threshold GARCH models, see Glosten, Jagganathan and Runkle (1993) and Bollerslev, Engle and Nelson (1994). In each case we find that the residuals are forecastable using stock message board information. We also find that if we tried directly entering message board variables as explanatory factors and model the error as EGARCH, GJR, or related process, then the message board variables are statistically significant. However, we also find that these models take a remarkably long number of iterations to converge. Schwert (1990), Jones, Kaul, and Lipson (1994) and Chan and Fong (2000) fit time series models to returns and then use the absolute values of the residuals from these equations as the measure of volatility. We have tried this approach and the results are similar to those reported. To save space we omit these tables.

content on the stock message boards. The rest of the paper focuses on obtaining a clearer understanding of the form of the connection between the message boards and the stock markets.

4.7 Simple Contemporaneous Regressions

The correlations reported in Tables (7) and (8) cannot address the question of the independence of the relationships. To address the independence issue Tables (9) and (10) report results from a number of simple regressions. Each of the financial variables is treated as a dependent variable, and is regressed on a set of company dummies and three variables from the message boards: the log number of words, the bullishness index, and the agreement index. We also employ a stock market index in the regressions. The R^2 tends to be very high due to the use of fixed effects.

In the regressions that explain stock returns both theory and evidence show that it is the difference of log SPY (i.e. return) that belongs in the regression, and so we do that. In the other regressions it is less clear whether to include the log SPY or the difference of the log SPY. Prior literature does not provide clear guidance on this issue. Empirically log SPY performs markedly better in these regressions, and so we report these results. [15] The results concerning the effects of the message boards do not depend on which approach is taken.

Because our contemporaneous regressions employ logarithmic transformations, we can conveniently interpret several of our estimates as elasticities. For example, in our very first regression in Table (9) an increase in posting activity by 10% is associated with an increase in trading volume by less than 3%. At the same time, a 10% increase in the market index is associated with a 24% decrease in trading volume. Scanning through the column with estimates relating to posting activity, it is apparent that the magnitudes of all of these estimates are empirically relevant, with the exception of the estimates relating to returns. Particularly noteworthy is that posting activity is associated with a much larger response on small trades than on large trades. The elasticity of the number of small trades with respect to message posting activity is almost three times as large as the corresponding elasticity for large trades.

Calculation of elasticities with respect to a change in the bullishness index is complicated by the fact that the agreement index is a non-linear transformation of the bullishness index. Using the Delta method (Green 1997, p. 124), the trading volume elasticity is given

[15] For our purposes this is a side issue. However it is actually an issue that might merit further study in its own right. It is not obvious to us why the market index performs better in log form rather than in the difference of log form.

by $\beta_\theta \theta + \beta_\eta \theta^2/\sqrt{1-\theta^2}$, where β_θ and β_η are the estimates corresponding to the bullishness index θ and the agreement index η, respectively. Thus, in our first regression an increase in the bullishness index by 20% from $\theta = 0.5$ to $\theta = 0.6$ is associated with an increase in trading volume by 0.24%, with the effect from β_η dampening the effect from β_θ. While the effect of β_θ remains dominant for low values of θ, the negative effect from the agreement index becomes dominant when θ exceeds about 0.83. Clearly, the economic importance of the magnitudes of the contemporaneous effect of the bullishness and agreement indices is relatively small compared to either the posting activity or market index measures.[16]

The results in Tables (9) and (10) show that when the market as a whole increases in value, the number of trades drop. This effect is strongest among smaller trades. Volatility also drops in response to an increase in the market index. The effect on the spread is much small in magnitude and in statistical significance.

The results on both Yahoo! Finance and on Raging Bull confirm that log messages has a strong role in accounting for log trading volume. This is stronger for small size trades than for large trades. It seems natural to think of this as loose confirmation of the hypothesis that the stock message boards have more to do with day traders than with institutional investors.

The bullishness indices play a significant role explaining the number of trades even after we control for the other factors. This is true for all six versions of the bullishness index, making this result robust with respect to the weighting schemes and classification methods we use. It is also true both for Yahoo! Finance and Raging Bull. In contrast to the number of trades, the effect of bullishness was not so strongly related to trade size.

On Yahoo! Finance the Naive Bayes bullishness indices help to account for price volatility. However, this finding disappears when we use the Support Vector versions of the bullishness indices. On Raging Bull none of the versions of the bullishness indices were statistically significant.

Recall that the simple correlations between agreement and number of trades is positive. However, in Tables (9) and (10) we find that once we control for a number of other factors, the sign on the agreement indices reverse. Since the signs reverse across statistical methodologies,

[16]The bullishness and agreement indices are contemporaneous with the financial variables in Tables (9) and (10). An alternative approach is to use the 1-hour, 4-hour, and 1-day lagged aggregates of these indices. When this approach is taken the magnitude and significance of the coefficients on bullishness and agreement increase. Perhaps more interesingly, the observation that the coefficient on bullishness declines with increasing trade size becomes more pronounced. This would reinforce the hypothesis that day traders play a particularly significant role on the message boards.

we are cautious about interpreting the meaning of the signs on these coefficients.[17]

5 Time Sequencing Tests

A great many messages assert that a particular stock is a good buy, or that it is a bad buy. The time horizon of such forecasts is rarely specified. How accurate are such claims? Since contemporaneous regressions cannot address this issue, we study short horizon assessments. To extract the time sequencing information we take a simple approach based on a version of the Granger causality test. We study the time sequencing effects relating market features to message board features on a pairwise basis. The market features are returns, trading volume, volatility and spreads. The message board features are the number of messages, number of words, bullishness, and agreement.

There is an important question of what other effects need to be controlled for. Table (3) demonstrates that there are significant differences in the cross sectional levels of message posting activity. Thus we include firm fixed effects. Table (7) demonstrates that there are time-of-day effects. Thus we also use time period dummy variables. As in the simple regressions, we include a proxy for the market factor.

It is well known that the first trading day after a weekend tends to have negative returns. There does not seem to be a consensus interpretation of this fact. When studying daily data we include a dummy variable for the first trading day of the week.

The role of news stories is potentially very important. Do the message boards simply reflect what was published in the Wall Street Journal? Or, do they have more of an effect? In the daily frequency tests we used count variables to control for this issue.

There are a number of closely related approaches to testing for time sequencing, see Hamilton (1994). We take a particularly simple approach. Let x_t be some financial measure at time

[17] In a standard model I do not observe your beliefs directly. If you and I disagree about the value of an asset, then I learn about our disagreement by observing your willingness to take the other side of my trade. On the stock message boards it is quite different. Disagreement is observable. Suppose that I am thinking about making a trade in some stock. I check the message boards for recent postings. If there is general agreement with my point of view I do not learn much and so I may carry out my trade. Suppose instead that I observe a fair bit of disagreement. Then I might be tempted to try to sort out who I think is right, and who I think is wrong. While I am doing this I delay my trade. I may even decide not to carry out the trade altogether as a result. If this interpretation is correct then it is not so much that the previous theories are wrong. Rather the empirical context differs from the context considered in the theories about disagreement and trading. There are also other possible interpretations that we have not ruled out. It could be that the agreement index is proxying for some other factor that we have not measured. It could also be that our measure of agreement is imperfect in some critical manner. We have looked at some alternative measures, but clearly more could be done. This is a topic for further research.

t, let y_t be some message board measure at time t. Let D_i represent a time of day dummy for time period i. Then estimate the following equations,

$$x_t = \sum_{i=1}^{26} \alpha_i^1 D_i + \sum_{i=1}^{p} \beta_i^1 x_{t-i} + \sum_{i=1}^{p} \gamma_i^1 y_{t-i} + u_t^1 \quad (9)$$

$$x_t = \sum_{i=1}^{26} \alpha_i^0 D_i + \sum_{i=1}^{p} \beta_i^0 x_{t-i} + u_t^0 \quad (10)$$

To test for significance one can either perform an F test or a χ^2 test. Let,

$$RSS_1 = \sum_{t=1}^{T} (u_t^1)^2, \quad \text{and} \quad RSS_0 = \sum_{t=1}^{T} (u_t^0)^2. \quad (11)$$

Then the test statistic

$$S = \frac{T(RSS_0 - RSS_1)}{RSS_1} \quad (12)$$

follows a χ^2 distribution with p degrees of freedom. We use this simple approach to testing. We also reverse the position of the x_t and the y_t terms to test for the reverse time sequencing.

These are essentially linear tests of independence. Of course being linear tests, they may not pick up nonlinear relationships. Adding nonlinear terms would increase the flexibility, but at the same time increase the risk of overfitting the data. We did not experiment with any nonlinear specifications. Omitted variables are always an issue with causality tests. If an omitted variable is truly significant, and it is correlated with an included variable, then the included variable will end up appearing to be "Granger-causal."

Our main focus is on panel regressions in which we pool the 45 companies in our sample and introduce both company and time period fixed effects into equations (9) and (10).[18] We also tried performing these tests on a firm by firm basis, applying (9) and (10) directly.

5.1 Results: 15 minute data

Results for Yahoo! Finance and Raging Bull are reported separately. Tables (11) and (12) provide results using a 15 minute time period definition. Many significant predictive effects are found in both directions.

Start with trading volume. Stock trading volume has a more significant predictive effect

[18] Equations (11) and (12) are suitably modified by replacing T with the total number of observations in the panel.

on message posting than is true in the reverse direction, but in both cases the magnitude of the effect is large. The results from Yahoo! Finance and from Raging Bull are similar.

The largest impact is from the small size trades. This is consistent with small traders first making a trade, and then almost immediately posting messages to tell others what they have just done. Notice the effect is both strong and very short-lived. Half an hour later the sign reverses although the magnitude is much smaller. If this was reflecting a "pump and dump" market manipulation strategy, then it would be surprising that the effect reverses sign so rapidly. Serious pump-and-dumpers ought to work at it for longer than half an hour. For example, the Michael Lebed example cited in footnote (1) describes about 3 hours of active message posting.

The number of messages posted does help to predict volatility. It is noteworthy that while the trading volume χ^2 are larger from the market to the messages, the reverse is true for volatility. Message posting has a more significant impact on subsequent volatility than volatility has on subsequent message posting. This is particularly true for Yahoo! Finance. The results for the number of words posted are very similar to the results for the number of messages posted.

The χ^2 tests suggest that there is some predictability for stock returns. But the significance is dramatically lower than for the volume and the volatility results. The magnitudes of the message board coefficients are also very close to zero, suggesting that these would be hard to use these results to earn excess returns. Similarly, the results for the prediction of the spreads are quite weak.

5.2 Results: daily data

Tables (13) and (14) report results using a 1 day time period definition. At daily frequency there is more time for contemplation by traders. Since there are also many fewer time periods, it is not surprising that the χ^2 are much lower.

The trading volume results are quite interesting. Rather remarkably the χ^2 are now routinely higher when we use the message count to explain trading volume rather than the reverse. This is true on both message boards and it is in stark contrast to the 15 minute evidence. The connection between message posting and trading activity remains very strong.

At daily frequency we reject both the hypothesis of returns predictability and of spread predictability in most tests. The volatility results are qualitatively similar to the 15 minute results. Once again messages are more significant for market volatility rather than the reverse.

In the daily data we are able to include two more types of control variables: a dummy for the first day of a new trading week (NWK), and variables for the days surrounding an article in the Wall Street Journal (WSJ_t). The WSJ_t variables are count variables that simply tabulate the number of stories about the firm in question that were published in the Wall Street Journal on the date in question.

On the first day of the new week there is a dramatic drop in the number of messages posted, and in the length of the messages. Many people are presumably more busy doing other things. One interpretation that has been occasionally offered for the Monday effect is psychological: people are more depressed when they return to work after the weekend. Consistent with this, the sign on bullishness is negative at the start of the new week. But this is only a very minor effect and it is statistically insignificant, in sharp contrast to the activity effects.

Lakonishok and Maberly (1990) found a reduction in trades by institutional investors and an increase in trades by small investors at the start of the week. If we interpret large trades as being institutional trades, then we also find a big drop in such trades at the start of the week. However, we also find more minor drops in the medium and small trades at the start of the new week – not increases. Apparently, small investor behavior has changed over time in this respect.

An important question is the extent to which the stock message boards merely repeat what was already known from press reports. In order to get at this issue we have count variables that tabulate the number of news stories about a firm on a given day in the Wall Street Journal. The variable for the day that the story appears is denoted WSJ_0. We also include variables for the two prior dates (WSJ_{-2},WSJ_{-1}) and for one subsequent date (WSJ_{+1}).

The most important point is that the significance of the message board activity for predicting financial market behavior survive the inclusion of these variables. This means that there is information on the message boards beyond that contained in the Wall Street Journal. Coinciding with the publication day (day 0) there is an increase in the word count, but less of an increase in the number of messages. On average the messages tend to be longer on these days. There is a stronger day zero effect on Raging Bull than on Yahoo! Finance. The day –1 effect is larger than the day 0 effect. On day -1 more messages are posted and they are a bit more bullish. On Raging Bull the messages get longer on day –1. The greater importance of day –1 is also found in the financial markets. On day –1 more trades take place, particularly large trades. The fact that so much of the effect is found on day –1 suggests that wire services and the internet may be very significant sources of information for traders.

Day –2 shows some elevated activity on the stock markets, but almost none on the message boards. If people are trading on inside information, they are not being so kind as to advertise the fact on the message boards. Very little of note takes place on date +1. The message boards exhibit no unusual activity. The stock markets register a slight decline in trading volume, which is mainly found among the largest trades.

Table (15) asks a somewhat different question about the role of the Wall Street Journal. This table reports our tests for whether the stock message board activity can be used to predict the occurrence of news stories in the Wall Street Journal. For this purpose we use a logit regression where the binary variable is based on whether or not a WSJ article appears on the following day (or the day thereafter). The regressors in this model are based on 1-day time-period definitions.

We test for predictability one day before the story appears (top two panels in Table 15), and also two days before the story appears (bottom two panels in Table 15). The results are quite clear. One day before a story appears there is substantial evidence of elevated posting activity and increased bullishness. Two days before a story appears neither of these are found. Nothing unusual is taking place in the stock message boards. The inference is that the message boards appear to reflect news stories very rapidly. But they do not appear helpful in predicting news stories. The results in Table (15) are very consistent with the results in Table (13) and (14). The message boards reflect public information very rapidly.

The business press itself has been making use of the fact that the message boards react to news so rapidly. It is not uncommon for stories about corporate events to include references to discussion on the message boards. Colorful quotes from posted messages sometimes find their way into news stories. On other occasions the message boards are used to identify when an idea about a firm became public information.[19]

We were concerned about the extent to which our results might have been affected by

[19]To find current examples, go to the Wall Street Journal's web page and then search for "message board". Typical recent examples include: "Speculation on Flooz's fate first popped up in online message boards, including dot-com rumor site F___dcompany.com and a shopping board run by Anandtech.com." ("Flooz.com Says It Is Seeking Merger, But Service Remain Offline for Now," by Stephanie Miles, *Wall Street Journal*, August 10, 2001.) and "Since mid-July, when rumors started to spread on Internet message boards that Fidelity was selling most of its 7% stake, the stock had dropped more than 25%." ("Skechers USA Up 9%; Buyers Step In After FMR Gets Out," by Victoria Marcinkowski, *Dow Jones News Wires*, August 14, 2001.)

sparseness of message postings. To address this concern we repeated the daily level tests for the sample of 10 firms with the greatest posting levels.[20] While some minor changes are observed, the basic patterns are very similar. To save space we do not report the results.

6 Conclusions.

Casual reading of the internet stock messages might easily lead one to believe that it is all just noise. Indeed that was our first reaction to reading such messages. As might have been expected, there is little evidence that the stock message boards successfully predict stock returns. However there is useful information present on the stock message boards. The magnitudes of some of the observed effects are quite large relative to other features of stock markets that have attracted attention in the finance literature.

- Message posting and trading volume are positively related. This is true in a correlation sense as well as in contemporaneous regressions that control for fixed firm effects and a market factor. We also study the strength of the time sequencing. At 15 minute frequency, the effect from trading volume to the number of messages posted is more significant than the reverse. At daily frequency the effect from message posting activity to trading volume is the more significant direction. In general, the number of messages posted are more closely connected to small trades than to large trades.

- Message posting and volatility are also positively related. Again, this effect is found in simple correlations as well as in the regression context. The time sequencing effects stand in contrast to the trading volume results. Message posting always has a more significant effect on market volatility than market volatility has on message posting.

- Bullishness indices are constructed using methods from computational linguistics. We study results using versions of both the Naive Bayes algorithm and the Support Vector Machine algorithm. The two approaches typically give similar results, but the Naive Bayes method seems to work somewhat better. The overall tone on the message boards is bullish. This bullishness is particularly impressive given the large decline in stock

[20] The top 10 firms are listed in order followed by ticker, number of messages posted (YF+RB) and number of words posted (YF + RB): 1. Microsoft 197530 12301312, 2. E*Trade 161649 8576062, 3. Amazon 120574 8154611, 4. Intel 94347 5356805, 5. Philip Morris 84198 6491652, 6. Yahoo 78008 3465359, 7. AT&T 76769 4394984, 8. Verticalnet 64293 3449676, 9. Priceline 61397 3052542,10. Boeing 56279 4593476.

market value of the firms during the period under study. The DJX firms declined by about 20% over the year, while the ECM firms declined by more than 70% over the year. The bullishness indices are positively related to trading volume. The more bullish the posted messages during a time period, the more trades take place in the stock market during that same time period. Bullishness also exhibits some predictive ability. However the more significant effect is from trading volume to bullishness rather than from bullishness to trading volume. This result is particularly true for the smaller-sized trades. While this can be interpreted in terms of market manipulation strategies, it is also consistent with small traders simply wishing to talk about purchases that they have just made.

- The reported effects survive even when we take into account the presence of news stories in the Wall Street Journal. The message boards provide a source of financially relevant information beyond that already present in the Wall Street Journal. Indeed, not infrequently the Wall Street Journal itself reports what has been posted on message boards. Message postings are particularly elevated one day before a story appears in the Wall Street Journal. This presumably reflects the use of news wires and internet news sources. There is no evidence of anomalous posting behavior two days prior to a Wall Street Journal story. In contrast, there is some evidence of anomalous trading activity two days prior to a story. There is no evidence that the message boards are able predict the news stories earlier than the day before.

We also find some interesting evidence related to the well known Monday effect. Consistent with Lakonishok and Maberly (1990) we find a reduction in the number of large trades at the start of the week. In contrast to their findings we also find some drop in the number of small trades at the start of the week. We observe a large reduction in message posting activity on the first trading day of the week. Consistent with the "Monday blues" hypothesis we observe a negative effect on bullishness on the first trading day of the week. But, in contrast to the activity level effects, it is a small effect and not statistically significant.

Our findings have potentially important implications for studies in which it is important to know what information is public at a particular moment in time. This is of interest in studies of insider trading, and in many event studies. The messages reflect news very rapidly, suggesting that the message boards can be used to determine when particular information in fact became public. Our findings are also potentially interesting for those who study stock market volatility, as well as for those who study trading volume. The message boards provide information that seems to be helpful for forecasting purposes. Given that the messages contain explanations, it might be possible to identify different classes of events that have different stock market effects.

Our use of the Naive Bayes and Support Vector Machine may also prove helpful in other ways. A great deal of text based information is becoming available to scholars due to the role of the internet. There may be research benefits from extracting the information from such sources. In such studies the Naive Bayes and the Support Vector Machine algorithms provide natural starting points. As an example of the type of study we have in mind, think of using news wire reports. It would be interesting to gather a large number of news stories and then try to determine which classes of events have particularly large effects for stock returns and for volatility.

The evidence clearly rejects the hypothesis that the stock message boards are merely noise. There is financially relevant information present. The information is particularly pertinent for trading volume and for volatility. In contrast to the Business Week story that motivated our study, we did not find any simple predictability of stock returns using this information.

References

[1] Allen, Franklin, and Douglas Gale, 1999, "Diversity of Opinion and Financing of New Technology," *Journal of Financial Intermediation*, 8, 68-89.

[2] Bacidore, Jeffrey Michael, and Marc L. Lipson, 2001, "The Effects of Opening and Closing Procedures on the NYSE and Nasdaq," Working paper.

[3] Bessembinder, Hendrik, Kalok Chan, and Paul J. Seguin, 1996, "An Empirical Examination of Information, Differences of Opinion, and Trading Activity," *Journal of Financial Economics*, 40, 105-134.

[4] Bollerslev, Tim, Robert F. Engle, and Daniel B. Nelson, 1994, "ARCH Models," chapter 49 in R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics, Volume IV*, Elsevier Science.

[5] Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay, 1997, *The Econometrics of Financial Markets*, Princeton University Press, Princeton N. J.

[6] Chan, Kalok and Wai-Ming Fong, 2000, "Trade Size, Order Imbalance, and the Volatility-Volume Relation," *Journal of Financial Economics*, 57, 247-273.

[7] Cover, T. M., and J. A. Thomas, 1991, *Elements of Information Theory*, Wiley, New York.

[8] Daniel, Kent D., David Hirshleifer and Avanidhar Subrahmanyam, 2001, "Overconfidence, Arbitrage, and Equilibrium Asset Pricing," *Journal of Finance*, 56, 921-965.

[9] Dewally, Michael, 2000, "Internet Investment Advice: Investing with a Rock of Salt," Working Paper, University of Oklahoma.

[10] Dumais, Susan T., J. Platt, D. Heckerman and M. Sahami, 1998, "Inductive Learning Algorithms and Representations for Text Categorization," in *Proceedings of ACM-CIKM98*, November, 148-155, ACM Press.

[11] French, K. R., G. W. Schwert, and R. F. Stambaugh, 1987, "Expected Stock Returns and Volatility," *Journal of Financial Economics*, 19, 3-29.

[12] Glosten, L. R., Ravi Jagannathan, and David Runkle, 1993, "On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *Journal of Finance*, 48, 1779-1801.

[13] Greene, William H., 1997, *Econometric Analysis*, 3rd edition, Prentice Hall, Upper Saddle River.

[14] Hamilton, James D., 1994, *Time Series Analysis*, Princeton University Press, Princeton N.J.

[15] Harris, Milton, and Artur Raviv, 1993, "Differences of Opinion Make a Horse Race," *Review of Financial Studies*, 6, 473-506.

[16] Hasbrouck, Joel, 1999, "The Dynamics of Discrete Bid and Ask Quotes," *Journal of Finance*, 54, 2109-2142.

[17] Hasbrouck, Joel, and Duane J. Seppi, 2001, "Common Factors in Pricing, Order Flows and Liquidity," *Journal of Financial Economics*, 59, 383-411.

[18] Joachims, Thorsten, 1998, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proceedings of the 10th European Conference on Machine Learning (ECML)*, 137-142, Springer-Verlag, New York.

[19] Joachims, Thorsten, 1999, "Making large-Scale SVM Learning Practical," in B. Schölkopf and C. Burges and A. Smola (eds.) *Advances in Kernel Methods - Support Vector Learning*, MIT Press, Cambridge, MA.

[20] Jones, Charles M., Gautam Kaul, and Marc L. Lipson, 1994, "Transactions, Volume, and Volatility," *Review of Financial Studies*, 7, 4, 631-651.

[21] Kandel, Eugene and N. D. Pearson, 1995, "Differential Interpretation of Public Signals and Trade in Speculative Markets," *Journal of Political Economy*, 103, 831-872.

[22] Lakonishok, Josef, and Edwin Maberly, 1990, "The Weekend Effect: Trading Patterns of Individual and Institutional Investors," *Journal of Finance*, 40, 231-243.

[23] Lewis, David D., 1998, "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, in *Proceedings of the 10th European Conference on Machine Learning (ECML)*, Springer-Verlag, New York.

[24] Madhavan, Anath, 2000, "Market Microstructure: A Survey," *Journal Of Financial Markets*, 3, 205-258.

[25] Manning, Christopher D., and Hinrich Schutze, 1999, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge Mass.

[26] McCallum, Andrew, 1996, "Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering," School of Computer Science, Carnegie-Mellon University.

[27] McCallum, Andrew and Kamal Nigam, 1998, "A Comparison of Event Models for Naive Bayes Text Classification," in *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41-48. Technical Report WS-98-05. AAAI Press.

[28] McCallum, Andrew, Kamal Nigam, Jason Rennie, and Kristie Seymore, 2000, "Automating the Construction of Internet Portals with Machine Learning," *Information Retrieval Journal*, 3, 127-163.

[29] Milgrom, Paul, and Nancy Stokey, 1982, "Information, Trade and Common Knowledge," *Journal of Economic Theory*, 26, 17-27.

[30] Schwert, G. W., 1990, "Stock Volatility and the Crash of '87," *Review of Financial Studies*, 3, 77-102.

[31] Tumarkin, Robert, and Robert F. Whitelaw, 2001, "News or Noise? Internet Message Board Activity and Stock Prices," *Financial Analysts Journal*, 57, 41-51.

[32] Weiss, Gary, 2000, "Where Everyone Is an Expert," *BusinessWeek*, May 22.

[33] Wysocki, Peter D., 1999, "Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards," Working Paper, November, University of Michigan.

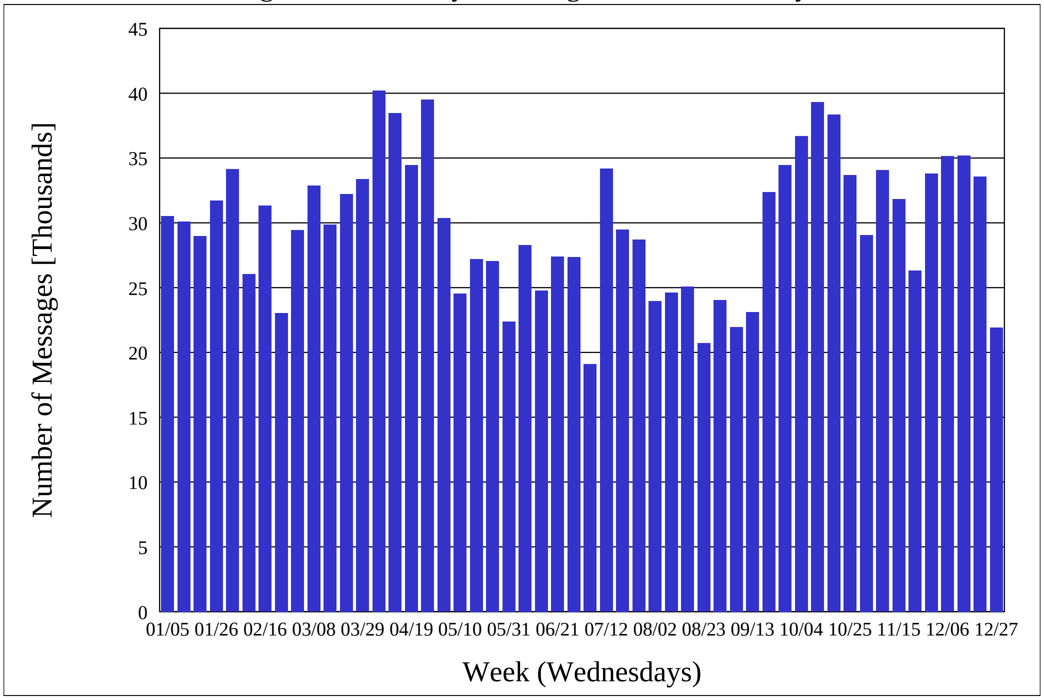
[34] Yang, Yiming, 1999, "An Evaluation of Statistical Approaches to Text Categorization," *Journal of Information Retrieval*, Vol. 1, No. 1/2, 67–88.

Figures

Figure 1: Samples of Bulletin Board Messages

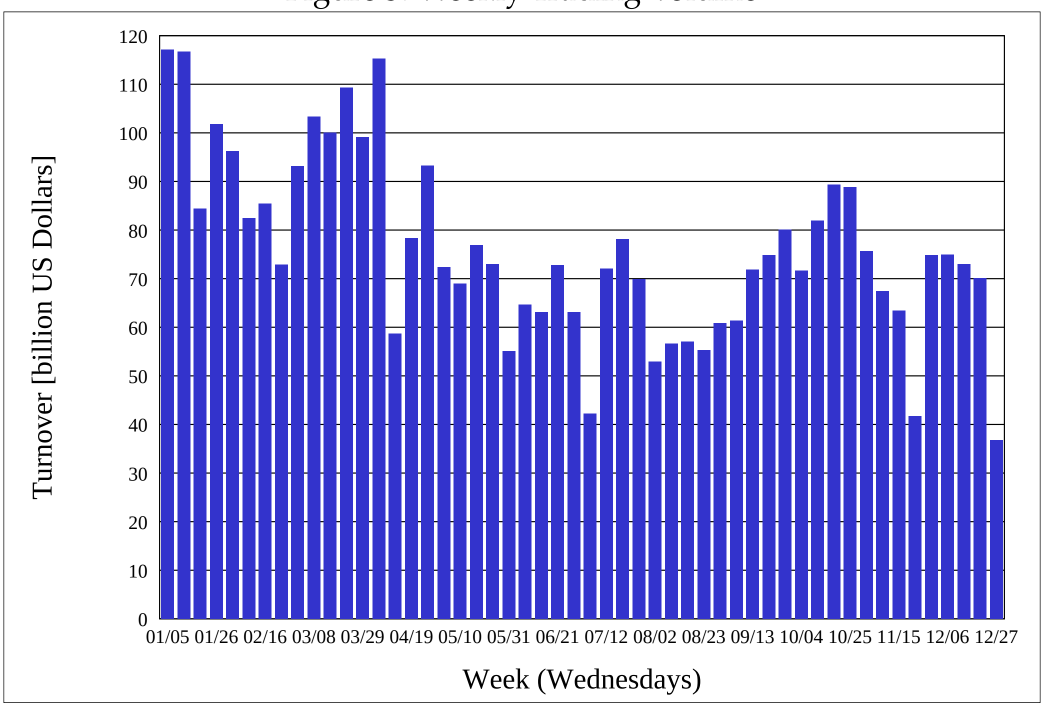
```
---------------------
FROM YF
COMP ETYS
MGID 13639
NAME CaptainLihai
LINK 1
DATE 2000/01/25 04:11
SKIP
TITL ETYS will surprise all pt II
SKIP
TEXT ETYS will surprise all when it drops to below 15$ a pop, and even then
TEXT it will be too expensive.
TEXT
TEXT If the DOJ report is real, there will definately be a backlash against
TEXT the stock. Watch your asses. Get out while you can.
---------------------
FROM YF
COMP IBM
MGID 43653
NAME plainfielder
LINK 1
DATE 2000/03/29 11:39
SKIP
TITL BUY ON DIPS - This is the opportunity
SKIP
TEXT to make $$$ when IBM will be going up again following this profit taking
TEXT bout by Abbey Cohen and her brokerage firm.
TEXT
TEXT IBM shall go up again after today.
-----------------
```

Figure 2: Weekly Message Board Activity



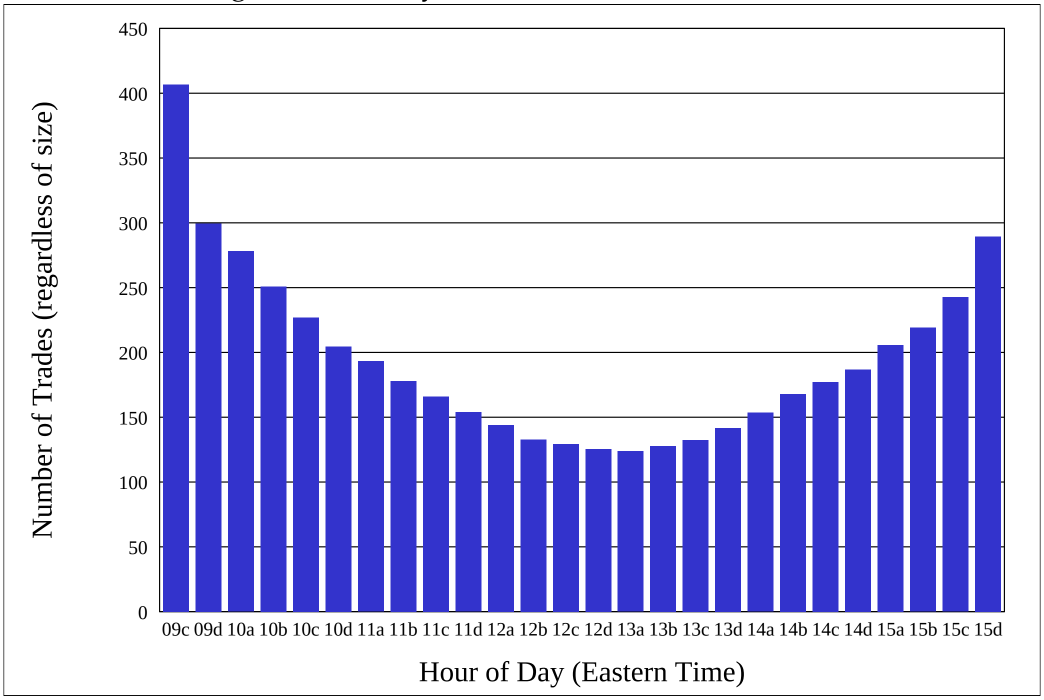
Note: Posting activity of 45 companies in ECM and DJX combined

Figure 3: Weekly Trading Volume



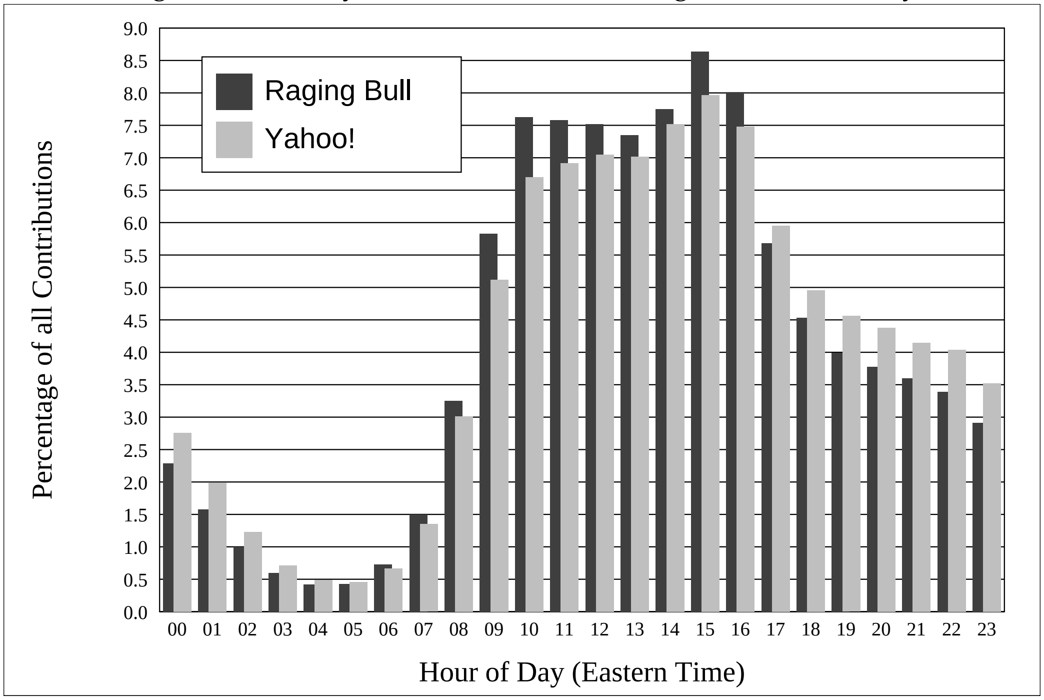
Note: Turnover of 45 companies in ECM and DJX combined

Figure 4: Hourly Distribution of Stock Trades



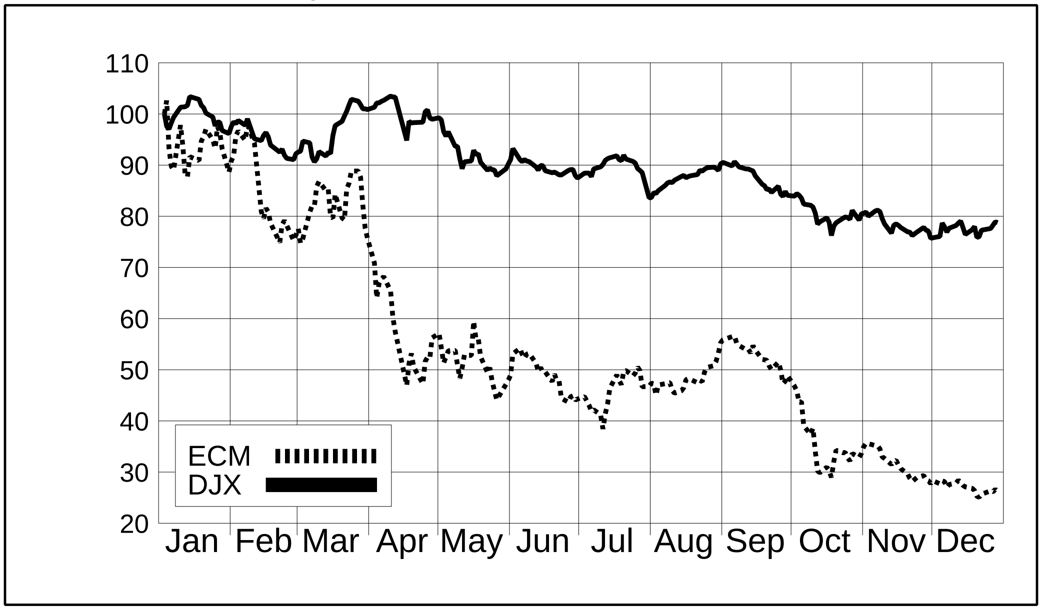
Note: Number of trades of all 45 companies in ECM and DJX combined. The letters a–d on the right side of an hour indicate fifteen minute intervals within each hour.

Figure 5: Hourly Distribution of Message Board Activity



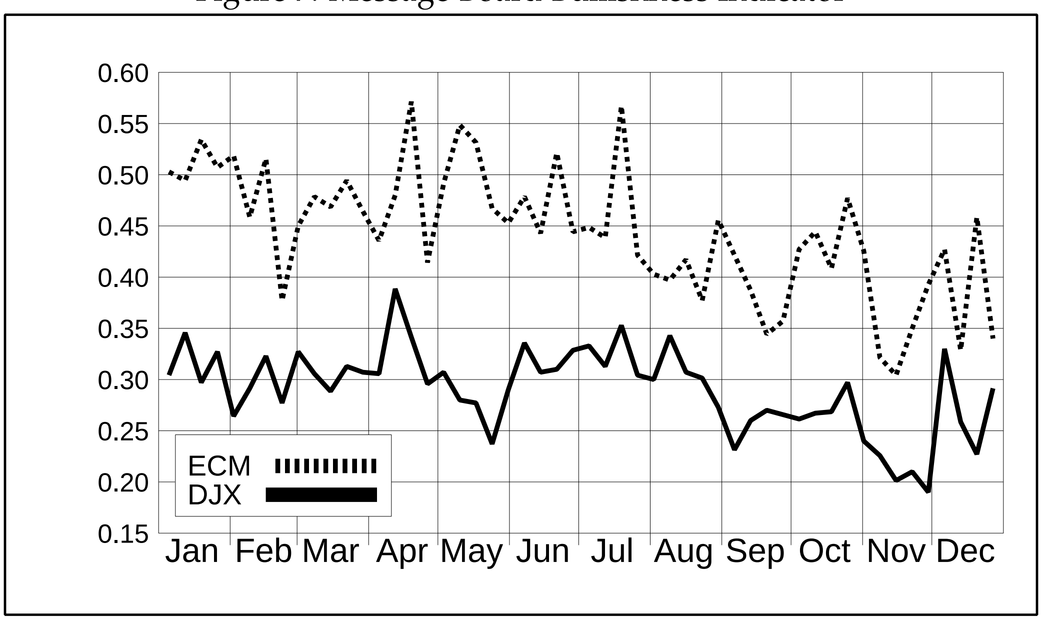
Note: Percentage of postings of 45 companies in ECM and DJX combined

Figure 6: Stock Index Performance



Note: Daily indices for DJX and ECM were computed using fixed weights. The indices were set to 100.0 for the first trading day of 2000. Index weights are based on percentage shares of market capitalization at the beginning of the year. The ECM index was adjusted when stocks ceased trading: Go2net on October 13; and Lycos on October 30.

Figure 7: Message Board Bullishness Indicator



Note: The bullishness signal is based on the Yahoo!-Finance message board only. Due to the high intra-week volatility of the bullishness signal, the above chart depicts weekly averages for the DJX and ECM indices.

Tables

Table 1: Frequency Distribution of Bulletin Board Contributors and Contributions

Postings per Contributor	Word Count Percentage		Percentage of Contributors		Words in Message	% of Postings	
	RB	YF	RB	YF		RB	YF
1	3.81	3.10	43.59	40.58	1–9	15.55	21.31
2	3.00	2.45	15.77	15.49	10–19	16.03	15.64
3–4	4.40	3.88	13.55	14.18	20–49	28.67	28.14
5–9	6.90	6.63	11.88	12.70	50–99	19.81	18.88
10–19	8.47	8.23	7.04	7.85	100–199	11.65	10.36
20–49	13.40	12.76	4.80	5.46	200–499	5.74	4.84
50+	60.01	62.96	3.38	3.74	500–999	1.84	0.83
					1000+	0.71	0.00

Note: The column 'Word Count Percentage' shows the fraction of the total words posted that were contributed by authors with the number of postings shown in the first column. The column 'Percentage of Contributors' shows the fraction of authors who contributed the number of messages indicated in the first column. RB and YF indicate the Raging Bull and Yahoo! Finance message boards. The second part of the table ('Words in Message') shows the length distribution of messages on the two message boards.

Table 2: Classification Accuracy Within Sample and Overall Classification Distribution

	Classification Results					
	NB			SV		
	B	H	S	B	H	S
Correct B	18.1	7.1	0.0	25.3	2.0	0.0
Correct H	3.4	65.9	0.0	0.2	66.6	0.0
Correct S	0.2	1.2	4.1	0.0	0.3	5.6
All Messages	20.0	78.8	1.3	20.3	77.0	2.7

Note: The first three rows lists the within sample classification for the 1,000 hand coded messages. The final row "all messages" provides summary statistics for the out-of-sample classification, showing the percentage of all 1,559,621 messages that were classified as buy (B), hold (H) or sell (S) by the Naive Bayes (NB) and Support Vector Machine (SV) algorithms.

Table 3: Summary Statistics By Company

Company Name	Bullishness[1] YF[2]	Bullishness[1] RB[2]	Activity[3] YF[2]	Activity[3] RB[2]	Intensity[4] YF[2]	Intensity[4] RB[2]	Return[5] [%]	Vola.[6] [%]	WSJ [#]
Philip Morris	0.570	0.421	78.4	5.8	74	115	86.5	0.24	45
Intel	0.541	0.346	80.2	14.1	52	85	-64.2	0.33	96
AT&T	0.525	0.340	64.9	11.9	53	78	-66.5	0.23	189
General Electric	0.491	0.355	40.1	15.8	72	90	-68.7	0.18	96
Microsoft	0.507	0.282	159.3	38.3	56	90	-63.0	0.25	397
Citigroup	0.337	0.545	4.4	2.5	60	97	-7.6	0.22	80
Hewlett Packard	0.378	0.296	16.0	0.8	63	125	-72.4	0.28	36
Johnson&Johnson	0.366	0.395	2.9	0.4	70	72	12.7	0.15	31
Wal Mart	0.357	0.443	20.5	3.2	82	79	-22.5	0.23	55
Walt Disney	0.354	0.412	18.3	2.0	71	96	-0.9	0.20	83
Home Depot	0.326	0.493	17.6	3.0	54	81	-33.5	0.22	26
United Technologies	0.354	0.395	2.0	0.1	79	70	20.9	0.19	18
Honeywell	0.352	0.365	12.1	0.7	76	81	-18.3	0.22	27
SBC Communications	0.353	0.277	16.6	1.5	65	88	-1.9	0.19	44
IBM	0.339	0.334	24.5	2.6	66	98	-24.6	0.22	108
Intn'l Paper	0.330	0.373	8.9	0.3	75	69	-28.4	0.24	19
Procter&Gamble	0.354	0.226	19.1	2.3	55	95	-27.2	0.23	54
JP Morgan	0.308	0.348	1.4	0.1	59	127	30.9	0.22	58
Eastman Kodak	0.307	0.278	2.7	0.2	72	129	-41.1	0.20	18
McDonalds	0.307	0.271	4.3	0.6	67	63	-15.0	0.20	32
Boeing	0.292	0.131	54.8	1.5	81	90	58.9	0.19	123
Minnesota Mining	0.261	0.414	1.5	0.2	66	101	25.6	0.18	6
American Express	0.269	0.377	3.4	0.2	58	70	-66.6	0.21	44
Alcoa	0.257	0.255	5.0	0.2	56	38	-59.3	0.22	26
Coca Cola	0.244	0.239	9.8	0.5	98	94	6.1	0.18	106
Du Pont	0.238	0.223	6.7	0.3	72	74	-26.8	0.21	0
Merck	0.217	0.186	8.8	0.5	75	87	36.8	0.17	8
General Motors	0.185	0.217	6.6	0.6	85	96	-31.3	0.17	181
Caterpillar	0.172	0.285	1.4	0.2	65	60	-3.8	0.20	8
Exxon	0.156	0.223	7.6	0.6	78	69	8.6	0.14	6
E*Trade	0.881	0.788	140.6	21.0	47	91	-72.4	0.48	8
Verticalnet	0.820	0.742	57.3	7.0	52	66	-96.3	0.81	3
Ameritrade	0.747	0.604	45.5	7.9	50	72	-68.7	0.54	0
Healtheon	0.624	0.559	40.5	6.6	58	83	-79.0	0.63	1
Yahoo!	0.629	0.462	63.8	14.2	41	60	-93.2	0.51	28
Etoys	0.597	0.520	32.4	8.9	58	79	-99.3	1.93	14
Lycos	0.578	0.545	13.5	5.8	49	96	-49.4	0.51	19
Priceline	0.564	0.428	48.5	12.9	46	65	-97.5	0.72	21
Ticketmaster	0.474	0.636	4.7	0.7	80	115	-79.6	0.56	4
Go2net	0.440	0.502	4.8	0.9	49	76	-63.1	0.53	3
CNet	0.435	0.440	12.3	3.7	49	73	-74.8	0.54	10
Webvan Group	0.393	0.402	20.2	3.1	84	96	-97.2	1.59	8
Amazon	0.397	0.306	103.8	16.8	66	76	-81.1	0.55	37
E-Bay	0.275	0.211	28.3	3.9	49	62	-74.5	0.54	63
MP3.com	0.154	0.069	14.4	5.4	68	128	-89.8	1.17	22

Note: The table is sorted by company group (DJX, ECM) and within each group by descending word-weighted bullishness index of the two bulletin boards. [1] Bullishness refers to our preferred length-weighted Naive Bayes classification. [2] RB and YF indicate Raging Bull and Yahoo Finance! [3] Activity is measured in thousands of messages. [4] Intensity is measured as the average number of words per message. [5] Return is the change in price between the first and the last trading day of the year. Lycos and Go2net stopped trading in late October. [6] Volatility is measures as the average daily standard deviation divided by price, expressed in percent.

Table 4: Market Data Summary Statistics

Variable	Mean	Std.Dev.	Mininum	Maximum
Trading Volume	217338.1	413460.3	0	21553200
Number of Small Trades	169.1461	405.1035	0	9833
Number of Medium Trades	23.8228	59.70418	0	2324
Number of Large Trades	1.56808	3.655313	0	184
Price Volatility	0.128186	0.211896	0	35.2113
Price	58.25154	40.55987	0.1709	496.0325
Bid-Ask Spread	0.18648	0.379847	0.0625	50.0625
Return (Log Diff. of Price)	-0.00014	0.010138	-1.31081	0.416248

Table 5: Yahoo! Finance Message Board Summary Statistics

Variable	Weight	No Lag	1-Hour Lag	4-Hour Lag	1-Day Lag
Messages		0.841 (2.508)	3.365 (9.176)	13.46 (33.33)	80.89 (152.9)
Words		49.84 (151.3)	199.4 (473.6)	797.6 (1627)	4790 (7644)
Naive Bayes	none	0.098 (0.317)	0.23 (0.444)	0.435 (0.514)	0.733 (0.411)
Naive Bayes	words	0.098 (0.318)	0.229 (0.449)	0.433 (0.525)	0.728 (0.437)
Naive Bayes	scarcity	0.098 (0.319)	0.231 (0.453)	0.439 (0.534)	0.745 (0.457)
Support Vector	none	0.081 (0.301)	0.198 (0.433)	0.385 (0.511)	0.661 (0.422)
Support Vector	words	0.08 (0.303)	0.193 (0.438)	0.368 (0.524)	0.613 (0.463)
Support Vector	scarcity	0.081 (0.304)	0.199 (0.444)	0.389 (0.54)	0.675 (0.489)

Table 6: Raging Bull Message Board Summary Statistics

Variable	Weight	No Lag	1-Hour Lag	4-Hour Lag	1-Day Lag
Messages		0.145 (0.68)	0.58 (2.266)	2.322 (7.798)	13.96 (34.45)
Words		12.16 (88.32)	48.63 (223.2)	194.6 (650.8)	1169 (2796)
Naive Bayes	none	0.023 (0.155)	0.069 (0.265)	0.178 (0.396)	0.443 (0.505)
Naive Bayes	words	0.023 (0.155)	0.069 (0.266)	0.178 (0.399)	0.444 (0.514)
Naive Bayes	scarcity	0.023 (0.155)	0.07 (0.266)	0.179 (0.4)	0.446 (0.518)
Support Vector	none	0.019 (0.169)	0.058 (0.28)	0.144 (0.406)	0.344 (0.498)
Support Vector	words	0.019 (0.169)	0.057 (0.284)	0.139 (0.415)	0.326 (0.525)
Support Vector	scarcity	0.019 (0.17)	0.058 (0.286)	0.145 (0.424)	0.348 (0.55)

Note: The time interval is taken as 15 minutes. There are 45 stocks in our sample, and there are 35,136 fifteen minute intervals (FMI) over the 366 days included. Multiplying these numbers gives the 1,581,120 company/FMI observations in our data. Some definitions: Price is the midpoint of the bid and the ask at interval end. Volatility is the standard deviation of the transaction prices within a time period. Trading volume is the total number of shares that were traded during a time period. Return is the log difference of the average transaction price from one period to the next. Spread is the difference between the bid and the ask at the end of the period. Number of Small Trades is a count of the Number of Trades below $100,000. Number of Medium Trades is a count of the Number of Trades that are at least $100,000 but below $1,000,000. Number of Large Trades is a count of the Number of Trades valued at least $1,000,000. Messages is a count of the number of of messages aggregated over the indicated time interval. Words is a count of the number of words posted over the indicated time interval. NB and SV indicate the classification methods Naive Bayes and Support Vector. Weights 'none,' 'words,' and 'scarcity' imply equal weighting, weighting by the length of contribution, and weighting by the inverse contribution frequency (scarcity) of the contributor.

Table 7: Pairwise Correlations for TAQ and Yahoo! Finance

	Volume	Small	Medium	Large	Volatility	Return	Spread
Small Trades	0.750						
Medium Trades	0.741	0.544					
Large Trades	0.654	0.421	0.677				
Volatility	0.063	0.218	-0.183	-0.093			
Return	0.016	0.008	0.023	0.016	-0.019		
Spread	-0.031	-0.086	0.063	0.043	-0.035		
09:30 Dummy	0.140	0.121	0.060	0.104	0.119	0.011	0.037
09:45 Dummy	0.075	0.061	0.065	0.048	0.071	-0.010	0.011
10:00 Dummy	0.067	0.052	0.058	0.043	0.046	-0.008	0.007
10:15 Dummy	0.047	0.038	0.041	0.027	0.026	-0.006	
15:00 Dummy	0.007	0.012	0.009				
15:15 Dummy	0.023	0.023	0.021	0.009			
15:30 Dummy	0.047	0.041	0.040	0.020	0.016		
15:45 Dummy	0.096	0.076	0.079	0.058	0.031		
Messages, no lag	0.322	0.484	0.146	0.173	0.132	-0.008	-0.069
Messages, 1h lag	0.325	0.500	0.142	0.171	0.122	-0.012	-0.072
Messages, 4h lag	0.300	0.474	0.124	0.155	0.097	-0.007	-0.072
Messages, 1d lag	0.351	0.518	0.179	0.211	0.108	-0.008	-0.074
Words, no lag	0.252	0.346	0.116	0.146	0.076	-0.007	-0.054
Words, 1h lag	0.301	0.420	0.136	0.169	0.090	-0.010	-0.067
Words, 4h lag	0.303	0.431	0.134	0.169	0.077	-0.007	-0.071
Words, 1d lag	0.372	0.507	0.191	0.224	0.097	-0.008	-0.078
NB simple, no lag	0.174	0.297	0.030	0.066	0.109		-0.048
NB simple, 1h lag	0.200	0.323	0.048	0.094	0.112		-0.051
NB simple, 4h lag	0.178	0.259	0.056	0.094	0.076		-0.043
NB simple, 1d lag	0.137	0.155	0.055	0.076	0.043		-0.019
NB length, no lag	0.173	0.295	0.030	0.065	0.108		-0.047
NB length, 1h lag	0.196	0.317	0.048	0.093	0.109		-0.051
NB length, 4h lag	0.174	0.253	0.055	0.091	0.073		-0.043
NB length, 1d lag	0.128	0.150	0.048	0.067	0.044		-0.018
NB scarce, no lag	0.173	0.295	0.030	0.066	0.108		-0.047
NB scarce, 1h lag	0.197	0.318	0.048	0.093	0.109		-0.051
NB scarce, 4h lag	0.175	0.254	0.057	0.094	0.073		-0.043
NB scarce, 1d lag	0.139	0.152	0.066	0.084	0.036		-0.018
SV length, no lag	0.152	0.259	0.046	0.068	0.077		-0.035
SV length, 1h lag	0.168	0.277	0.052	0.084	0.084		-0.042
SV length, 4h lag	0.132	0.204	0.037	0.072	0.061		-0.033
SV length, 1d lag	0.079	0.092	0.019	0.047	0.028		-0.014
Agreement, no lag	0.190	0.326	0.034	0.073	0.118		-0.053
Agreement, 1h lag	0.196	0.325	0.042	0.091	0.116		-0.052
Agreement, 4h lag	0.136	0.200	0.034	0.071	0.066		-0.034
Agreement, 1d lag	-0.005	-0.025	-0.023				

Note: Only correlations that are significantly different from zero at the 99% confidence level are reported. Missing values are deleted on a case by case basis rather than for the entire data set. The time dummies for the first and last hour are reported. Midday time dummies tended to be insignificant and economically small. The time preiod length is 15 minutes.

Table 8: Pairwise Correlations for Raging Bull

	Volume	Small	Medium	Large	Volatility	Return	Spread
Messages, no lag	0.249	0.383	0.157	0.179	0.109		-0.050
Messages, 1h lag	0.275	0.433	0.170	0.193	0.108	-0.011	-0.060
Messages, 4h lag	0.272	0.436	0.162	0.189	0.092	-0.007	-0.066
Messages, 1d lag	0.311	0.482	0.203	0.233	0.104	-0.007	-0.069
Words, no lag	0.143	0.214	0.087	0.107	0.048		-0.032
Words, 1h lag	0.210	0.314	0.125	0.154	0.065	-0.005	-0.049
Words, 4h lag	0.246	0.373	0.144	0.181	0.066		-0.063
Words, 1d lag	0.308	0.465	0.194	0.238	0.088	-0.005	-0.074
NB simple, no lag	0.120	0.207	0.043	0.062	0.075		-0.029
NB simple, 1h lag	0.169	0.300	0.051	0.087	0.097	-0.006	-0.049
NB simple, 4h lag	0.194	0.353	0.053	0.103	0.110		-0.063
NB simple, 1d lag	0.224	0.390	0.064	0.126	0.145		-0.061
NB length, no lag	0.119	0.206	0.042	0.061	0.075		-0.029
NB length, 1h lag	0.168	0.298	0.050	0.086	0.098	-0.006	-0.048
NB length, 4h lag	0.192	0.349	0.049	0.099	0.111		-0.062
NB length, 1d lag	0.218	0.381	0.058	0.117	0.147		-0.060
NB scarce, no lag	0.119	0.206	0.043	0.062	0.075		-0.029
NB scarce, 1h lag	0.170	0.301	0.053	0.089	0.097	-0.006	-0.048
NB scarce, 4h lag	0.197	0.353	0.056	0.107	0.107		-0.062
NB scarce, 1d lag	0.225	0.384	0.070	0.131	0.139		-0.059
SV length, no lag	0.104	0.164	0.060	0.070	0.042		-0.021
SV length, 1h lag	0.136	0.224	0.066	0.085	0.061		-0.030
SV length, 4h lag	0.131	0.229	0.049	0.079	0.067		-0.031
SV length, 1d lag	0.138	0.218	0.049	0.079	0.067		-0.029
Agreement, no lag	0.131	0.227	0.048	0.068	0.081		-0.032
Agreement, 1h lag	0.178	0.322	0.054	0.091	0.106	-0.006	-0.053
Agreement, 4h lag	0.191	0.361	0.044	0.097	0.118		-0.067
Agreement, 1d lag	0.189	0.351	0.036	0.101	0.144		-0.058

Note: See table 7 for explanations.

Table 9: Contemporaneous Regressions – Yahoo! Finance Message Board

	Log of Messages		Bullishness Index		Agreement Index		Market Index[1]		R^2
Naive Bayes, Unweighted									
Log Trading Volume	0.276^c	(89.71)	0.040^c	(5.149)	-0.027^b	(3.218)	-2.386^c	(60.52)	0.995
Log Small Trades	0.243^c	(112.4)	0.047^c	(8.683)	-0.040^c	(6.818)	-1.510^c	(54.5)	0.984
Log Medium Trades	0.134^c	(49.95)	0.043^c	(6.406)	-0.023^b	(3.125)	-0.439^c	(12.78)	0.931
Log Large Trades	0.086^c	(40.15)	0.017^b	(3.064)	0.000	(0.077)	-0.210^c	(7.638)	0.642
Return	-0.000	(0.151)	0.000	(0.916)	0.000	(0.573)	0.797^c	(136.6)	0.063
Price Volatililty	0.044^c	(38.81)	0.014^c	(5.111)	-0.018^c	(6.08)	-1.169^c	(81.31)	0.537
Spread	0.002	(1.315)	0.002	(0.536)	0.000	(0.125)	-0.045^b	(2.619)	0.244
Naive Bayes, Weighted by Message Length									
Log Trading Volume	0.276^c	(89.15)	0.035^c	(4.646)	-0.020^a	(2.47)	-2.386^c	(60.53)	0.995
Log Small Trades	0.243^c	(111.7)	0.044^c	(8.231)	-0.035^c	(5.981)	-1.510^c	(54.52)	0.984
Log Medium Trades	0.134^c	(49.77)	0.042^c	(6.375)	-0.021^b	(2.973)	-0.439^c	(12.78)	0.931
Log Large Trades	0.086^c	(39.91)	0.014^b	(2.65)	0.003	(0.581)	-0.210^c	(7.636)	0.642
Return	-0.000	(0.001)	0.000	(0.998)	0.000	(0.351)	0.797^c	(136.6)	0.063
Price Volatililty	0.044^c	(38.7)	0.014^c	(5.036)	-0.018^c	(5.932)	-1.169^c	(81.31)	0.537
Spread	0.002	(1.32)	0.002	(0.734)	-0.000	(0.039)	-0.045^b	(2.622)	0.244
Naive Bayes, Weighted by Inverse Contribution Frequency									
Log Trading Volume	0.276^c	(88.78)	0.028^c	(3.831)	-0.014	(1.74)	-2.385^c	(60.51)	0.995
Log Small Trades	0.243^c	(111.3)	0.038^c	(7.195)	-0.030^c	(5.17)	-1.510^c	(54.49)	0.984
Log Medium Trades	0.134^c	(49.59)	0.036^c	(5.607)	-0.016^a	(2.305)	-0.438^c	(12.76)	0.931
Log Large Trades	0.086^c	(39.54)	0.012^a	(2.275)	0.007	(1.153)	-0.210^c	(7.636)	0.642
Return	-0.000	(0.062)	0.000	(1.481)	0.000	(0.029)	0.797^c	(136.6)	0.063
Price Volatililty	0.044^c	(38.57)	0.012^c	(4.433)	-0.016^c	(5.395)	-1.169^c	(81.31)	0.537
Spread	0.002	(1.34)	0.001	(0.382)	0.001	(0.226)	-0.045^b	(2.613)	0.244
Support Vector, Unweighted									
Log Trading Volume	0.286^c	(96.05)	0.037^c	(5.489)	-0.056^c	(7.612)	-2.379^c	(60.39)	0.995
Log Small Trades	0.250^c	(119.8)	0.033^c	(7.008)	-0.052^c	(10.05)	-1.504^c	(54.32)	0.984
Log Medium Trades	0.141^c	(54.37)	0.042^c	(7.144)	-0.042^c	(6.569)	-0.433^c	(12.6)	0.931
Log Large Trades	0.091^c	(43.91)	0.019^c	(4.177)	-0.017^b	(3.29)	-0.206^c	(7.491)	0.642
Return	0.000	(0.239)	0.000	(1.172)	0.000	(0.29)	0.797^c	(136.6)	0.063
Price Volatililty	0.045^c	(41.49)	-0.000	(0.164)	-0.011^c	(4.1)	-1.168^c	(81.26)	0.537
Spread	0.001	(0.549)	0.003	(1.032)	0.004	(1.235)	-0.045^b	(2.631)	0.244
Support Vector, Weighted by Message Length									
Log Trading Volume	0.286^c	(95.82)	0.031^c	(4.833)	-0.050^c	(6.937)	-2.379^c	(60.39)	0.995
Log Small Trades	0.250^c	(119.5)	0.028^c	(6.069)	-0.046^c	(9.05)	-1.505^c	(54.33)	0.984
Log Medium Trades	0.141^c	(54.32)	0.038^c	(6.676)	-0.038^c	(6.043)	-0.433^c	(12.61)	0.931
Log Large Trades	0.091^c	(43.82)	0.017^c	(3.703)	-0.014^b	(2.788)	-0.206^c	(7.493)	0.642
Return	0.000	(0.487)	0.000	(0.895)	0.000	(0.161)	0.797^c	(136.6)	0.063
Price Volatililty	0.045^c	(41.35)	-0.001	(0.499)	-0.010^c	(3.807)	-1.168^c	(81.26)	0.537
Spread	0.001	(0.572)	0.003	(1.019)	0.004	(1.241)	-0.045^b	(2.63)	0.244
Support Vector, Weighted by Inverse Contribution Frequency									
Log Trading Volume	0.286^c	(95.3)	0.031^c	(4.852)	-0.049^c	(6.838)	-2.380^c	(60.4)	0.995
Log Small Trades	0.250^c	(118.8)	0.026^c	(5.724)	-0.044^c	(8.563)	-1.505^c	(54.34)	0.984
Log Medium Trades	0.141^c	(54.04)	0.036^c	(6.463)	-0.037^c	(5.861)	-0.433^c	(12.61)	0.931
Log Large Trades	0.091^c	(43.57)	0.017^c	(3.861)	-0.014^b	(2.864)	-0.206^c	(7.496)	0.642
Return	0.000	(0.458)	0.000	(0.818)	0.000	(0.247)	0.797^c	(136.6)	0.063
Price Volatililty	0.045^c	(41.07)	0.000	(0.094)	-0.011^c	(4.079)	-1.168^c	(81.28)	0.537
Spread	0.001	(0.53)	0.003	(1.068)	0.004	(1.233)	-0.045^b	(2.629)	0.244

Note: All regressions use company fixed effects. A coefficient that is signficant at 95% level is indicated with ^{a}, while ^{b} and ^{c} denote significance at the 99% level and 99.9% level, respectively. T-statistics are shown in parentheses. [1] Market index denotes the log of the SPY price except in the case of the 'Return' regressions where it denotes the return (difference of the log price) of the SPY.

Table 10: Contemporaneous Regressions – Raging Bull Message Board

	Log of Messages		Bullishness Index		Agreement Index		Market Index[1]		R^2
Naive Bayes, Unweighted									
Log Trading Volume	0.348^c	(68.98)	0.049^b	(2.893)	-0.008	(0.44)	-2.518^c	(63.52)	0.995
Log Small Trades	0.311^c	(87.32)	0.060^c	(4.961)	-0.022	(1.725)	-1.626^c	(58.18)	0.983
Log Medium Trades	0.244^c	(56.04)	0.084^c	(5.699)	-0.021	(1.365)	-0.497^c	(14.52)	0.931
Log Large Trades	0.142^c	(40.71)	0.049^c	(4.175)	-0.029^a	(2.355)	-0.248^c	(9.05)	0.642
Return	0.000	(1.709)	0.000^a	(2.17)	-0.000	(1.357)	0.797^c	(136.6)	0.063
Price Volatililty	0.056^c	(30.81)	0.010	(1.675)	-0.013^a	(1.983)	-1.189^c	(82.73)	0.536
Spread	0.005^a	(2.263)	-0.003	(0.452)	0.006	(0.746)	-0.046^b	(2.648)	0.244
Naive Bayes, Weighted by Message Length									
Log Trading Volume	0.348^c	(68.86)	0.047^b	(2.777)	-0.005	(0.271)	-2.518^c	(63.52)	0.995
Log Small Trades	0.310^c	(87.19)	0.058^c	(4.885)	-0.020	(1.585)	-1.626^c	(58.18)	0.983
Log Medium Trades	0.244^c	(55.97)	0.081^c	(5.595)	-0.018	(1.217)	-0.497^c	(14.53)	0.931
Log Large Trades	0.142^c	(40.64)	0.047^c	(4.013)	-0.026^a	(2.158)	-0.248^c	(9.051)	0.642
Return	0.000	(1.716)	0.000^a	(2.271)	-0.000	(1.437)	0.797^c	(136.6)	0.063
Price Volatililty	0.056^c	(30.77)	0.009	(1.51)	-0.012	(1.827)	-1.189^c	(82.73)	0.536
Spread	0.005^a	(2.253)	-0.003	(0.461)	0.006	(0.758)	-0.045^b	(2.648)	0.244
Naive Bayes, Weighted by Inverse Contribution Frequency									
Log Trading Volume	0.348^c	(68.85)	0.040^a	(2.407)	0.001	(0.036)	-2.518^c	(63.52)	0.995
Log Small Trades	0.310^c	(87.14)	0.049^c	(4.118)	-0.011	(0.896)	-1.626^c	(58.17)	0.983
Log Medium Trades	0.244^c	(55.92)	0.074^c	(5.108)	-0.011	(0.753)	-0.497^c	(14.52)	0.931
Log Large Trades	0.142^c	(40.62)	0.043^c	(3.674)	-0.022	(1.853)	-0.248^c	(9.046)	0.642
Return	0.000	(1.711)	0.000^a	(1.97)	-0.000	(1.163)	0.797^c	(136.6)	0.063
Price Volatililty	0.056^c	(30.75)	0.009	(1.485)	-0.011	(1.808)	-1.189^c	(82.73)	0.536
Spread	0.005^a	(2.253)	-0.003	(0.435)	0.006	(0.735)	-0.046^b	(2.648)	0.244
Support Vector, Unweighted									
Log Trading Volume	0.367^c	(72.46)	0.032^c	(3.764)	-0.045^c	(4.631)	-2.516^c	(63.47)	0.995
Log Small Trades	0.329^c	(92.09)	0.025^c	(4.171)	-0.042^c	(6.084)	-1.624^c	(58.09)	0.983
Log Medium Trades	0.261^c	(59.66)	0.042^c	(5.787)	-0.028^c	(3.308)	-0.494^c	(14.45)	0.931
Log Large Trades	0.151^c	(43.14)	0.034^c	(5.887)	-0.036^c	(5.4)	-0.248^c	(9.02)	0.642
Return	0.000	(1.141)	0.000	(0.653)	0.000	(1.471)	0.797^c	(136.6)	0.063
Price Volatililty	0.059^c	(32.12)	0.004	(1.243)	-0.014^c	(3.89)	-1.189^c	(82.73)	0.536
Spread	0.006^b	(2.82)	0.004	(1.141)	-0.004	(1.009)	-0.045^b	(2.647)	0.244
Support Vector, Weighted by Message Length									
Log Trading Volume	0.367^c	(72.16)	0.028^c	(3.302)	-0.042^c	(4.285)	-2.516^c	(63.46)	0.995
Log Small Trades	0.329^c	(91.66)	0.021^c	(3.512)	-0.038^c	(5.547)	-1.624^c	(58.09)	0.983
Log Medium Trades	0.262^c	(59.48)	0.038^c	(5.205)	-0.025^b	(2.977)	-0.494^c	(14.44)	0.931
Log Large Trades	0.152^c	(43.05)	0.031^c	(5.363)	-0.034^c	(5.08)	-0.247^c	(9.018)	0.642
Return	0.000	(1.179)	0.000	(0.554)	0.000	(1.452)	0.797^c	(136.6)	0.063
Price Volatililty	0.059^c	(32.13)	0.004	(1.319)	-0.014^c	(4.119)	-1.189^c	(82.73)	0.536
Spread	0.006^b	(2.843)	0.004	(1.15)	-0.004	(1.039)	-0.045^b	(2.647)	0.244
Support Vector, Weighted by Inverse Contribution Frequency									
Log Trading Volume	0.367^c	(71.73)	0.029^c	(3.53)	-0.042^c	(4.313)	-2.516^c	(63.46)	0.995
Log Small Trades	0.329^c	(91.06)	0.021^c	(3.6)	-0.037^c	(5.356)	-1.624^c	(58.09)	0.983
Log Medium Trades	0.260^c	(58.91)	0.038^c	(5.307)	-0.022^b	(2.625)	-0.494^c	(14.44)	0.931
Log Large Trades	0.152^c	(42.76)	0.032^c	(5.549)	-0.034^c	(5.052)	-0.247^c	(9.017)	0.642
Return	0.000	(1.164)	0.000	(1.078)	0.000	(1.133)	0.797^c	(136.6)	0.063
Price Volatililty	0.060^c	(32.05)	0.004	(1.422)	-0.015^c	(4.267)	-1.189^c	(82.72)	0.536
Spread	0.006^b	(2.832)	0.004	(1.107)	-0.004	(1.011)	-0.045^b	(2.646)	0.244

Note: See table 9 for explanations.

Table 11: Time Sequencing Tests — Yahoo! Finance (15 minutes)

X	Y	$X \Rightarrow Y$						$Y \Rightarrow X$					
		X_{-1}	X_{-2}	X_{-3}	X_{-4}	SPY	χ^2	Y_{-1}	Y_{-2}	Y_{-3}	Y_{-4}	SPY	χ^2
messages	volume	0.025^c	0.024^c	0.005^a	0.000	-0.666^c	$614.^c$	0.147^c	0.007^a	-0.017^c	-0.023^c	-0.016	$5855.^c$
messages	small	0.003^a	0.010^c	0.003^a	0.003^a	-0.312^c	$153.^c$	0.231^c	0.012^a	-0.027^c	-0.031^c	-0.021	$6818.^c$
messages	medium	0.035^c	0.016^c	-0.007^c	-0.014^c	-0.059	$1048.^c$	0.165^c	-0.005	-0.038^c	-0.047^c	-0.277^c	$4448.^c$
messages	large	0.020^c	0.013^c	0.002	-0.004^a	-0.068^a	$491.^c$	0.097^c	-0.001	-0.012^c	-0.014^c	-0.278^c	$1300.^c$
messages	return	-0.000^a	-0.000	-0.000	-0.000	0.823^c	21.0^b	-0.112	-0.586^c	-0.495^b	-0.680^c	-0.231	60.8^c
messages	volatility	0.000^c	0.000^b	-0.000^c	-0.000^c	-0.004^c	$416.^c$	7.281^c	0.613	-1.581^a	-0.933	-0.220^c	$248.^c$
messages	spread	0.004^c	0.002	0.001	0.002	-0.009	52.0^c	0.011^a	-0.007	-0.003	-0.001	-0.286^c	10.9
words	volume	0.007^c	0.007^c	0.002^a	0.001	-0.658^c	$398.^c$	0.276^c	0.021^a	-0.019^a	-0.020^b	-0.047	$3313.^c$
words	small	0.001^b	0.003^c	0.001^a	0.001^a	-0.309^c	$137.^c$	0.409^c	0.026	-0.025	-0.023	-0.096	$3588.^c$
words	medium	0.009^c	0.004^c	-0.001^a	-0.003^c	-0.063^a	$545.^c$	0.309^c	-0.008	-0.064^c	-0.070^c	-0.655^c	$2177.^c$
words	large	0.005^c	0.003^c	0.000	-0.000	-0.074^a	$240.^c$	0.187^c	0.014	-0.016	-0.004	-0.659^c	$752.^c$
words	return	-0.000	-0.000	0.000	-0.000	0.823^c	11.2	-0.522	-1.055^a	-1.088^a	-1.233^b	2.367	31.7^c
words	volatility	0.000^c	0.000^b	-0.000	-0.000	-0.004^c	$220.^c$	12.200^c	2.184	0.151	1.001	-0.494^c	$144.^c$
words	spread	0.001^b	0.000	0.000	0.001^a	-0.010	41.7^c	0.015	-0.014	-0.002	0.016	-0.686^c	5.56
bullishness	volume	0.033^c	0.032^c	0.025^c	0.027^c	-0.690^c	$446.^c$	0.025^c	0.007^c	0.007^c	0.007^c	0.230^c	$2160.^c$
bullishness	small	0.020^c	0.017^c	0.014^c	0.019^c	-0.331^c	$405.^c$	0.041^c	0.011^c	0.010^c	0.009^c	0.228^c	$2873.^c$
bullishness	medium	0.023^c	0.015^c	0.009^a	0.009^b	-0.084^b	$175.^c$	0.022^c	0.005^b	0.004^a	0.003	0.114^c	$945.^c$
bullishness	large	0.019^c	0.015^c	0.013^c	0.014^c	-0.094^c	$215.^c$	0.019^c	0.009^c	0.007^c	0.007^c	0.113^c	$688.^c$
bullishness	return	-0.000	-0.000	-0.000	-0.000	0.823^c	2.38	0.164	-0.051	-0.125	-0.179^a	0.542	18.0^a
bullishness	volatility	0.000^c	0.000^c	0.000	0.000	-0.004^c	82.5^c	2.309^c	0.574	0.990^b	0.796^a	0.162^c	$321.^c$
bullishness	spread	0.005^a	0.007^b	0.002	0.002	-0.016	30.1^c	0.002	0.003	0.002	0.003	0.102^c	7.42
agreement	volume	0.034^c	0.034^c	0.033^c	0.036^c	-0.683^c	$575.^c$	0.015^c	0.008^c	0.009^c	0.009^c	0.165^c	$1848.^c$
agreement	small	0.027^c	0.022^c	0.018^c	0.023^c	-0.328^c	$635.^c$	0.026^c	0.012^c	0.012^c	0.014^c	0.163^c	$2489.^c$
agreement	medium	0.015^c	0.011^c	0.016^c	0.015^c	-0.080^a	$151.^c$	0.009^c	0.006^c	0.006^c	0.006^c	0.060^b	$637.^c$
agreement	large	0.014^c	0.014^c	0.015^c	0.019^c	-0.089^b	$207.^c$	0.012^c	0.009^c	0.009^c	0.008^c	0.061^b	$584.^c$
agreement	return	0.000	-0.000	0.000	-0.000	0.823^c	5.08	0.092	-0.031	-0.111	-0.204^a	0.517	16.3^a
agreement	volatility	0.000^b	0.000^c	0.000^b	0.000^a	-0.004^c	74.2^c	1.756^c	0.512	1.092^c	1.127^c	0.109^c	$313.^c$
agreement	spread	0.004	0.005^a	0.002	0.003	-0.015	21.8^b	0.002	0.002	0.002	0.006^a	0.052^a	13.8^a

Note: The direction of Granger causality is indicated as follows: $X \Rightarrow Y$ indicates X Granger-causes Y, and $Y \Rightarrow X$ indicates Y Granger-causes X. The regressors X_i and Y_i are subscripted by their lags. SPY is a variable with the log of the price of the Standard & Poors Depositary Receipt S&P 500 Tracking Fund, except in the 'Return' regressions where this variable is the time-differenced log of the price. A coefficient that is signficant at the 99% level is indicated with ^{a}, while ^{b} and ^{c} denote significance at a 99.9% level and a 99.99% level, respectively.

Table 12: Time Sequencing Tests — Raging Bull (15 minutes)

X	Y	$X \Rightarrow Y$						$Y \Rightarrow X$					
		X_{-1}	X_{-2}	X_{-3}	X_{-4}	SPY	χ^2	Y_{-1}	Y_{-2}	Y_{-3}	Y_{-4}	SPY	χ^2
messages	volume	0.036^{c}	0.026^{c}	0.004	0.000	-0.676^{c}	$391.^{c}$	0.075^{c}	0.006^{c}	-0.006^{c}	-0.009^{c}	0.149^{c}	$4261.^{c}$
messages	small	0.003	0.009^{c}	-0.000	0.004	-0.313^{c}	53.6^{c}	0.125^{c}	0.008^{a}	-0.015^{c}	-0.013^{c}	0.151^{c}	$5362.^{c}$
messages	medium	0.045^{c}	0.022^{c}	-0.004	-0.010^{c}	-0.074^{a}	$682.^{c}$	0.088^{c}	0.006^{a}	-0.012^{c}	-0.018^{c}	-0.007	$3833.^{c}$
messages	large	0.036^{c}	0.021^{c}	0.005	-0.004	-0.083^{b}	$613.^{c}$	0.055^{c}	0.004	0.001	-0.001	-0.012	$1290.^{c}$
messages	return	-0.000^{c}	-0.000	-0.000	-0.000	0.823^{c}	46.4^{c}	0.036	-0.216	-0.196	-0.371^{c}	-0.784^{a}	31.6^{c}
messages	volatility	0.000^{c}	0.000	-0.000^{c}	-0.000	-0.004^{c}	$270.^{c}$	4.598^{c}	0.302	-1.774^{c}	-0.023	0.015	$242.^{c}$
messages	spread	0.004^{a}	0.001	0.001	0.003	-0.014	19.9^{b}	0.007^{a}	0.001	0.002	0.001	-0.024	9.89
words	volume	0.008^{c}	0.007^{c}	0.002	0.001	-0.675^{c}	$304.^{c}$	0.211^{c}	0.023^{c}	-0.010	-0.010	0.613^{c}	$3388.^{c}$
words	small	0.001	0.003^{c}	0.000	0.002^{c}	-0.315^{c}	74.5^{c}	0.339^{c}	0.037^{c}	-0.036^{c}	-0.009	0.597^{c}	$4111.^{c}$
words	medium	0.010^{c}	0.006^{c}	-0.000	-0.002^{a}	-0.077^{a}	$466.^{c}$	0.238^{c}	0.019^{a}	-0.022^{b}	-0.037^{c}	0.092	$2675.^{c}$
words	large	0.008^{c}	0.005^{c}	0.001	-0.000	-0.086^{b}	$368.^{c}$	0.157^{c}	0.011	0.013	0.009	0.075	$1004.^{c}$
words	return	-0.000^{a}	-0.000	-0.000	-0.000	0.823^{c}	18.7^{b}	-0.217	-0.694	-0.659	-1.067^{b}	0.700	25.7^{c}
words	volatility	0.000^{c}	0.000	-0.000^{b}	0.000	-0.004^{c}	$164.^{c}$	12.897^{c}	2.614	-4.608^{b}	1.600	0.192	$208.^{c}$
words	spread	0.001	0.001	0.000	0.001	-0.014	17.0^{a}	0.032^{b}	0.006	-0.000	-0.001	0.033	14.4^{a}
bullishness	volume	0.046^{c}	0.038^{c}	0.022^{c}	0.026^{c}	-0.671^{c}	$216.^{c}$	0.016^{c}	0.005^{c}	0.002	0.003^{c}	0.109^{c}	$2011.^{c}$
bullishness	small	0.017^{c}	0.020^{c}	0.009^{a}	0.022^{c}	-0.318^{c}	$143.^{c}$	0.028^{c}	0.007^{c}	0.001	0.005^{c}	0.109^{c}	$2816.^{c}$
bullishness	medium	0.046^{c}	0.030^{c}	0.013^{a}	0.011^{a}	-0.081^{a}	$220.^{c}$	0.017^{c}	0.006^{c}	0.002	0.002^{a}	0.045^{c}	$1634.^{c}$
bullishness	large	0.037^{c}	0.030^{c}	0.019^{c}	0.015^{b}	-0.090^{b}	$225.^{c}$	0.013^{c}	0.006^{c}	0.004^{c}	0.005^{c}	0.043^{c}	$755.^{c}$
bullishness	return	-0.000^{b}	-0.000	0.000	-0.000^{b}	0.823^{c}	36.5^{c}	0.026	-0.023	-0.101	-0.085	0.193	11.8
bullishness	volatility	0.000^{c}	0.000	0.000	0.000^{a}	-0.004^{c}	47.0^{c}	1.104^{c}	0.799^{c}	0.060	0.425^{a}	0.067^{c}	$231.^{c}$
bullishness	spread	0.008^{a}	0.000	0.002	0.004	-0.015	11.2	0.002	0.003	0.002	0.003	0.036^{b}	19.5^{b}
agreement	volume	0.049^{c}	0.045^{c}	0.030^{c}	0.036^{c}	-0.671^{c}	$295.^{c}$	0.011^{c}	0.005^{c}	0.003^{c}	0.005^{c}	0.092^{c}	$1730.^{c}$
agreement	small	0.029^{c}	0.025^{c}	0.018^{c}	0.028^{c}	-0.320^{c}	$290.^{c}$	0.020^{c}	0.008^{c}	0.002	0.008^{c}	0.092^{c}	$2438.^{c}$
agreement	medium	0.038^{c}	0.028^{c}	0.021^{c}	0.020^{c}	-0.079^{a}	$205.^{c}$	0.011^{c}	0.006^{c}	0.003^{c}	0.004^{c}	0.033^{a}	$1293.^{c}$
agreement	large	0.033^{c}	0.031^{c}	0.020^{c}	0.018^{c}	-0.088^{b}	$208.^{c}$	0.009^{c}	0.006^{c}	0.005^{c}	0.005^{c}	0.031^{a}	$655.^{c}$
agreement	return	-0.000	-0.000	-0.000	-0.000^{c}	0.823^{c}	23.0^{b}	-0.118^{a}	-0.020	-0.125^{a}	-0.078	0.189	23.8^{c}
agreement	volatility	0.000^{c}	0.000	0.000^{a}	0.000^{a}	-0.004^{c}	57.9^{c}	1.013^{c}	0.754^{c}	-0.021	0.646^{c}	0.056^{c}	$245.^{c}$
agreement	spread	0.007	0.002	0.001	0.002	-0.014	6.74	0.002	0.002	0.001	0.003^{a}	0.025	15.7^{a}

Note: See table 11 for explanations.

Table 13: Time Sequencing Tests — Yahoo! Finance (1 day)

X	Y	$X \Rightarrow Y$									$Y \Rightarrow X$								
		X_{-1}	X_{-2}	WSJ_{-2}	WSJ_{-1}	WSJ_0	WSJ_1	NWK	SPY	χ^2	Y_{-1}	Y_{-2}	WSJ_{-2}	WSJ_{-1}	WSJ_0	WSJ_1	NWK	SPY	χ^2
messages	volume	0.108^c	-0.031^c	0.029^a	0.076^c	0.024^a	-0.025^a	-0.164^c	-0.983^c	268.^{c}	-0.025	0.056^b	-0.011	0.043^b	0.027	-0.001	-0.540^c	-0.011	14.9^b
messages	small	0.072^c	-0.028^c	0.028^b	0.051^c	0.016	-0.016	-0.049^c	-0.495^c	190.^{c}	-0.051^a	0.094^c	-0.011	0.043^b	0.027	-0.001	-0.543^c	-0.027	25.0^c
messages	medium	0.048^c	-0.053^c	0.032^a	0.062^c	0.008	-0.027^a	-0.109^c	0.220	57.8^c	-0.027	0.009	-0.013	0.042^b	0.027	0.001	-0.539^c	-0.062	6.41
messages	large	0.094^c	-0.067^c	0.039^a	0.086^c	0.035^a	-0.043^a	-0.215^c	0.052	88.1^c	-0.023	0.011	-0.013	0.042^b	0.027	0.002	-0.537^c	-0.076	6.21
messages	return	-0.001	0.001	0.001	-0.002	0.000	-0.001	-0.004^a	1.257^c	1.89	-0.098	0.135	-0.012	0.042^b	0.026	0.001	-0.534^c	0.924	2.10
messages	volatility	0.000^c	-0.000^b	0.000	0.000^a	0.000	0.000	-0.000	-0.006^c	26.5^c	-3.915	5.537	-0.012	0.042^b	0.027	0.001	-0.535^c	-0.058	3.89
messages	spread	0.002	-0.002	0.003	0.004	0.000	0.001	0.000	-0.041	2.00	-0.030	-0.041	-0.013	0.042^b	0.027	0.001	-0.534^c	-0.075	1.06
words	volume	0.043^c	-0.006	0.029^a	0.079^c	0.031^b	-0.025^a	-0.133^c	-0.945^c	114.^{c}	0.038	0.083^b	-0.027	0.034	0.065^b	0.009	-0.600^c	-0.227	33.8^c
words	small	0.025^c	-0.007	0.027^b	0.053^c	0.021^b	-0.016	-0.024^a	-0.484^c	61.4^c	0.010	0.126^c	-0.028	0.034	0.067^b	0.010	-0.600^c	-0.322	35.1^c
words	medium	0.017^b	-0.020^c	0.030^a	0.062^c	0.011	-0.029^a	-0.092^c	0.213	20.8^c	-0.015	0.016	-0.029	0.034	0.070^b	0.016	-0.604^c	-0.484	0.79
words	large	0.035^c	-0.023^b	0.037	0.088^c	0.041^a	-0.045^b	-0.186^c	0.035	30.8^c	0.001	0.013	-0.029	0.034	0.069^b	0.015	-0.602^c	-0.483	1.38
words	return	-0.001	0.000	0.001	-0.002	0.000	-0.001	-0.004^a	1.260^c	1.28	-0.297	0.212	-0.030	0.033	0.068^b	0.017	-0.600^c	0.281	4.01
words	volatility	0.000^a	-0.000	0.000	0.000^a	0.000	0.000	-0.000	-0.006^c	7.18	-0.028	6.547	-0.029	0.034	0.069^b	0.016	-0.602^c	-0.403	3.09
words	spread	0.001	-0.000	0.003	0.004	0.001	0.001	0.001	-0.041	0.36	-0.012	-0.020	-0.029	0.034	0.069^b	0.016	-0.601^c	-0.484	0.09
bullishness	volume	0.079^c	-0.023	0.030^a	0.081^c	0.035^c	-0.022	-0.098^c	-0.978^c	18.8^c	0.025^c	0.010	0.014^a	0.033^c	0.010	0.008	-0.011	0.416^c	61.3^c
bullishness	small	0.051^b	-0.032	0.027^b	0.054^c	0.023^b	-0.014	-0.002	-0.504^c	15.9^b	0.045^c	0.006	0.014^a	0.032^c	0.010	0.008	-0.009	0.411^c	102.^{c}
bullishness	medium	0.037	-0.090^c	0.030^a	0.062^c	0.012	-0.028^a	-0.075^c	0.233	18.9^c	0.013^a	0.003	0.014^a	0.033^c	0.012^a	0.010	-0.011	0.319^c	30.0^c
bullishness	large	0.098^b	-0.031	0.036	0.088^c	0.043^b	-0.046^b	-0.155^c	-0.021	12.4^a	0.006	0.008	0.014^a	0.034^c	0.012^a	0.009	-0.012	0.330^c	20.9^c
bullishness	return	0.001	-0.001	0.001	-0.002	0.000	-0.001	-0.004^b	1.263^c	0.38	-0.005	0.030	0.014^a	0.033^c	0.012^a	0.010	-0.012	0.054	0.61
bullishness	volatility	0.000	-0.000	0.000	0.000^a	0.000	0.000	-0.000	-0.006^c	7.15	2.150	0.730	0.014^a	0.033^c	0.012^a	0.010	-0.012	0.364^c	10.5^a
bullishness	spread	0.014^a	-0.001	0.003	0.004	0.000	0.000	0.002	-0.047	9.13	0.014	0.019	0.014^a	0.033^c	0.012^a	0.010	-0.012	0.324^c	1.93
agreement	volume	0.129^c	-0.062	0.030^a	0.081^c	0.036^c	-0.022	-0.095^c	-0.963^c	18.2^b	0.012^c	0.004	0.005	0.019^c	0.005	0.004	0.004	0.195^c	37.9^c
agreement	small	0.081^a	-0.056	0.027^b	0.054^c	0.023^b	-0.014	-0.001	-0.500^c	14.1^b	0.025^c	0.003	0.005	0.019^c	0.004	0.004	0.005	0.200^c	89.8^c
agreement	medium	0.095	-0.141^b	0.029^a	0.062^c	0.011	-0.029^a	-0.073^c	0.217	18.3^b	0.009^b	0.003	0.005	0.019^c	0.006	0.005	0.004	0.151^c	47.3^c
agreement	large	0.216^c	-0.060	0.036	0.088^c	0.042^b	-0.045^b	-0.152^c	-0.025	19.8^c	0.003	0.005^a	0.005	0.019^c	0.005	0.005	0.003	0.159^c	23.3^c
agreement	return	-0.002	-0.004	0.001	-0.002	0.000	-0.001	-0.004^b	1.263^c	1.65	-0.001	0.026	0.005	0.019^c	0.006	0.005	0.003	0.066	1.40
agreement	volatility	0.000	-0.000	0.000	0.000^a	0.000	0.000	-0.000	-0.006^c	4.39	1.180	-0.095	0.005	0.019^c	0.006	0.005	0.004	0.171^c	5.79
agreement	spread	0.027^b	-0.002	0.003	0.004	0.000	0.000	0.002	-0.047	12.2^a	0.014	0.031^a	0.005	0.019^c	0.006	0.005	0.003	0.155^c	11.1^a

Note: See table 11 for explanations. In addition, variables WSJ_{-2}, WSJ_{-1}, WSJ_0, WSJ_{+1} indicate how many stories were released in the *Wall Street Journal* on a given day: two days after, one day after, on the same day, an the day before the current day, respectively. NWK is an indicator variable for a day being the first trading day after a weekend or holiday.

Table 14: Time Sequencing Tests — Raging Bull (1 day)

X	Y	$X \Rightarrow Y$									$Y \Rightarrow X$								
		X_{-1}	X_{-2}	WSJ_{-2}	WSJ_{-1}	WSJ_0	WSJ_1	NWK	SPY	χ^2	Y_{-1}	Y_{-2}	WSJ_{-2}	WSJ_{-1}	WSJ_0	WSJ_1	NWK	SPY	χ^2
messages	volume	0.070^{c}	-0.018^{a}	0.029^{a}	0.079^{c}	0.030^{b}	-0.023^{a}	-0.129^{c}	-1.050^{c}	$148.^{c}$	0.072^{c}	0.033	-0.013	0.067^{c}	0.040^{a}	-0.016	-0.354^{c}	0.713^{c}	43.5^{c}
messages	small	0.043^{c}	-0.011	0.027^{b}	0.053^{c}	0.019^{a}	-0.015	-0.023^{a}	-0.546^{c}	92.7^{c}	0.091^{c}	0.071^{a}	-0.015	0.066^{c}	0.041^{a}	-0.016	-0.348^{c}	0.697^{c}	71.0^{c}
messages	medium	0.042^{c}	-0.042^{c}	0.030^{a}	0.062^{c}	0.009	-0.028^{a}	-0.097^{c}	0.210	54.4^{c}	0.018	0.025	-0.014	0.068^{c}	0.044^{a}	-0.012	-0.359^{c}	0.432^{a}	17.2^{b}
messages	large	0.073^{c}	-0.039^{c}	0.036	0.086^{c}	0.037^{a}	-0.045^{b}	-0.188^{c}	-0.008	68.7^{c}	0.039^{b}	0.003	-0.013	0.067^{c}	0.042^{a}	-0.013	-0.357^{c}	0.465^{a}	20.5^{c}
messages	return	-0.000	0.001	0.001	-0.002	0.000	-0.001	-0.004^{b}	1.261^{c}	1.51	-0.126	0.081	-0.014	0.067^{c}	0.043^{a}	-0.011	-0.360^{c}	1.005	1.30
messages	volatility	0.000^{c}	-0.000^{a}	0.000	0.000^{a}	0.000	0.000	-0.000	-0.006^{c}	18.0^{b}	-1.861	9.843^{a}	-0.016	0.067^{c}	0.044^{a}	-0.012	-0.361^{c}	0.555^{a}	10.9^{a}
messages	spread	0.005^{a}	-0.002	0.003	0.004	-0.000	0.000	-0.000	-0.043	11.0^{a}	0.171^{a}	-0.011	-0.016	0.066^{c}	0.043^{a}	-0.012	-0.362^{c}	0.454^{a}	7.44
words	volume	0.015^{c}	-0.001	0.029^{a}	0.081^{c}	0.034^{c}	-0.023^{a}	-0.112^{c}	-1.003^{c}	53.4^{c}	0.268^{c}	0.114^{a}	-0.037	0.145^{b}	0.125^{b}	-0.050	-0.555^{c}	2.029^{c}	91.9^{c}
words	small	0.008^{c}	-0.000	0.026^{b}	0.054^{c}	0.022^{b}	-0.015	-0.011	-0.520^{c}	27.4^{c}	0.303^{c}	0.199^{b}	-0.044	0.141^{b}	0.129^{b}	-0.049	-0.542^{c}	1.830^{c}	$111.^{c}$
words	medium	0.008^{b}	-0.008^{a}	0.029^{a}	0.062^{c}	0.011	-0.030^{a}	-0.085^{c}	0.208	16.1^{b}	0.066	0.062	-0.041	0.151^{b}	0.146^{c}	-0.029	-0.566^{c}	1.013	22.4^{c}
words	large	0.014^{c}	-0.006	0.035	0.088^{c}	0.042^{b}	-0.045^{b}	-0.169^{c}	-0.004	18.1^{b}	0.152^{c}	0.003	-0.036	0.148^{b}	0.139^{b}	-0.036	-0.564^{c}	1.124	42.4^{c}
words	return	0.000	0.000	0.001	-0.002	0.000	-0.001	-0.004^{b}	1.263^{c}	0.60	-0.659	-0.014	-0.041	0.149^{b}	0.146^{c}	-0.025	-0.568^{c}	1.726	3.86
words	volatility	0.000	-0.000	0.000	0.000^{a}	0.000	0.000	-0.000	-0.006^{c}	2.21	11.889	17.455	-0.044	0.148^{b}	0.146^{c}	-0.029	-0.569^{c}	1.467^{a}	13.2^{a}
words	spread	0.002^{b}	-0.000	0.003	0.004	0.000	0.000	-0.000	-0.044	13.7^{a}	0.386	0.024	-0.045	0.147^{b}	0.145^{c}	-0.027	-0.575^{c}	1.079	5.52
bullishness	volume	0.047^{b}	-0.018	0.029^{a}	0.082^{c}	0.037^{c}	-0.022	-0.100^{c}	-0.959^{c}	13.2^{a}	0.024^{b}	0.006	0.002	0.017^{a}	0.004	0.001	-0.012	0.392^{c}	28.0^{c}
bullishness	small	0.038^{b}	-0.016	0.026^{b}	0.054^{c}	0.024^{b}	-0.014	-0.004	-0.504^{c}	14.9^{b}	0.040^{c}	0.008	0.001	0.017^{a}	0.003	0.000	-0.010	0.387^{c}	47.8^{c}
bullishness	medium	0.049^{a}	-0.034	0.028^{a}	0.062^{c}	0.012	-0.030^{a}	-0.078^{c}	0.203	13.0^{a}	0.016^{a}	0.008	0.002	0.018^{a}	0.005	0.002	-0.012	0.304^{c}	32.7^{c}
bullishness	large	0.059^{a}	-0.011	0.035	0.088^{c}	0.045^{b}	-0.045^{b}	-0.158^{c}	-0.011	8.20	0.017^{c}	0.002	0.003	0.018^{a}	0.005	0.002	-0.013	0.321^{c}	25.7^{c}
bullishness	return	0.001	0.000	0.001	-0.002	0.000	-0.001	-0.004^{b}	1.263^{c}	0.93	-0.013	-0.010	0.002	0.018^{a}	0.006	0.003	-0.012	0.429	0.10
bullishness	volatility	0.000	-0.000	0.000	0.000^{a}	0.000	0.000	-0.000	-0.006^{c}	0.30	0.199	1.912	0.002	0.018^{a}	0.006	0.003	-0.013	0.342^{c}	3.35
bullishness	spread	0.009^{a}	-0.001	0.003	0.004	0.000	0.001	0.001	-0.045	7.14	0.038	0.027	0.002	0.018^{a}	0.006	0.003	-0.013	0.314^{c}	3.99
agreement	volume	0.028	-0.023	0.030^{a}	0.082^{c}	0.037^{c}	-0.021	-0.099^{c}	-0.945^{c}	3.74	0.008	-0.001	-0.001	0.007	0.001	0.001	-0.002	0.202^{b}	3.79
agreement	small	0.028	-0.016	0.026^{b}	0.054^{c}	0.024^{b}	-0.014	-0.003	-0.496^{c}	5.08	0.013	-0.001	-0.001	0.007	0.001	0.001	-0.002	0.202^{c}	6.26
agreement	medium	0.049	-0.027	0.029^{a}	0.062^{c}	0.012	-0.030^{a}	-0.077^{c}	0.204	6.74	0.008	0.001	-0.001	0.007	0.002	0.002	-0.002	0.180^{b}	8.15
agreement	large	0.022	-0.004	0.035	0.089^{c}	0.045^{b}	-0.045^{b}	-0.157^{c}	0.004	0.63	0.009^{a}	-0.001	-0.001	0.007	0.002	0.001	-0.002	0.186^{b}	9.25^{a}
agreement	return	0.002	0.000	0.001	-0.002	0.000	-0.001	-0.004^{b}	1.263^{c}	0.87	0.017	0.015	-0.001	0.007	0.002	0.002	-0.003	0.163	0.36
agreement	volatility	-0.000	0.000	0.000	0.000^{a}	0.000	0.000	-0.000	-0.006^{c}	0.12	0.176	-0.210	-0.001	0.007	0.002	0.002	-0.003	0.184^{b}	0.05
agreement	spread	0.008	-0.003	0.003	0.004	0.001	0.001	0.001	-0.043	3.40	-0.001	0.003	-0.001	0.007	0.002	0.002	-0.003	0.184^{b}	0.03

Note: See table 13 for explanations.

Table 15: Does Posting Activity Predict News Releases in the Wall Street Journal?

	Yahoo! Finance, 1 day lag				
	(1)	(2)	(3)	(4)	(5)
Log Messages	0.5694^{c} (224.2)	0.5681^{c} (219.7)			
Log Words			0.3527^{c} (124.8)	0.3719^{c} (144.1)	0.3488^{c} (120.1)
Bullishness	0.2982^{a} (5.017)	0.3528 (2.266)	0.4796^{c} (14.24)		0.7134^{b} (9.372)
Agreement		−0.115 (0.08)		0.5022^{a} (5.646)	−0.482 (1.449)
Log Likelihood	−4403	−4403	−4452	−4456	−4451
Observations	14235	14235	14235	14235	14235
Companies	39	39	39	39	39

	Raging Bull, 1 day lag				
	(1)	(2)	(3)	(4)	(5)
Log Messages	0.4367^{c} (169.3)	0.4386^{c} (159.4)			
Log Words			0.1519^{c} (102.6)	0.1589^{c} (118.7)	0.1487^{c} (96.55)
Bullishness	0.1637 (2.68)	0.1318 (0.539)	0.1165 (1.389)		0.476^{b} (6.667)
Agreement		0.0508 (0.045)		−0.061 (0.215)	-0.575^{a} (5.549)
Log Likelihood	−4452	−4452	−4486	−4486	−4483
Observations	14235	14235	14235	14235	14235
Companies	39	39	39	39	39

	Yahoo! Finance, 2 day lag				
	(1)	(2)	(3)	(4)	(5)
Log Messages	0.0524 (2.236)	0.0512 (2.087)			
Log Words			0.049 (3.797)	0.045 (3.381)	0.0483 (3.669)
Bullishness	−0.147 (1.31)	−0.104 (0.252)	−0.157 (1.546)		−0.111 (0.3)
Agreement		−0.096 (0.069)		−0.261 (1.305)	−0.105 (0.085)
Log Likelihood	−4544	−4544	−4543	−4543	−4543
Observations	14196	14196	14196	14196	14196
Companies	39	39	39	39	39

	Raging Bull, 2 day lag				
	(1)	(2)	(3)	(4)	(5)
Log Messages	0.0085 (0.066)	0.0157 (0.214)			
Log Words			0.0155 (1.18)	0.0123 (0.785)	0.0162 (1.287)
Bullishness	−0.041 (0.164)	−0.165 (0.95)	−0.072 (0.482)		−0.185 (1.273)
Agreement		0.2022 (0.817)		−0.003 (0)	0.1906 (0.773)
Log Likelihood	−4545	−4545	−4545	−4545	−4544
Observations	14196	14196	14196	14196	14196
Companies	39	39	39	39	39

Note: Estimates are obtained from logit regressions with company fixed effects where the dependent variable is a binary response variable which is 1 when the Wall Street Journal has published an article on a particular company on a given day, and 0 otherwise. The first two panels predict today's news release based on yesterday's posting activity, while the last two panels predict today's news release based on the posting activity on the day before yesterday. Companies which zero or fewer than 5 WSJ releases were dropped from this analysis. A coefficient that is signficant at 95% level is indicated with ^{a}, while ^{b} and ^{c} denote significance at a 99% level and a 99.9% level respectively. Numbers in parentheses provide the Wald χ^2 statistics on which the significance determination was based.