Can Blog Communication Dynamics be correlated with Stock Market Activity?

Munmun De Choudhury Hari Sundaram Ajita John Dorée Duncan Seligmann

Arts Media & Engineering, Arizona State University Collaborative Applications Research, Avaya Labs

Email: {munmun.dechoudhury,hari.sundaram}@asu.edu, {ajita,doree}@avaya.com

ABSTRACT

In this paper, we develop a simple model to study and analyze communication dynamics in the blogosphere and use these dynamics to determine interesting correlations with stock market movement. This work can drive targeted advertising on the web as well as facilitate understanding community evolution in the blogosphere. We describe the communication dynamics by several simple contextual properties of communication, e.g. the number of posts, the number of comments, the length and response time of comments, strength of comments and the different information roles that can be acquired by people (early responders / late trailers, loyals / outliers). We study a "technology-savvy" community called Engadget (http://www.engadget.com). There are two key contributions in this paper: (a) we identify information roles and the contextual properties for four technology companies, and (b) we model them as a regression problem in a Support Vector Machine framework and train the model with stock movements of the companies. It is interestingly observed that the communication activity on the blogosphere has considerable correlations with stock market movement. These correlation measures are further cross-validated against two baseline methods. Our results are promising yielding about 78% accuracy in predicting the magnitude of movement and 87% for the direction of movement.

Keywords

Blogosphere, information roles, communication dynamics, Support Vector Regression, Stock market movement.

1. INTRODUCTION

In this paper, we develop a simple model to study and analyze communication dynamics in the blogosphere and use these dynamics to mine interesting correlations with stock market movement. The problem is important because it provides insights into understanding communication patterns of people; for example, how context affects these patterns; how the information roles of people affect communication, temporal and topical dynamics of information roles etc. The communication dynamics further seem to yield correlations with certain external events as well, justifying their predictive power. Often, these dynamics are useful to corporate organizations who are interested in identifying the 'moods' of people on external communities in response to product releases and company related events. The dynamics can also drive targeted advertising on the web as well as enable understanding community evolution in the blogosphere.

There has been prior work on modeling communication dynamics [5,7,8] and their correlation with external events [4,6]. The authors in [4] determine correlations between activity in Internet message boards (through frequency counts of relevant messages) and stock volatility and trading volume. In [6] the authors attempt to determine if blog data exhibit any recognizable pattern prior to spikes in the ranking of the sales of books on Amazon.com. They present hand-crafted predicates to show that correlation measures indicate visible blog mentions ahead of any evidence in sales rank. There has also been some work on identifying important bloggers based on communication activity in [9]. The authors attempt to determine 'hot' conversations in the blogosphere through agitators and summarizers by establishing discriminants. In [10] the authors identify opinion leaders who are responsible for disseminating important information to the blog network using a variation of the PageRank algorithm. However, modeling information roles and communication dynamics in the prior work have been done in a context-independent manner.

The main contribution of this paper is a simple contextual framework to model the communication dynamics among people and understand how they can be correlated with events external to the blogosphere. In this work, we have specifically looked at the impact on stocks of technology companies due postings in a gadget-discussing blog; however, the framework can be extended easily to other scenarios.

We define stock movement of a company as the normalized difference between the returns on two consecutive days. We assume that the stock movement on a certain weekday can be correlated with the communication dynamics in the past week. This is reasonable because blog communication is often found to precede the occurrence of a real world event [6]. Hence we characterize the communication dynamics in a blog through several contextual features for a particular company. These contextual features are: the number of posts, the number of comments, the length and response time of comments, strength of comments and the different information roles that can be acquired by people (early responders / late trailers, loyals / outliers). We use these features and stock market movement of the company over N weeks for training an SVM regressor. The trained parameters are used to predict the movement at the $(N+1)^{th}$ week. These results are validated using two baseline methods: firstly by comparing with a non-context aware case and secondly using a linear combination of the contextual features. Our technique supersedes both with error of 22 % and 13% in predicting the magnitude and direction of movement respectively.

The rest of the paper is organized as follows. In sections 2 and 3, we present a computational model for information roles and determining contextual features for stock movement prediction. Section 4 discusses the method of determining correlation between communication and stock movements. In section 5, we discuss some experimental results followed by conclusions and future work in section 6.

2. MODELING INFORMATION ROLES

In this section, we describe a set of information roles that can be assumed by people involved in communication in the blogosphere. Roles of people affect their communication and are therefore a part of the social context. We categorize posters of blog comments into the following information roles: early responders / late trailers corresponding to their response time; and loyals / outliers corresponding to the measure of communication activity (frequency count of posts or comments authored).

2.1 Roles due to response behavior

In this section we describe the information roles of people with respect to the response times of their comments (on blog posts).

Intuitively, people who are regularly involved in communication in the blogosphere develop certain structures which characterizes their response behavior or when they would write comments on a certain post. Analysis of such patterns can define the information role of the person over a long period of time. To determine roles due to response time, we define a normalized response time frequency distribution for each poster in the following manner.

Normalized Response Time: Response time of a comment is defined as the time (in seconds) elapsed between the publishing of the original blog and the publishing of the comment. We define the normalized response time of a comment such that it depends on (a) the time at which it was published and (b) the rank of the comment, defined as a metric that depends on its relative position among the set of all comments. If a comment is posted very soon after the blog post, its rank is considered high and vice versa. Our motivation for this definition of normalized response time follows from Figure 1. We consider the two comments shown in green dots. We notice that the 38th comment (first green dot) has a short response time while the 43rd comment (second green dot) has a very long response time. However, we also notice that the difference in their ranks is very low. Hence the two comments have been posted to the blog in a comparatively short span of time. The effective normalized response time thus incorporates the rank metric in order to curb the skewness due to response times.

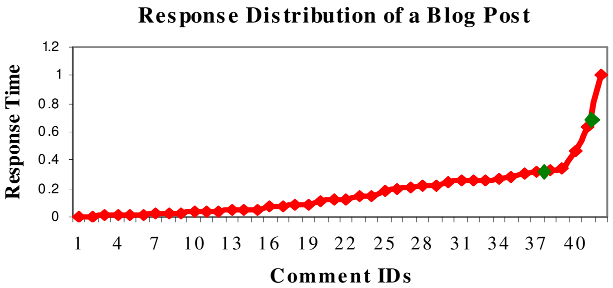


Figure 1: Skewness in normalizing response time.

We define the normalized response time as follows. Let t_m be the time at which a comment was posted by a person Alice to a blog post p_x. Let us further assume that t_s and t_e are, respectively, the publishing time of the post and the last comment. Also, let κ be the rank of Alice's comment. The response time $r^c(x)$ of post p_x is defined as,

$$r^c(x) = \theta_1\left(1 - \frac{(t_m - t_s)}{(t_e - t_s)}\right) + \theta_2\left(1 - \frac{\kappa}{n^c(x)}\right), \qquad < 1 >$$

where θ_1 and θ_2 are two chosen weights and $n^c(x)$ is the number of comments on post p_x. The equation suggests that the normalized response time of the comment is minimized when Alice has responded early and her comment has low rank κ.

Early Responders / Late Trailers: We now define two categories of behavior: early responders and late trailers. *Early Responders* are people who respond to messages or blog posts quickly. *Late Trailers* are people who catch up with an on-going discussion towards the end of communication on the topic. If the mean response time r^c over all comments in a period of time is less than a threshold ρ, then the behavior in that time period is taken to be Early Responder. If it is greater than ρ, then the behavior is defined as a Late Trailer.

2.2 Roles due to measure of activity

In this section we describe two different information roles of people with respect to their overall past communication activity. They are: *loyals* and *outliers*. People who are noticed to author large numbers of comments or posts on a certain topic can be considered as loyals to that topic; while, outliers are all the people who are not characterized by any structure in their communication activity. For example, they are the people who appear to comment on blog posts sporadically.

Assume that a person has written a total number of C comments on all posts in a large time period (say 50 weeks) about a certain company. We construct an activity distribution (Figure 2) for all such people. In order to determine the information role of a particular person using this distribution, we define a suitable threshold θ over the maximum number of comments in the distribution. If C is greater than θ, then the person's role would be a loyal while it will be an outlier if C is less than or equal to θ. For example in the figure, Charles is a loyal while Brian is an outlier with suitably chosen θ.

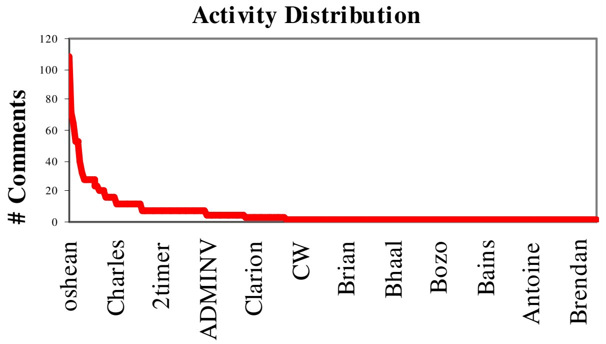


Figure 2: Activity distribution.

3. CONTEXTUAL MODELING OF COMMUNICATION DYNAMICS

In this section we develop several contextual features which characterize the communication dynamics of people. For the set of all features, we assume that the stock movement y_t of a company on a certain weekday depends on the communication activity in the past week (about the same company), in the time period (t – 6) to t. These features are then used to predict stock movement and determine its correlation with communication activity at a future time.

We now introduce some features of our dataset on the Engadget blog. The blog is characterized by two modes of communication: posts and comments written in response to posts. Every comment on a certain post is also marked for its significance by other users. There are five levels: *highest ranked, highly ranked, neutral, low ranked* and *lowest ranked*. It might be noted that these strength levels are different from the 'rank of a comment' described in section 2.1. The level associated with a comment at any instant of time represents the composite significance indicated by all users. Let us now discuss the features.

Number of posts: The higher the number of posts about a certain company, the more impact that particular day has on a future event. Hence the first contextual feature is the number of posts per day in the past week, n^p_{t-6}, n^p_{t-5}, …, n^p_t. Let at day (t – i) where $0 \leq i \leq 6$, p_1, p_2, …, p_{ki} be the k_i posts on a particular company.

Number of comments: The higher the number of comments a certain company, the more impact that particular day has on a future event. Hence the second contextual feature is the number of comments in all the posts per day in the past week, n^c_{t-6}, n^c_{t-5}, …, n^c_t .

$$n^c_{t-i} = \sum_{x=1}^{k_i} n^c_{t-i}(x), \qquad < 2 >$$

where $n^c_{t-i}(x)$ is the number of comments on post p_x, $1 \leq x \leq k_i$ and $0 \leq i \leq 6$.

Length and Normalized response times of comments: The *mean* and *standard deviation* of the length and response times (normalized between 0 and 1) of comments per day might also affect the stock movements. For example, high mean and high standard deviation might reveal that the interest in the posts that day was high, but was peaky or fluctuating. While high mean and low standard deviation would reveal that the interests were high but were mostly flat or consistent among the posters. Let at day (t – i), $l^c_{t-i}(x)$ be the average length of comments on post p_x, where $1 \leq x \leq k_i$ and $0 \leq i \leq 6$. Therefore we define our third contextual feature as the tuple (μ^l_{t-i}, σ^l_{t-i}),

$$\mu^l_{t-i} = \frac{1}{k_i} \sum_{x=1}^{k_i} l^c_{t-i}(x),$$

$$\sigma^l_{t-i} = \frac{1}{k_i} \sqrt{\sum_{x=1}^{k_i} \left(l^c_{t-i}(x) - l^{\bar{c}}_{t-i} \right)^2} \qquad < 3 >$$

Again let at day (t – i), $r^c_{t-i}(x)$ be the average normalized response time of comments on post p_x, where $1 \leq x \leq k_i$ and $0 \leq i \leq 6$. Therefore we define our fourth contextual feature as the tuple (μ^r_{t-i}, σ^r_{t-i}) where the mean and the standard deviation can be computed similar to equation < 3 >.

Strength of comments: The significance of comments on each day of the week is a useful indicator of impact of the comments on external events. If several people have *highly ranked* comments on a certain day, those comments are likely to impact the stock movements of the corresponding company more than less ranked comments. Let R^1_{t-i} be the number of comments that are *highest ranked* in all the posts p_1, p_2, …, p_{ki} at day (t – i) where $0 \leq i \leq 6$. Similarly, we can define the other four sets as R^2_{t-i} , R^3_{t-i} , R^4_{t-i} , R^5_{t-i} respectively. Then the 5th feature is the 5-tuple (R^1_{t-i} , R^2_{t-i} , R^3_{t-i} , R^4_{t-i} , R^5_{t-i}) corresponding to each day (t – i) for all the seven days in the week.

Size of the Early Responder / Late Trailer set: This features takes into account the habitual behavior of the people involved in communication on a certain day (t – i). The sizes of these information roles on a particular day are useful because it contains useful information. If the size of the early responders' set is large, it means that the corresponding posts are likely newer and so the impact on stock movement might be high; while if the size of late trailer set is large, it might reflect an older post whose impact might have already happened or less likely to happen in the future. The 6th feature is therefore the tuple (E, L) where E is the set of all people on day (t – i) who are early responders while L is the set of all late trailers on day (t – i).

Size of the Loyals / Outliers set: We conjecture that the impact of a particular post on the stock movement of a company in the future also depends on *who* is posting comments to that post (based on extent of communication activity in the past). We use the set sizes of the information roles: loyals and outliers that we discussed in the previous section.

Let us consider the activity distribution for all the posts p_1, p_2, …, p_{ki} at day (t – i) where $0 \leq i \leq 6$. Let S^L_{t-i} and S^O_{t-i} be respectively the set of loyals and outliers at day (t – i).

$$S^L_{t-i} = \{x : C(x) > \theta\},$$

$$S^O_{t-i} = \{x : C(x) < \theta\}, \qquad < 4 >$$

where $C(x)$ is the total number of comments written by poster x at day (t – i) where $0 \leq i \leq 6$ and θ is a suitably chosen threshold.

Let us further assume that S^L and S^O be the set of loyals and outliers over the whole training period. There are several interesting implications of these sets. If the cardinality of the set $S^L_{t\text{-i}} \cap S^L$ is large, it means that most of the posters who are otherwise loyal, have responded to posts on day (t – i). It might indicate regularity in communication activity which might further mean that the posts on that day might have low impact on external events. On the other hand, if the cardinality of the set $S^O_{t\text{-i}} \cap [S^L \cup S^O]$ is large, it indicates several outliers posting comments. Such large attention focus from external users might imply the pre-occurrence of a big event. Hence the 8th feature is the tuple ($|S^L_{t\text{-i}} \cap S^L|$, $|\, S^O_{t\text{-i}} \cap [S^L \cup S^O] \,|$) corresponding to each day (t – i).

4. DETERMINING CORRELATION

In this section, we present a Support Vector regression framework to predict the stock movements for a company which would reveal the extent of correlation with the communication dynamics.

First of all we discuss the method of computing stock market movement. In order to determine their correlation with communication dynamics, it is important to take into account the effect of the overall stock market sentiment as well. For example, a negative movement of the stock returns of a particular company may be attributed due to negative movements in the overall stock market index (e.g. NASDAQ, ISE, S&P500 etc).

We define the stock movement of a company c at a day t to be the change in stock return from the closing value of the past day, normalized by the return of the past day. Closing value for a company is the value of stock which exists at the end of the accounting period (one day). The movement is determined as follows,

$$y^{c}_{t} = \frac{(\varphi_t - \varphi_{t-1})}{\varphi_{t-1}},$$

where φ_t is the stock return of the company at day t. Similarly, we determine the overall market movement as,

$$y^{\eta}_{t} = \frac{(\psi_t - \psi_{t-1})}{\psi_{t-1}},$$

where ψ_t is the stock return of the NASDAQ index at day t (NASDAQ because we are focusing on technology companies). Hence the net stock movement for the company is,

$$y_t = y^{c}_{t} - y^{\eta}_{t}\,. \qquad < 5 >$$

Now we present an SVM regression framework to predict stock movement. Let us represent the communication data (comprising the contextual feature vectors) as x_t, $t = 1, 2, \ldots, N$ where N is the number of weeks over the past for a certain company. Also let us assume, the stock movements data be, y_t , $t = 1, 2, \ldots, N$ for the corresponding N weeks for the same company. SVM regression function $f(x)$ is trained on $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. It is tested on the incremental sample (x_{N+1}, y_{N+1}) to get the predicted movement $\hat{y}_{N+1}$. The error in prediction is computed as, $E = (y_{N+1} - \hat{y}_{N+1}) / y_{N+1}$.

5. EXPERIMENTAL RESULTS

In this section we present the experimental results. First we present two baseline frameworks and then describe the nature of the Engadget dataset. This is followed by experimental results.

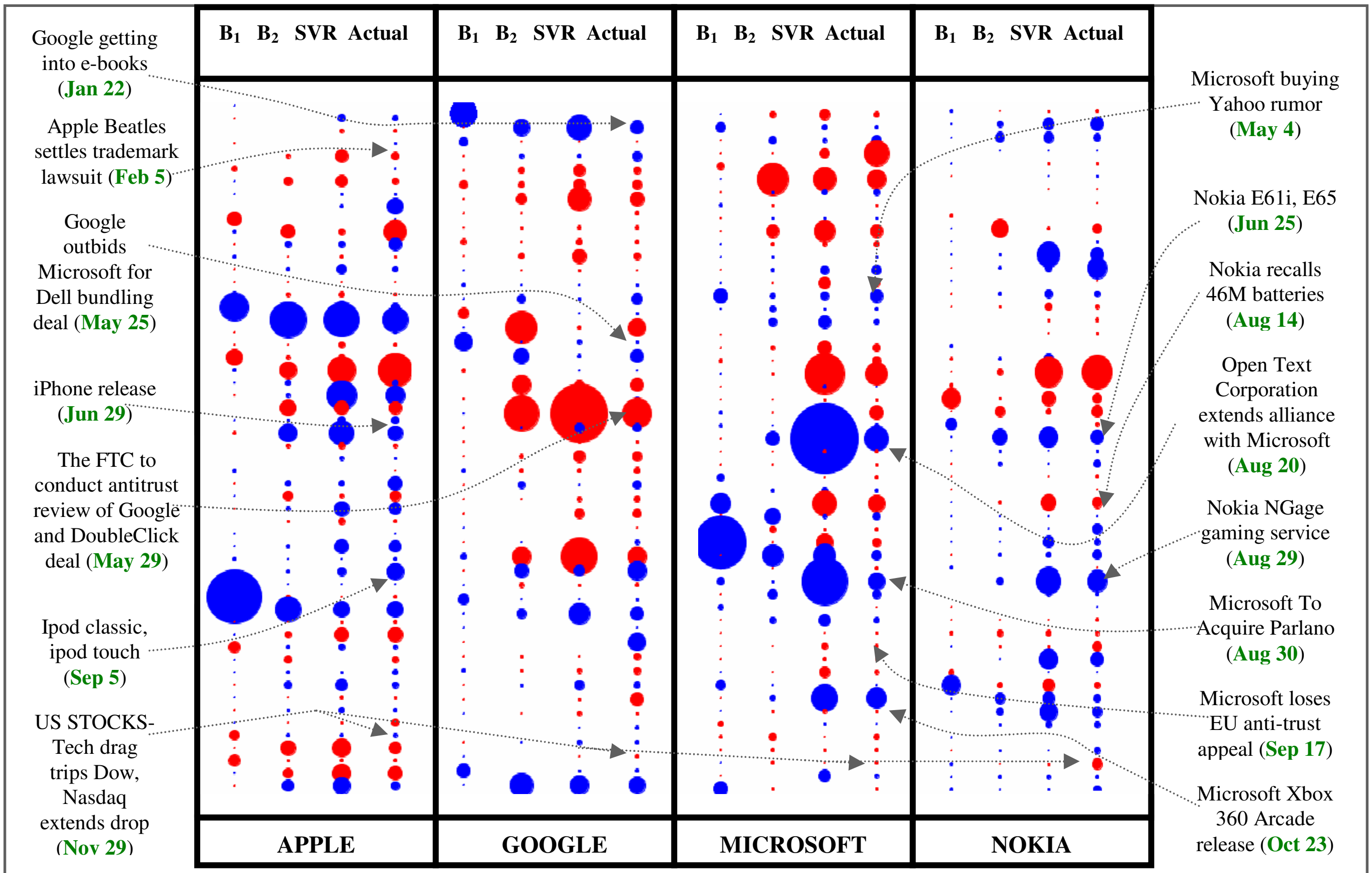
5.1 Baseline Methods

Comment Frequency: The first baseline method for determining correlation of stock market movements uses the frequency of comments per day. We assume again that the stock movement y_t on a certain weekday depends on the number of comments in the past week, in the time period $(t - 6)$ - t. We further use a linear regressor to learn the correlation coefficients incrementally based on stock movements and number of comments.

Linear Relationship among features: In this method, we assume a linear relationship between the contextual features and use a linear regressor to learn the correlation coefficients incrementally. The regression fit is given as, $\mathbf{Y} = \boldsymbol{\alpha} \cdot \mathbf{X}$, where $\boldsymbol{\alpha}$ is the vector of $\alpha_1, \alpha_2, \ldots, \alpha_n$, the correlation coefficients and $\mathbf{Y}$ and $\mathbf{X}$ are the vectors of movements and communication features. Hence for a given training set of n weeks, the regressor predicts the stock movement on a particular weekday at the $(n+1)^{\text{th}}$ week (using the correlation coefficients learnt during training).

5.2 Results of SVM Regressor

In this section we discuss our dataset and the results of prediction using SVR. Our dataset comprises the following data items



Google getting into e-books (Jan 22)

Apple Beatles settles trademark lawsuit (Feb 5)

Google outbids Microsoft for Dell bundling deal (May 25)

iPhone release (Jun 29)

The FTC to conduct antitrust review of Google and DoubleClick deal (May 29)

Ipod classic, ipod touch (Sep 5)

US STOCKS-Tech drag trips Dow, Nasdaq extends drop (Nov 29)

Microsoft buying Yahoo rumor (May 4)

Nokia E61i, E65 (Jun 25)

Nokia recalls 46M batteries (Aug 14)

Open Text Corporation extends alliance with Microsoft (Aug 20)

Nokia NGage gaming service (Aug 29)

Microsoft To Acquire Parlano (Aug 30)

Microsoft loses EU anti-trust appeal (Sep 17)

Microsoft Xbox 360 Arcade release (Oct 23)

Figure 3: Visualization of stock movements with time on vertical scale. B_1 and B_2 are the two baseline techniques discussed in section 5.1. Blue bubbles indicate positive movement and red bubbles negative movements. Sizes of the bubbles represent magnitude of movement. The SVR prediction is found to follow the *movement trend* very closely with an error of 13.41 %.

crawled from the Engadget [1] site: a set of blog posts, corresponding comments, length and strength measures as well as who posted them and at what time. These data were collected for four different companies: Apple, Microsoft, Google and Nokia to capture diversity of patterns. There were a total number of 2,469 blog posts, 41,372 comments and 862 users in the dataset in a time period starting January 2007 to November 2007. The stock market returns for the companies (as per the NASDAQ index) were collected from Google Finance [2].

Now we present the results of the experiments (Figure 3) performed using the two baseline methods and the SVM regression technique. We compare them against actual stock movement for the four companies. In Figure 3, the predicted and actual movements have been shown across time on a vertical scale with blue bubbles indicating positive movement and red bubbles negative movements. The correlations between the communication activity and stock movement have been shown through several representative events. Each of these events has been collected from the New York Times [3] website. For ease of reading and constraints of space, the movements in the figure have been chosen to be representative days in which changes are significant. They span over a span of 50 days and are chosen using suitable thresholds for each company. Hence the same row in Figure 3 might not imply the same day across companies. However, the error has been predicted over all 11 months of training data.

The results of the experiments are revealing. We observe that the two baseline methods are not able to adequately capture the subtleties in variation of stock movements. It is only after the occurrence of a 'big' event that the baseline methods try to compensate for it by showing a large movement later. However, the SVR is able to capture the fluctuations much better. This is because certain discussions on Engadget often occur regarding future events. The SVR technique, being able to capture a wide array of contextual features and also being able to learn their relationships dynamically, follows the actual stock movement better. It is therefore observed that in majority of the cases the SVR method is closest to the magnitude of actual movement (error: 22%) compared to the two baseline methods (error in 1st baseline method: 48%; in 2nd baseline method: 33%). It is interesting to note that the SVR does a better job in following the *movement trend* (correspondence in color of bubbles) with an error of 13.41% (error in 1st baseline method: 36%; in 2nd baseline method: 29%). This might be attributed to the fact that our context aware model can predict the direction of movement very well; but the magnitude of movement is often affected by unprecedented factors, e.g. how the event affects other companies, company statements etc which might not have traces in the past.

It might also be noted here that prediction of stock market movement is an extremely challenging problem which not only depends upon the discussions in a community (e.g. messages exchanged in forums, blogs etc) but are affected by several unforeseen factors which might not be captured in discussions at an earlier point of time. For example, the events "Google outbids Microsoft for Dell bundling deal (May 25)"and "Open Text Corporation extends alliance with Microsoft (Aug 20)" are not found to be present in Engadget discussions in the week long activity used for prediction; this explains the discrepancy and higher error in predicted movement for that day. We also emphasize that this work is attempted to mine interesting correlations of blog communication with stock market activity. The existence of any *causal relationship* between the two remains an open question and is beyond the scope of this paper.

6. CONCLUSIONS

In this paper, we have developed a simple model to study and analyze communication dynamics in blogosphere and use those dynamics to determine interesting correlations with stock market movement. We characterized the communication dynamics in a blog through several contextual features for a particular company. These contextual features were: the number of posts, the number of comments, the length of comments, response time of comments, strength of comments and the different information roles that can be acquired by people (early responders / late trailers, loyals / outliers). We used these features and stock market movement of the company over *N* weeks for training an SVM regressor. We predicted the stock movement using an incremental sample at *N*+1. Our technique supersedes two baseline methods with a mean prediction error of 22 % for magnitude and 13.41% for predicting the direction of movement.

There are several interesting directions to future work. We would like to improve our analysis of the information roles to identify people with variable consequences of their communication activity. The contextual model can also be refined by incorporating clustering of tags of companies, characterizing people by identifying response regions of their comments etc. It might also be interesting to see if there is an implicit macro property that underlies communication dynamics on the blogosphere and if that macro property is accounted for by a vocal minority or majority.

7. REFERENCES

[1] *Engadget* http://www.engadget.com/.

[2] *Google Finance* http://finance.google.com/finance.

[3] *New York Times* http://www.nytimes.com/.

[4] W. Antweiler and M. Z. Frank (2004). *Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.* Journal of Finance **59**(3): 1259-1294.

[5] M. D. Choudhury, H. Sundaram, A. John, et al. (2007). *Contextual Prediction of Communication Flow in Social Networks* Proceedings of the IEEE / ACM / WIC International Conference on Web Intelligence (WI '07). San Jose, CA

[6] D. Gruhl, R. Guha, R. Kumar, et al. (2005). *The predictive power of online chatter* Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining Chicago, Illinois, USA 78-87

[7] R. Kumar, J. Novak, P. Raghavan, et al. (2003). *On the bursty evolution of blogspace*. Proceedings of the 12th international conference on World Wide Web. Budapest, Hungary, ACM**:** 568-576.

[8] B. Li, S. Xu and J. Zhang (2007). *Enhancing clustering blog documents by utilizing author/reader comments*. Proceedings of the 45th annual southeast regional conference. Winston-Salem, North Carolina, ACM**:** 94-99.

[9] S. Nakajima, J. Tatemura, Y. Hara, et al. (2005). *Discovering Important Bloggers based on Analyzing Blog Threads*, Proceedings of WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan,

[10] X. Song, Y. Chi, K. Hino, et al. (2007). *Identifying opinion leaders in the blogosphere,* Proceedings of the sixteenth ACM conference on Conference on information and knowledge management Lisbon, Portugal ACM**:** 971-974.