Some Experiments on Modeling Stock Market Behavior Using Investor Sentiment Analysis and Posting Volume from Twitter

Nuno Oliveira
Centro Algoritmi
Dep. Sistemas de Informação
Universidade do Minho
4800-058 Guimarães,
Portugal
nunomroliveira@gmail.com

Paulo Cortez
Centro Algoritmi
Dep. Sistemas de Informação
Universidade do Minho
4800-058 Guimarães,
Portugal
pcortez@dsi.uminho.pt

Nelson Areal
Department of Management
School of Economics and
Managment
University of Minho
4710-057 Braga, Portugal
nareal@eeg.uminho.pt

ABSTRACT

The analysis of microblogging data related with stock markets can reveal relevant new signals of investor sentiment and attention. It may also provide sentiment and attention indicators in a more rapid and cost-effective manner than other sources. In this study, we created several indicators using Twitter data and investigated their value when modeling relevant stock market variables, namely returns, trading volume and volatility. We collected recent data from nine major technological companies. Several sentiment analysis methods were explored, by comparing 5 popular lexical resources and two novel lexicons (emoticon based and the merge of all 6 lexicons) and sentiment indicators produced using two strategies (based on daily words and individual tweet classifications). Also, we measured posting volume associated with tweets related to the analyzed companies. While a short time period is considered (32 days), we found scarce evidence that sentiment indicators can explain these stock returns. However, interesting results were obtained when measuring the value of using posting volume for fitting trading volume and, in particular, volatility.

Categories and Subject Descriptors

Information Systems [**Information systems applications**]: Data mining; Information Systems [**World Wide Web**]: Web applications, Social networks

General Terms

Economics, Experimentation, Languages

Keywords

Text Mining, Sentiment Analysis, Microblogging Data, Returns, Trading Volume, Volatility.

WIMS'13, June 12-14, 2013 Madrid, Spain

1. INTRODUCTION

The analysis and forecasting of stock market behavior has been a focus of academics and practitioners alike. A model that accounts for investor sentiment and attention can provide a better explanation of the cross-section of stock market returns and can also contain useful information to forecast stock market volatility and perhaps even stock market returns. Some studies have shown that individual's financial decisions are significantly affected by their emotion and mood [13, 11]. Investors' emotions can affect the way they process and react to new information which explains why their decisions in some circumstances can depart from the rational behavior that is generally assumed. If emotions can affect decisions at the individual level it can be argued that investors' collective sentiment can affect market returns and their dynamics.

Several arguments can be made for the use of microblogging data as a valuable source to predict investor sentiment:

- The community of users that utilizes these services to communicate and share information about stock market issues has grown and is potentially more representative of all investors. The analysis of the data they generate can allow data mining of investor sentiment in several dimensions.
- Microblogging data is readily available at low cost permitting a faster and less expensive creation of indicators, compared to traditional sources (e.g. large-scale surveys) and can also contain new information that is not present in historical quantitative financial data.
- The small size of the message (maximum 140 characters) and the usage of cashtags (a hashtag identifier for financial stocks) can make it a less noisy source of data.
- Users post very frequently, reacting to events in real-time. This regularity allows a real-time sentiment assessment that can be exploited during the trading day.

Mining microblogging data to forecast stock market behavior is a very recent research topic that has already presented promising results [3, 16, 9, 12]. Bollen et al. [3] measured collective mood states ("positive", "negative", "calm", "alert", "sure", "vital", "kind", and "happy") through senti-

ment analysis applied to large scale Twitter data. Tweets were filtered by some generic sentiment expressions (e.g. "I'm feeling") and were not directly related to stock market. They analyzed the text by two mood tracking tools, namely OpinionFinder [19] that classifies tweets as positive or negative mood, and Google-Profile of Mood States that measures mood in the other 6 dimensions. They found an accuracy of 86.7% in the prediction of the Dow Jones Industrial Average daily directions. Sprenger and Welpe [16] have used sentiment analysis on stock related tweets collected during a 6-month period. To reduce noise, they selected Twitter messages containing cashtags of S&P 100 companies. Each message was classified by a Naïve Bayes method trained with a set of 2,500 tweets. Results showed that sentiment indicators are associated with abnormal returns and message volume is correlated to the trading volume. Mao et al. [9] surveyed a variety of web data sources (Twitter, news headlines and Google search queries) and tested two sentiment analysis methods used for the prediction of stock market behavior. They used a random sample of all public tweets and defined a tweet as bullish or bearish only if it contained the terms "bullish" or "bearish". They showed that their Twitter sentiment indicator and the frequency of occurrence of financial terms on Twitter are statistically significant predictors of daily market returns. Oh and Sheng [12] resorted to a microblogging service exclusively dedicated to stock market. They collected 72,221 micro blog postings from `Stocktwits.com`, over a period of three months. The sentiment of the messages was classified by a bag of words approach [15] that applies a machine learning algorithm J48 classifier to produce a learning model. They verified that the extracted sentiment appears to have strong predictive value for future market directions.

Our study aims at testing whether Twitter data sentiment variables have any correlation with stock market variables and if they are related to stock market dynamics. Five different popular lexical resources are compared and two novel lexicons are proposed for extracting a sentiment indicator, emoticon based (e.g. ":)") and ALL, which merges all six previous lexicons. We test different sentiment classifications, under two proposed strategies (S1 – sentiment word based and S2 – individual tweet sentiment based) and their respective impact in explaining stock returns. We also explore the use of posting volume and their value in modeling the volume and volatility financial indicators. These methods are applied for nine large technological companies, for which very recent Twitter data was collected using their respective cashtags. While a short period was analyzed (32 days), promising results were achieved for fitting the volume and in particular volatility.

The rest of the paper is organized as follows. Section 2 describes the data and methods. Next, Section 3 presents and discusses the research results. Finally, Section 4 concludes with a summary and discussion of the main results.

2. MATERIALS AND METHODS

2.1 Data Overview

Data was collected for nine large US technological companies: AMD (AMD), Amazon (AMZN), Dell (DELL), Ebay (EBAY), HP (HPQ), Google (GOOG), IBM (IBM), Intel (INTC) and Microsoft (MSFT). These companies were chosen because they belong to a sector that has a substantial posting volume on Twitter and therefore can be indicative of investors' level of attention on these stocks. For each company, we collected Twitter and stock market data, on a daily basis (considering working days) from December 24, 2012 to February 8, 2013. Figure 1 plots the total number of tweets collected for each technological company.

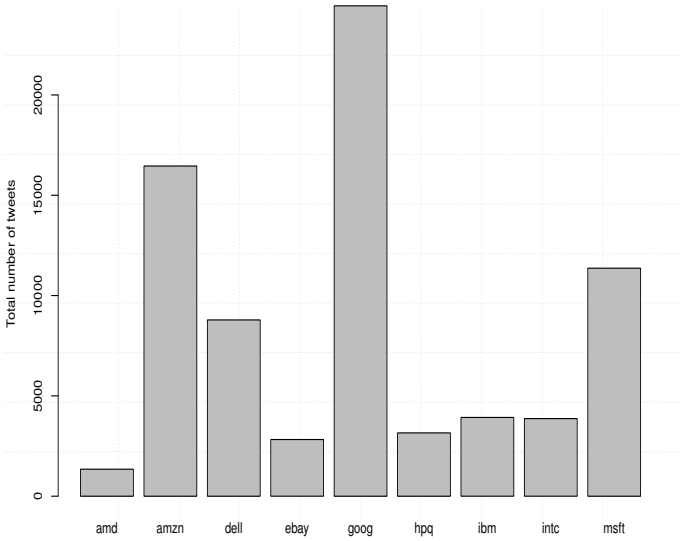

Figure 1: Total number of tweets collected for the nine selected technological companies

2.2 Twitter Data Collection

We selected Twitter as a source of microblogging data since it is by far the most popular service and also because their API enables the collection of a large number of messages. All daily tweets were collected by using the Twitter REST API[1]. Messages were filtered by the company cashtags, i.e., $AMD, $AMZN, $DELL, $EBAY, $HPQ, $GOOG, $IBM, $INTC, $MSFT. Cashtags are composed by the company ticker preceded by the "$" symbol. These symbols are commonly used by the investor community in discussions related to the respective company. Concentrating on only these messages reduces the amount of irrelevant messages, resulting in a less noisy data set.

2.3 Stock Market Data

The stock market variables here considered are daily returns, trading volume and volatility. The data was collected from Thompson Reuters Datastream[2].

Return is the percentage change in the asset value. We used the adjusted close prices to calculate returns. Adjusted close price is the official closing price adjusted for capital actions and dividends. We computed returns (r_t) using the following formulae:

$$r_t = (P_t - P_{t-1})/P_{t-1} \quad (1)$$

where P_t is the adjusted close price of day t and P_{t-1} is the adjusted close price of the preceding day. There is scarce

[1] `https://dev.twitter.com/docs/api`
[2] `http://online.thomsonreuters.com/datastream/`

evidence of return predictability [18]. Nevertheless returns provide useful information about the probability distribution of asset prices. This is essential for investors and portfolio managers as they use this information to value assets and manage their risk exposure.

Volatility (σ_t, for day t) is a measure of total risk associated with a given investment. Volatility can be estimated using several different approaches. Previous studies have found that implied volatility contained in option prices is an appropriate estimator of volatility [5]. The average of the implied volatility for a 30-day to maturity Call and Put options contracts for each stock is here used to measure volatility. Estimates of volatility are essential for portfolio selection, financial assets valuation and risk management.

Trading volume is the number of shares traded in each day during a trading session. Volume can be used to measure stock liquidity, which in turn has been shown to be useful in assset pricing as several theoretical and empirical studies have identified a liquidity premium. Liquidity can help to explain the cross-section of expected returns, [3]

2.4 Sentiment Analysis Methods

We describe the lexical resources and the sentiment analysis methods in this subsection.

2.4.1 Lexical Resources

In our sentiment analysis methods, we exploited 5 different and popular lexical resources to evaluate the usefulness of each resource, as well as their complementarity. The lexical resources are:

1. **Harvard General Inquirer (GI)** [17] – This resource comprise 11788 words classified in 182 categories. These categories come from four sources: the Harvard IV-4 dictionary; the Lasswell value dictionary; categories recently constructed, and "marker" categories containing syntactic and semantic markers. We exploited this resource by producing a list with the "positive" category words and another list with the "negative" category words. The syntactic information was discarded because we did not analyze the text syntactically.
2. **Opinion Lexicon (OL)** [8] – This lexicon contains two lists of positive and negative opinion words for English, including misspelled words that appear frequently in social media contents. We applied this lexicon without any transformation.
3. **Macquarie Semantic Orientation Lexicon (MSOL)** [10]. – classifies more than 75 thousand n-grams, as positive or negative. In this study, we only considered the unigrams (1-grams).
4. **MPQA Subjectivity Lexicon (MPQA)** [20] – this lexicon is part of OpinionFinder, a system that identifies various aspects of subjectivity (e.g. sources of opinion, sentiment expressions) in text. MPQA Subjectivity Lexicon has more than 8000 entries with the following attributes:
   - Type: The word is classified as strongsubj if it is subjective in most contexts and it is considered weaksubj if it only has certain subjective usages.
   - Len: Refers to the number of words in the entry. This attribute can be discarded because all entries are composed by single words.
   - Word1: Contains the word or its stem.
   - Pos1: Identifies the part of speech of the word (i.e. noun, verb, adverb or adjective). It may be anypos, meaning that part of speech is irrelevant for the polarity.
   - Stemmed1 - Indicates whether the word is stemmed (y) or not (n).
   - Priorpolarity - Classifies the out of context polarity of the word. It may be positive, negative, both or neutral.

   In this paper we only needed to use the attributes "word1" e "priorpolarity" because we did not perform part of speech tagging or stemming.
5. **SentiWordNet (SWN) 3.0** [2] – a lexical resource that assigns, to each synset of WordNet, a positivity and a negativity score, varying from 0 to 1. A synset is a group of word or expressions that are semantically equivalent in some context. Each word may appear multiple times with different scores in this lexical resource because it can belong to various synsets of Wordnet. In this paper, we used the average positivity and negativity score for each word because we did not analyze the contextual polarity.

We propose two additional lexicons, termed Emoticons and ALL. The former is based on the simpler analysis of positive (":-)' or ":)") and negative (":-(" or ":(") emoticons. If a positive emoticon is present in the text, then we add 1 to the positivity score and similarly we increase (+1) the negative score if a negative emoticon is detected. The latter lexicon merges all 6 previous lexicons (GI, OL, MSOL, MPQA, SWN, Emoticons) by producing a union of all positive, negative and neutral score rules.

2.4.2 Sentiment Analysis Methods

While more complex parsing techniques could potentially lead to better results (e.g. use of semantic knowledge), in this paper we propose and explore two simple and global sentiment analysis approaches. The rational for this choice is that the proposed approaches are very easy to implement and test. For example, a tweet that contains "I do not really like $GOOG prices" will have a neutral effect under both proposed strategies and it is not wrongly classified as positive, while correctly identifying equivalent or even more complex negative posts would require a quite sophisticated parsing that is out of scope of this work. Moreover, since we analyze a very large number of daily tweets (e.g. several thousands of posts for GOOG, see Figure 1), these simple approaches should produce good global results.

In the first approach (S1), we count the daily total number of words that are considered positive and negative by each lexical resource (total of two sentiment variables). As an example, if a tweet has 2 positive words and 3 negative words, we add 2 to the daily positivity score and 3 to the daily negativity score. In the SentiWordNet situation, we add

[3]For more on liquidity please refer to [1] and references therein.

the positivity and negativity score of each word. Regarding the second sentiment approach (S2), we classify each individual tweet, by considering the "positive" and "negative" words that it contains. A message is considered: positive, if the number of "positive" words is higher than the number of "negative" words; negative, if the number of "negative" words is higher than the number of "positive" words; and neutral, if the number of both word polarity types is equal. In the SentiWordNet approach, we compared the total positivity and total negativity score for each tweet. The total of sensitive variables measured is thus two for S1, total number of positive and negative words, and three for S2 (total number of positive, neutral and negative classified tweets).

2.5 Analysis of the relationship between microblogging features and stock market variables

In this subsection we verify the statistical relationship between some microblogging sentiment variables and each stock return, volatility and trading volume. This will allow us to infer if there is any connection between microblogging data and stock market variables, and if so, assess whether it can contribute to a better modeling of stock market dynamics.

2.6 Multiple regression model

Given that we have a small number of samples to fit (around 30), we explicitly opt for a very simple predictive model, under the Occam's razor principle giving that it has few internal parameters, thus it less prone to overfit the data: the multiple regression model. Such model is defined by the equation [7]:

$$\hat{y} = \beta_0 + \sum_{i=1}^{I} \beta_i x_i \tag{2}$$

where $\hat{y}$ is the predicted value for the dependent variable y (target output), x_i are the independent variables (inputs) and $\beta_0, \ldots, \beta_i$ are the set of parameters to be adjusted, usually by applying a least squares algorithm. Due to its additive nature, this model is easy to interpret and has been widely used in several areas, including the Finance domain.

2.7 Evaluation

To measure the quality of fit of the regression models we use two metrics: coefficient of determination R^2 and Relative Absolute Error (RAE). These are given by [21]:

$$\begin{aligned} R^2 &= 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2} \\ RAE &= \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|y_i - \overline{y_i}|} \end{aligned} \tag{3}$$

where y_i and $\hat{y}_i$ are the target and fitted value for the i-th day, N is the number of days considered and $\overline{y_i}$ is the average of the target values. Both metrics are scale independent. The ideal regression will produce a R^2 of 1.0, while an R^2 closer to 0 indicates a bad fit. The lower the RAE, the better the model, where 1.0 means that the regression method has similar performance as the constant average predictor. When compared with RAE, R^2 is more sensitive to high individual errors. Depending on the regression model input variables, the value of N corresponds to 31, if only previous day values are used ($d-1$), or 30, if a lag of two days ($d-2$) is included. As a baseline method, we adopt a regression model that has one input: the target financial index but from the previous day ($t-1$). For all metrics, we measure the value of Twitter based data if the respective regression model is better than the baseline method.

2.7.1 Returns

The relationship between the information content of microblogging data and daily returns was tested by regressing today's return ($y = r_t$) for each company on several combinations of microblogging variables. We modeled S1 and S2 sentiment variables for each lexical resource (GI, OL, MSOL, MPQA, SWN, Emoticons and ALL). For all these models, there is only input $x_1 = Pos - Neg$, where Pos and Neg denote the positive and negative counts (according to method S1 or S2). The baseline uses the input $x_1 = r_{t-1}$. We also test a regression model that combines the sentiment variable with the baseline ($x_1 \oplus r_{t-1}$).

2.7.2 Trading Volume

Trading volume is usually correlated with investor attention. The number of tweets is the microblogging data variable that is more closely related with investors attention. We tested this relationship for each stock by measuring the regression between the previous day total number of related tweets (n_{t-1}) and today's stock trading volume (v_t).

2.7.3 Volatility

In this study, we assess the contribution of microblogging data to explain volatility using linear regressions. For each stock, we used the following combination of independent variables:

- previous day number of tweets (n_{t-1});
- previous day volatility (σ_{t-1});
- previous day volatility and number of tweets ($\sigma_{t-1} \oplus n_{t-1}$); and
- previous day volatility and number of tweets for the previous two days $t-1$ and $t-2$ ($\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$).

2.8 Computational Environment

All experiments here reported were conducted using the open source **R** tool [14] running on Linux server. We adopted several **R** packages. The Twitter posts were collected using the **RCurl** library, which allows tweets to be fetched under the AOuth[4] secure authorization protocol. The posts were stored using the MongoDB database format by adopting the **rmongodb** package. Some text preprocessing (e.g. tokenization and removal of extra spaces) was performed using the **tm** text mining package [6]. Finally, the R^2 evaluation metric was computed using the **rminer** package [4], while the regression models were computed using the `lm` **R** function.

3. RESULTS

In this section, we present and discuss the results related with the analysis of the relationship between microblogging features and stock market variables.

[4] `http://oauth.net/`

3.1 Returns

Different sentiment indicators using seven lexical resources and the application of two approaches of aggregation of daily indicators (i.e. S1 and S2) are used to infer the regression relationships between daily returns and previous day sentiment data. Tables 1 and 2 present the regression error metric values resulting from the regressions using S1 and S2 methodologies respectively. The column titled **Average** contains the mean of the error metric for all assets and thus is used to assess the overall value of the lexicon and sentiment approach tested. The last two columns are related with a regression model that includes two inputs: the best lexicon based variable (signaled in bold) and the baseline (r_{t-1}). Given that similar results were achieved for both R^2 and RAE, we opted to only show the R^2 metric in Tables 1 and 2. The exception is the last row of each table, which contains the RAE values.

Overall, the baseline method shows an almost null effect in estimating the next day returns, with R^2 values close to zero. Also, there is no added value when joining the baseline input to the best sentiment method result (best $\oplus r_{t-1}$). For a few companies, such as IBM and INTC, the sentiment features seem to have an relevant contribution (e.g. R^2 values of 0.47 and 0.38) for explaining the daily returns. Nevertheless, the overall sentiment results are only slightly better than the baseline, with an average impact of 0.1 points in terms of the R^2 values for most lexicons. When comparing the sentiment methods, the results similar performances for both S1 and S2 strategies. Also, few differences are found between distinct lexicons. MSOL presents the best average result for both S1 and S2. However, the overall R^2 values (0.15 and 0.14) as still low.

For demonstration purposes, Figure 2 shows the quality of the fitted results for the best model (S2 strategy and MSOL lexicon). The model is particularly good at estimating the lowest r_t value.

3.2 Trading Volume

In this subsection, we assess the trading volume regressions. The results are presented in Tables 3 (R^2 values) and 4 (RAE values). Here, the baseline (v_{t-1}) contains some predictive information, with average values of $R^2 = 0.27$ and RAE=0.84. More interestingly, Twitter posting volume data seems quite useful. When used by its own (n_{t-1}), the regression model outperforms in general the baseline for both error metrics (average $R^2 = 0.33$ and RAE=0.85). Moreover, when both inputs are combined ($v_{t-1} \oplus n_{t-1}$), the global results improve (average $R^2 = 0.41$ and RAE=0.81). For some companies, such as AMD, quite interesting modeling results are achieved. Overall, better regression models were obtained when compared with the models fitted for the returns (described in previous subsection).

In Figure 3, we present the quality of the fit for the model with best R^2 value (AMD and $v_{t-1} \oplus n_{t-1}$). The fitted values follow the diagonal line (bottom of Figure 3), suggesting an interesting fit. In particular, the raise of the highest value is correctly fitted by the model.

3.3 Volatility

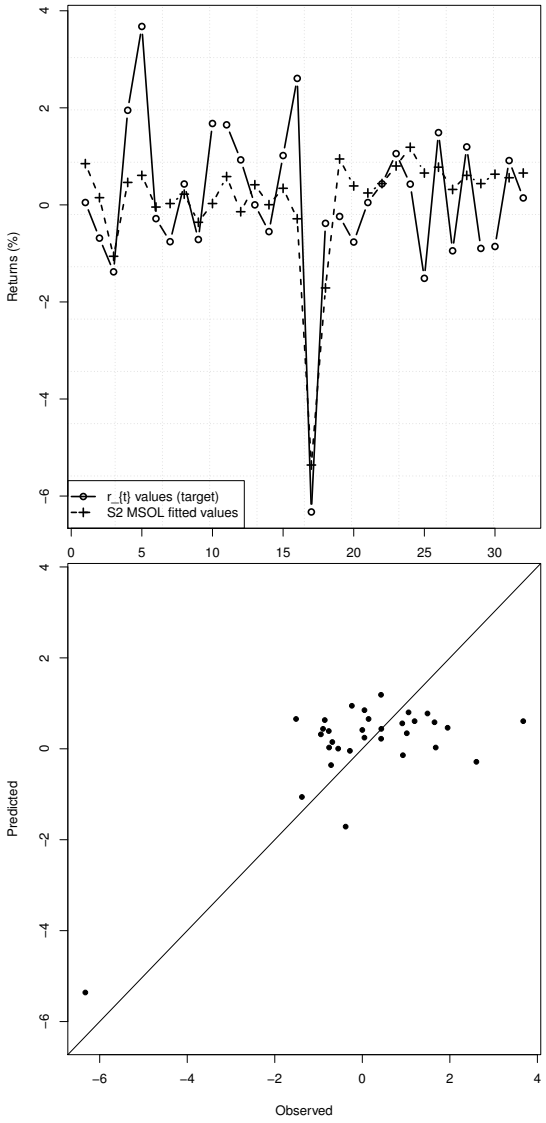

Figure 2: INTC returns (r_t) and fitted values (top, x-axis denotes time in days) and scatter plot of observed versus predictive values (bottom, diagonal line denotes the perfect fit)

The regressions between microblogging data and future volatility was assessed by linear regressions using four different specifications as previously described. Table 5 exhibits the R^2 values of these regressions, while Table 6 presents the RAE errors. Given that, in general, better results were achieved when compared with the trading volume and returns regressions, we highlight the results that are better than the 0.5 threshold, for both R^2 and RAE metrics.

Here, the baseline (σ_{t-1}) is quite informative for fitting the next day volatility, with overall $R^2 = 0.57$ and RAE=0.50. By its own, the Twitter posting volume (n_{t-1}) does not seem useful, with an average $R^2 = 0.04$ and RAE=0.96. However, when combined with the baseline input ($\sigma_{t-1} \oplus n_{t-1}$ and $\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$), there is an increase in the fitted performance. Overall, the best results are achieved by the second

Table 1: Returns using S1 features results (R^2 values except for last row, which includes RAE values, best R^2value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (r_{t-1})	0.01	0.05	0.00	0.01	0.00	0.02	0.00	0.01	**0.04**	0.02
GI	0.03	0.01	0.00	0.02	**0.19**	0.00	0.11	0.16	0.00	0.06
OL	0.02	**0.16**	0.01	0.02	0.08	0.05	0.32	0.10	0.02	0.09
MSOL	**0.20**	0.05	0.02	0.04	0.17	0.01	**0.38**	0.45	0.01	0.15
MPQA	0.06	0.08	0.01	0.02	0.11	**0.08**	0.25	0.33	0.01	0.11
SWN	0.12	0.03	**0.06**	**0.04**	0.03	0.05	0.26	0.40	0.02	0.11
Emoticons	0.11	0.06	0.01	0.00	0.00	0.06	0.03	0.43	0.01	0.08
ALL	0.13	0.04	0.02	0.03	0.16	0.01	0.32	**0.46**	0.01	0.13
best $\oplus$ r_{t-1} (R^2)	0.20	0.16	0.07	0.11	0.21	0.08	0.44	0.46	0.04	0.20
best $\oplus$ r_{t-1} (RAE)	0.91	0.91	0.99	0.93	0.96	0.97	0.82	0.83	0.98	0.92

Table 2: Returns using S2 features results (R^2 values except for last row, which includes RAE values, best R^2value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (r_{t-1})	0.01	0.05	0.00	0.01	0.00	0.02	0.00	0.01	**0.04**	0.02
GI	0.05	0.02	0.00	0.02	**0.20**	0.00	0.10	0.08	0.00	0.05
OL	0.01	**0.15**	**0.03**	0.02	0.08	0.05	0.30	0.06	0.02	0.08
MSOL	**0.24**	0.03	0.00	**0.05**	0.11	0.02	**0.34**	**0.47**	0.01	0.14
MPQA	0.05	0.08	0.00	0.02	0.07	**0.12**	0.25	0.29	0.01	0.10
SWN	0.19	0.06	0.02	0.03	0.06	0.04	0.32	0.37	0.03	0.13
Emoticons	0.11	0.06	0.01	0.00	0.00	0.06	0.03	0.43	0.01	0.08
ALL	0.14	0.06	0.00	0.03	0.13	0.04	0.28	0.38	0.01	0.12
best $\oplus$ r_{t-1} (R^2)	0.25	0.16	0.04	0.06	0.23	0.12	0.44	0.48	0.04	0.20
best $\oplus$ r_{t-1} (RAE)	0.93	0.91	0.98	0.94	0.91	0.93	0.82	0.82	0.98	0.91

Table 3: Volume R^2 results (best value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (v_{t-1})	0.24	0.38	0.07	0.25	0.24	0.40	0.19	0.32	0.30	0.27
n_{t-1}	0.57	**0.48**	0.12	0.39	**0.41**	0.09	0.46	0.28	0.19	0.33
$v_{t-1} \oplus n_{t-1}$	**0.58**	**0.48**	**0.14**	**0.40**	**0.41**	**0.41**	**0.54**	**0.38**	**0.32**	**0.41**

Table 4: Volume RAE results (best value in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (v_{t-1})	0.88	0.86	0.90	0.85	0.85	**0.74**	0.90	0.78	**0.80**	0.84
n_{t-1}	**0.73**	**0.80**	0.87	**0.82**	**0.82**	0.92	**0.88**	0.82	0.98	0.85
$v_{t-1} \oplus n_{t-1}$	0.74	**0.80**	**0.86**	0.84	**0.82**	**0.74**	0.93	**0.76**	0.81	**0.81**

Table 5: Volatility R^2 results (values higher than 0.5 are in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (σ_{t-1})	0.32	0.36	**0.69**	**0.86**	**0.79**	0.12	**0.60**	**0.93**	0.42	**0.57**
n_{t-1}	0.00	0.05	0.00	0.03	0.09	0.01	0.14	0.00	0.07	0.04
$\sigma_{t-1} \oplus n_{t-1}$	0.36	**0.60**	**0.69**	**0.96**	**0.87**	0.12	**0.75**	**0.94**	**0.51**	**0.64**
$\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$	0.37	**0.71**	**0.73**	**0.97**	**0.88**	0.16	**0.77**	**0.94**	**0.53**	**0.67**

Table 6: Volatility RAE results (values lower 0.5 are in bold)

Method	AMD	AMAZN	DELL	EBAY	GOOG	HPQ	IBM	INTC	MSFT	Average
Baseline (σ_{t-1})	0.73	0.63	**0.45**	**0.20**	**0.27**	0.91	**0.44**	**0.21**	0.64	0.50
n_{t-1}	1.00	0.93	1.00	0.97	0.91	1.01	0.88	0.99	0.95	0.96
$\sigma_{t-1} \oplus n_{t-1}$	0.76	0.54	**0.45**	**0.15**	**0.28**	0.92	**0.41**	**0.19**	0.61	**0.48**
$\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$	0.76	**0.49**	**0.44**	**0.13**	**0.27**	0.92	**0.40**	**0.19**	0.59	**0.47**

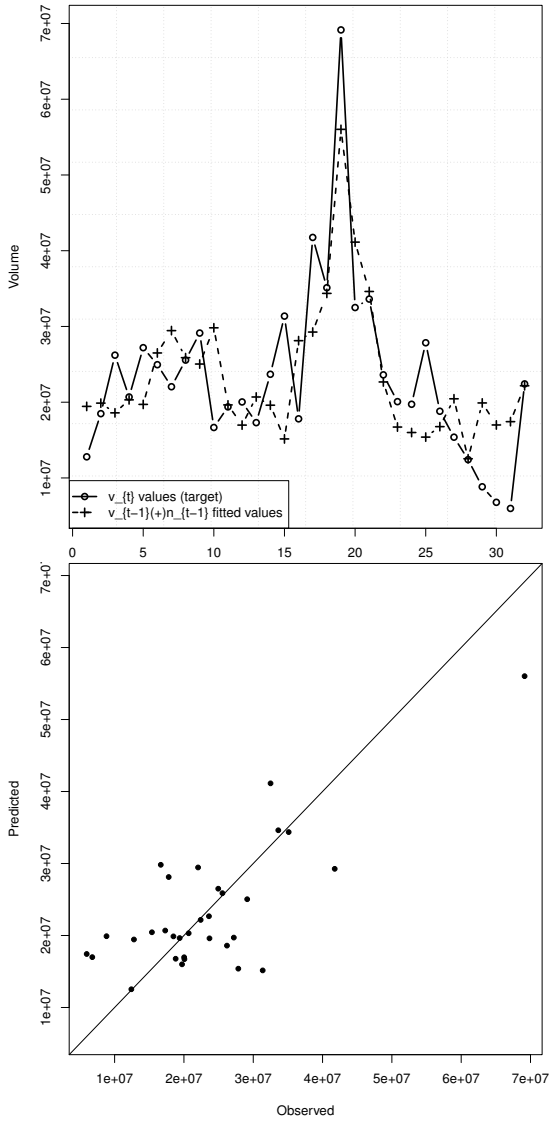

Figure 3: Trading volumes and fitted values for AMD (top, x-axis denotes time in days) and scatter plot of observed versus predictive values (bottom, diagonal line denotes the perfect fit)

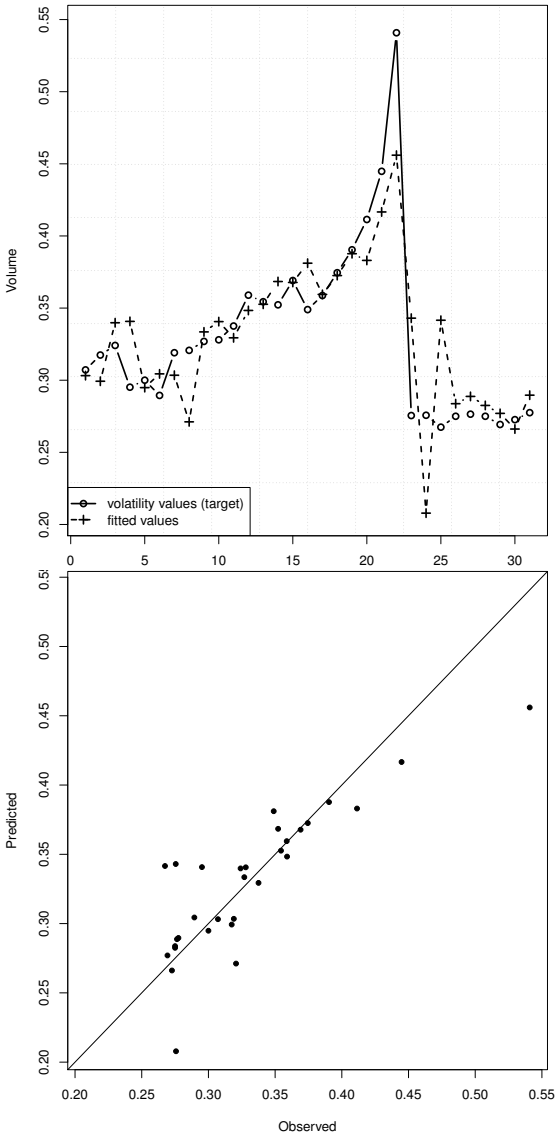

Figure 4: Volatility and fitted values for AMAZN (top, x-axis denotes time in days) and scatter plot of observed versus predictive values (bottom, diagonal line denotes the perfect fit)

combination model ($\sigma_{t-1} \oplus n_{t-1} \oplus n_{t-2}$). For this model, an average of $R^2 = 0.67$ and RAE=0.47 was achieved, meaning that it has a significant predictive capacity for the next day volatility. In effect, the obtained volatility fitting results are of high quality, with several results better than the threshold (AMAZN, DELL, EBAY, GOOG, IBM, INTC and MSFT).

Figure 4 shows the implied volatility and fitted values of best regression model for Amazon (AMZN). We can observe that the use of the lagging values of volatility and number of tweets produces the best fit. In this particular case, the results are interesting not widthstanding the short period analysed. As observed in the figure, the fitted model correctly identifies the raise of the highest peak (at 22 day) and the subsequent fall (at day 23).

4. CONCLUSIONS

The main purpose of this study is to provide a preliminary assessment of the information content of microblogging data for explaining some stock market variables. We focused on very recent Twitter data related to nine large technological companies and tested the modeling ability of this dataset of messages in relation to valuable financial indexes: returns, trading volume and volatility. Two types of data were extracted, sentiment indicators and posting volume. Regarding the former, several indicators were constructed using five popular lexical resources and a two new proposed lexicons: emoticons, based on ":)", ":-)", ":(" and ":-(" terms; and ALL, which merges the six remaining lexicons. We also propose and explore two simple sentiment analysis strategies: S1 - based on the daily number of positive and negative words for a given stock; and S2 - based on the daily classification

of individual tweets, each tweet classified as positive, negative or neutral. Given that we analyzed a recent but small dataset, with 32 days, we opted for simple regression models, under the Occam's razor principle and to avoid overfitting the data.

Confirming the scarce evidence of return predictability [18], the explored sentiment indicators did not, in general, provide significant information about the following day return. However, we found some evidence that Twitter posting volume is relevant for modeling the next day trading volume. Moreover, the same source of data can substantially improve the modeling of volatility, provided it is used in conjunction with the previous day volatility.

The results presented here are promising, showing that information from social networks, in particular microblogging posting volume, can be relevant for modeling the dynamics of useful stock market indicators, such as volume and in particular volatility. However, given the preliminary nature and scope of the study, and the fact that all analysis is performed in-sample, the conclusions need to be analyzed with some caution. While the results are interesting, they merit further research, for a much larger period of time with a more thorough forecasting exercise. Furthermore, the use of more sophisticated parsers and lexicons, more adjusted to stock market terminology, could also improve the investors' sentiment indicators.

5. ACKNOWLEDGMENTS

This work is funded by FEDER, through the program COMPETE and the Portuguese Foundation for Science and Technology (FCT), within the project FCOMP-01-0124-FEDER-022674.

6. REFERENCES

[1] Y. Amihud, H. Mendelson, and L. H. Pedersen. Liquidity and Asset Prices. *Foundations and Trends in Finance*, 1(4):269–364, Aug. 2007.

[2] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Language Resources Association (ELRA)*, 2010.

[3] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[4] P. Cortez. Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In P. Perner, editor, *Advances in Data Mining – Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*, pages 572–583, Berlin, Germany, July 2010. LNAI 6171, Springer.

[5] L. Ederington and W. Guan. Is implied volatility an informationally efficient and effective predictor of future volatility? *Journal of Risk*, 4:29–46, 2002.

[6] I. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5), March 2008.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY, USA, 2nd edition, 2008.

[8] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[9] H. Mao, S. Counts, and J. Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*, 2011.

[10] S. Mohammad, C. Dunne, and B. Dorr. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 599–608. Association for Computational Linguistics, 2009.

[11] J. Nofsinger. Social mood and financial economics. *The Journal of Behavioral Finance*, 6(3):144–160, 2005.

[12] C. Oh and O. Sheng. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *ICIS 2011 Proceedings*, 2011.

[13] R. Peterson. Affect and financial decision-making: How neuroscience can inform market participants. *The Journal of Behavioral Finance*, 8(2):70–78, 2007.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[15] R. Schumaker and H. Chen. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):12, 2009.

[16] T. Sprenger and I. Welpe. Tweets and trades: The information content of stock microblogs. *Social Science Research Network Working Paper Series*, pages 1–89, 2010.

[17] P. Stone, D. Dunphy, and M. Smith. *The General Inquirer: A Computer Approach to Content Analysis*. MIT press, 1966.

[18] A. Timmermann. Elusive return predictability. *International Journal of Forecasting*, 24(1):1–18, Jan. 2008.

[19] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

[20] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.

[21] I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, 2005.