By Kayvan Shah

Unveiling Market Whispers: Deciphering Blog Communication Dynamics for Stock Predictions

—

Exploring the Synergy between Blogosphere Insights and Stock Market Movements

Background

Objective:

- Develop a model to analyze blogosphere communication dynamics.
- Uncover correlations with stock market movements.

Motivation:

- Understand communication patterns' impact on external events.
- Relevant for corporate insights, targeted advertising, and community evolution.

Contributions:

- Identify information roles and contextual properties for tech companies.
- Model as a regression problem using Support Vector Machine (SVM).

Results:

- SVM outperforms baselines in predicting stock movements.
- Promising correlations - 78% accuracy in magnitude, 87% in direction.

Introduction

- Understanding communication patterns is crucial for insights into how context, information roles, and temporal dynamics affect interactions in the blogosphere.
- The study aims to provide predictive power for corporate organizations interested in gauging public sentiment, especially in response to product releases and company-related events.
- Potential applications include targeted web advertising and understanding community evolution.
- Previous research has explored communication dynamics and correlations with external events, such as stock volatility and book sales spikes on platforms like Amazon.

Introduction

- The paper introduces contextual features for stock movement prediction, including the number of posts, comments, comment length, response time, comment strength, and different information roles.
- The proposed framework uses these features and stock market movement data to train an SVM regressor for predicting future stock movements.
- Validation is done against baseline methods, including a non-context-aware case and a linear combination of contextual features, with the proposed technique outperforming both.
- The main contribution of the paper is a contextual framework to model communication dynamics and correlate them with external events, specifically looking at the impact on technology company stocks based on a gadget-discussing blog.

Modeling Information Roles

Roles Based on Response Behavior

- Focuses on defining information roles of individuals in the blogosphere based on their response times to blog posts.
- The response time is the elapsed time between the publishing of the original blog and the publishing of the comment.

Normalized Response Time:

- It considers the time at which a comment was posted, the publishing time of the post, the last comment, and the rank of the comment among all comments.
- The normalized response time incorporates the rank metric to address skewness due to response times.
- The formula for normalized response time involves weights (θ_1 and θ_2) and considers the number of comments on the post.

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

Definition the normalized response time:

- tm = time at which a comment was posted
- ts = publishing time of the post
- te = the last comment time
- px = a blog post
- κ = rank of comment
- rc(x) = response time
- θ1 and θ2 are chosen weights
- nc(x) = number of comments on post px

Example Illustration

- An example is provided with two comments represented by green dots.
- The 38th comment has a short response time, while the 43rd comment has a very long response time, but the difference in their ranks is low.
- The effective normalized response time incorporates the rank metric to balance the influence of response times.

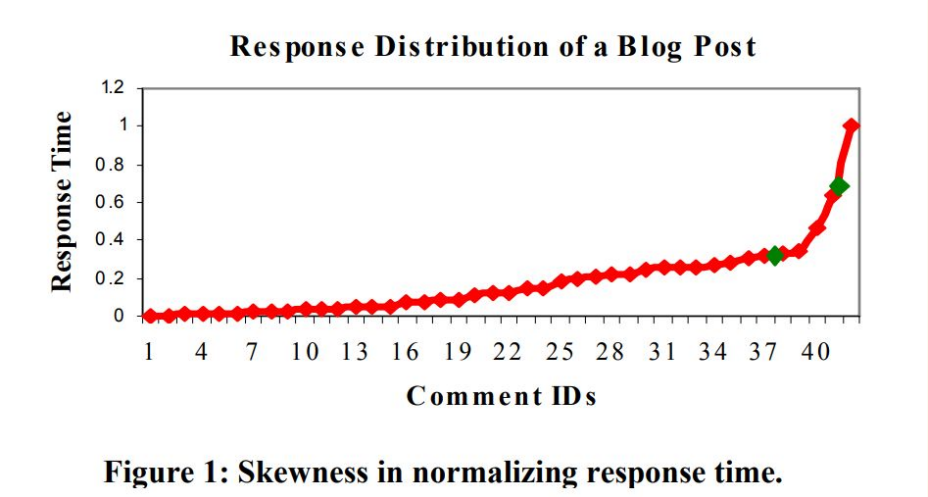

Figure 1: Skewness in normalizing response time.

Behavior Categories Determination

Early Responders / Late Trailers:

- Two categories of behavior are defined based on mean response time over a period:
  - Early Responders: Individuals responding quickly to messages or blog posts.
  - Late Trailers: Individuals catching up with an ongoing discussion towards the end of communication on a topic.
- The determination is made by comparing the mean response time to a threshold ρ.
- Early Responder or Late Trailer behavior is determined by comparing the mean response time over a period to a threshold ρ.
- If mean response time is less than ρ, it is classified as Early Responder behavior; if greater, it is classified as Late Trailer behavior.

Roles Based on Measure of Activity

This section introduces two information roles based on individuals' overall past communication activity:

1. **Loyals**: Individuals who author large numbers of comments or posts on a specific topic, indicating a consistent and dedicated engagement with that topic.
2. **Outliers**: Individuals lacking a structured communication pattern, commenting sporadically without a clear commitment to a specific topic.

Activity Distribution:

- Assume a person has written a total of C comments on all posts about a specific company over an extended period (e.g., 50 weeks).
- An activity distribution is constructed for all individuals based on their total comments.
- Threshold Definition: A threshold θ is established over the maximum number of comments in the distribution.
- If an individual's total comments (C) exceed θ, they are categorized as loyals; if C is less than or equal to θ, they are considered outliers.

Illustrative Example

For example - representing individuals and their total comments.
A threshold θ is chosen, and individuals like Charles, exceeding this threshold, are identified as loyals, while those like Brian, falling below the threshold, are labeled as outliers.

Role Assignment:

- The person's role is determined based on a simple comparison with the threshold θ.
- Loyal Assignment: If the total comments exceed the threshold ($C > \theta$).
- Outlier Assignment: If the total comments are less than or equal to the threshold ($C \leq \theta$).

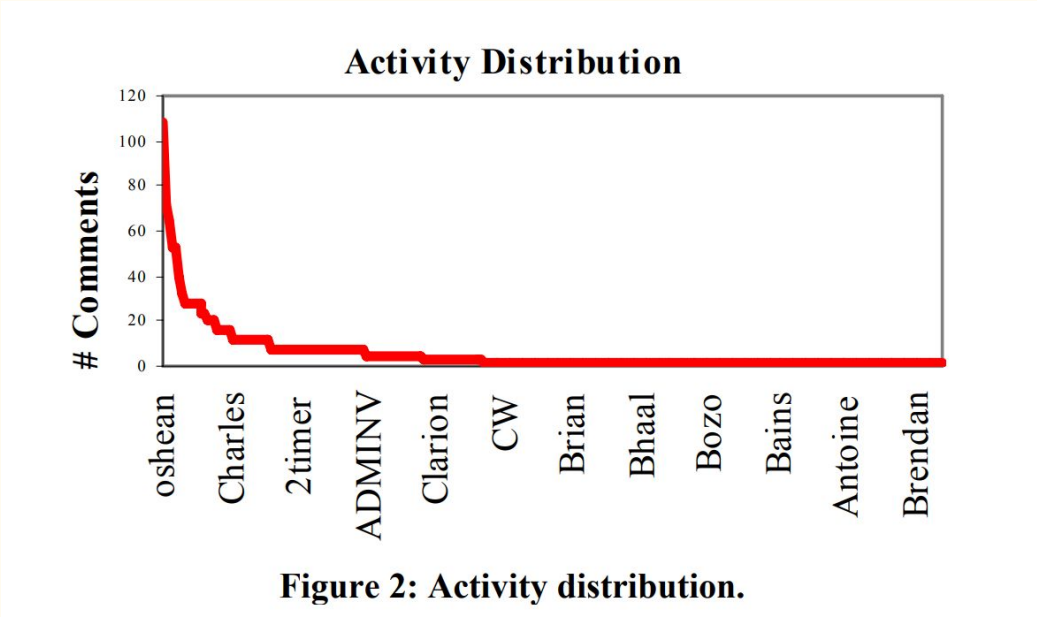

Figure 2: Activity distribution.

Contextual Modeling of Communication Dynamics

Feature Description

Number of Posts:
- Count of posts about a specific company on a given day.
- Indicates impact on future events.

Number of Comments:
- Total count of comments across all posts about a company on a given day.
- Reflects the impact of comments on future events.

Length and Response Times of Comments:
- Average length and normalized response times of comments.
- Insight into engagement nature – peaky, fluctuating, flat, or consistent.

Strength of Comments:
- Count of comments ranked highest to lowest.
- Indicates significance on external events.

Size of Early Responder / Late Trailer Set:
- Tuple representing sizes of early responders and late trailers.
- Considers habitual behavior for potential impact.

Size of Loyals / Outliers Set:
- Tuple representing sizes of loyals and outliers.
- Impact depends on communication activity - loyalty vs. lack of structure.

Activity Distribution Thresholds:
- Define thresholds for loyals and outliers based on total comments by an individual.
- Classifies based on communication activity level.

Set Intersections for Loyals and Outliers:
- Tuples representing intersections of sets of loyals/outliers at specific intervals.
- Analyzes regularity in communication and identifies potential impact on external events.

Determining Correlation

Support Vector Regression Framework for Stock Movement Prediction

The framework integrates SVM regression to predict stock movements based on contextual feature vectors derived from communication data. The computed error assesses the accuracy of the predictions.

Stock Movement Calculation:
- *Objective*: Reveal correlation with communication dynamics by predicting stock movements.
- *Considerations*: Account for overall stock market sentiment impact on individual stock returns.

Stock Movement Definition:
- *Formula*: Where ϕ_t is the stock return of company c at day t.
- *Normalization*: Closing value represents the day's end stock value.

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

Compute Overall Market Movement:
- Formula:Where ψ_t is the stock return of NASDAQ index at day t.
- *Focus*: Technology companies, hence NASDAQ index is considered.

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

Compute Net Stock Movement:

$$y_t = y^c{}_t - y^\eta{}_t$$

SVM Regression Framework:
- *Objective*: Predict stock movement using SVM regression.
- *Data*: Represent communication as x_t and stock movement as y_t.
- *Training*: Train SVM regression function f(x) on past data.
- *Testing*: Predict stock movement for future data.

$$\{(x_1, y_1), \ldots, (x_N, y_N)\}$$

Error Computation:

$$E = (y_{N+1} - \hat{y}_{N+1}) \,/\, y_{N+1}$$

Machine Learning Pipeline

Experimental Results

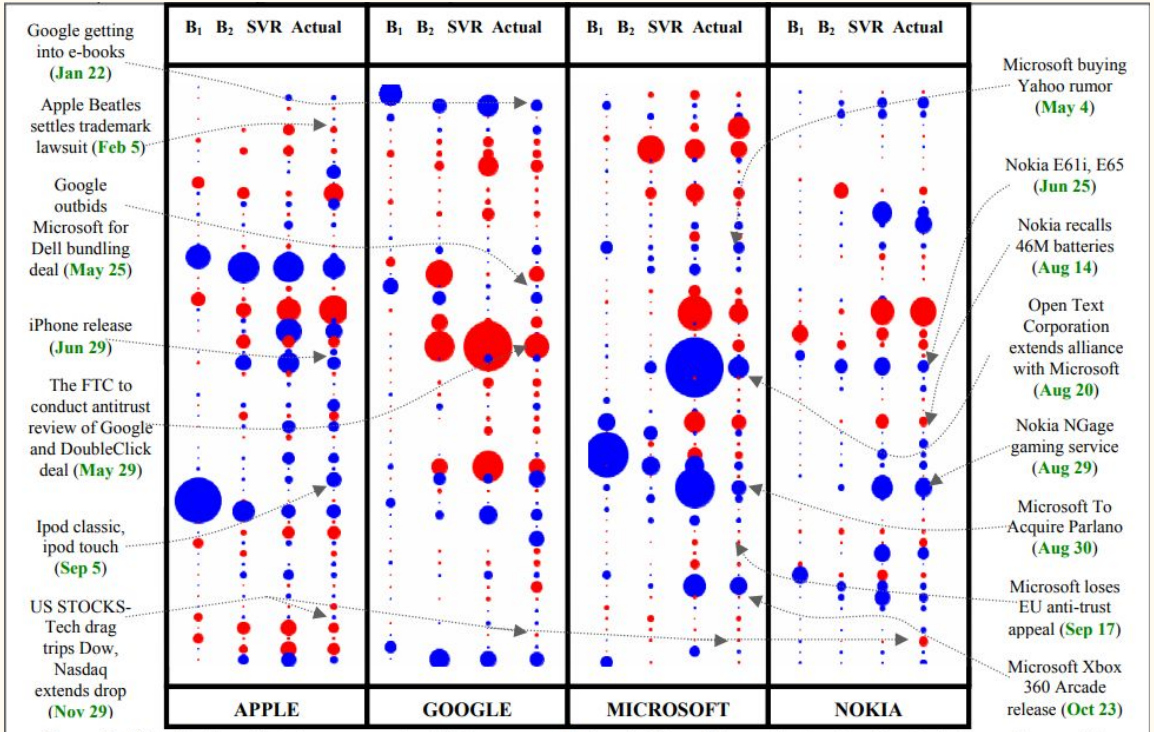

Figure 3: Visualization of stock movements with time on vertical scale. B_1 and B_2 are the two baseline techniques discussed in section 5.1. Blue bubbles indicate positive movement and red bubbles negative movements. Sizes of the bubbles represent magnitude of movement. The SVR prediction is found to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

Visualization of Stock Movement

Baseline Methods for Correlation Assessment

Baseline 1: Comment Frequency

- Approach: Evaluate correlation based on comment frequency per day.
- Assumption: Stock movement y_t depends on the number of comments in the past week (t−6 to t).

Baseline 2: Linear Relationship among Features

- Approach: Assume a linear relationship between contextual features.
- Regressor: Utilize a linear regressor to learn correlation coefficients incrementally.
- Formula: $Y = \alpha \cdot X$, where
  - α is the vector of correlation coefficients $(\alpha_1, \alpha_2, \ldots, \alpha_n)$
  - Y and X are vectors representing movements and communication features.

These baseline methods serve as comparison approaches for assessing the correlation of stock market movements with communication dynamics

Dataset Overview

Data from Engadget site:

- Blog posts, comments, length, strength measures, user details.

Companies:

- Apple, Microsoft, Google, Nokia.

Dataset Size:

- 2,469 blog posts, 41,372 comments, 862 users (Jan 2007 - Nov 2007).

Stock Market Returns:

- Collected from Google Finance (NASDAQ index).

Results & Observations

Experiment Results (Slide 16):

- Predictions using SVM regression, compared with two baseline methods.
- Vertical scale shows predicted and actual movements with blue and red bubbles.
- Correlations illustrated with representative events from the New York Times.

Observations:

- Baseline methods struggle to capture subtle variations in stock movements.
- SVR excels in capturing fluctuations, closely matching actual movement magnitudes (error: 22%).
- SVR better follows movement trends, showing a color correspondence (error: 13.41%).

Challenges and Considerations:

- Stock market prediction is complex, influenced by unforeseen factors and community discussions.
- Context-aware model enhances trend prediction, while magnitude is affected by unprecedented factors.

Key Takeaways

Conclusion

- Communication dynamics in a blog are characterized using various contextual features for a specific company, including the number of posts, comments, comment length, response time, comment strength, and different information roles (early responders/late trailers, loyals/outliers).
- An SVM regressor is trained using these features and stock market movement data over N weeks, with predictions made for the stock movement at N+1.
- The proposed technique outperforms two baseline methods, with a mean prediction error of 22% for magnitude and 13.41% for predicting the direction of movement.

Future Directions

- Improvement in the analysis of information roles to identify individuals with variable consequences of their communication activity.
- Refinement of the contextual model by incorporating clustering of company tags, characterizing individuals based on response regions of their comments, etc.
- Exploration of implicit macro properties underlying communication dynamics in the blogosphere and investigating whether these properties are influenced by a vocal minority or majority.

Considering the results, what inference can be drawn about the effectiveness of context-aware models in predicting stock movements?

1. Context-aware models are ineffective for stock predictions.
2. Baseline methods outperform context-aware models.
3. Context-aware models, especially SVM, perform well in capturing stock fluctuations.
4. Context-aware models only work well for short-term predictions.

3

Context-aware models, especially SVM, perform well in capturing stock fluctuations.

THANK YOU