Exploiting Social Relations and Sentiment for Stock Prediction

Jianfeng Si∗, Arjun Mukherjee†, Bing Liu‡, Sinno Jialin Pan‡, Qing Li‡, Huaiyi Li†

∗ Institute for Infocomm Research, Singapore
† Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
‡ Department of Computer Science, City University of Hong Kong

Abstract

• The Web has seen a tremendous rise in social media.
• Information in social media text (e.g., Twitter, Facebook ) not only contains Opinion, but also Relations.
• The goal of this paper is to exploit social relations and social sentiment for stock market prediction.
• We build a Semantic Stock Network (SSN) from the co-occurrence statistics of cash-tags in Twitter messages. This SSN summarized discussion topics about stocks and stock relations.
• Experimental results demonstrate that topic sentiments from close neighbours are able to help improve the prediction of a stock markedly.

Key Tasks

• Data collection.
• Build the stock network.
• Derive the topics over nodes and edges.
• Regress stock price with sentiment time-series derived from the network in an autoregressive framework for market prediction.

Data Collection

• Collected streaming tweets using Twitter’s REST API.
• Query keywords: ticker symbols from S&P100 stocks.
• $AAPL$, $GOOG$, SAMZ, SMSFT…
• “$AAPL$ is loosing customers. everybody is buying android phones! $GOOG.”

Tweets in Relation to the Stock Market

• We define the stock network as an undirected graph: $G = (V, E)$.
• $V$ comprises stocks.
• $e_{u,v} \in E$ stands for the edge between stock nodes $u$ and $v$.
• For a tweet, $d$ with three cash-tags: $\{v_1, v_2, v_3\}$, we annotate $d$ with the label set as $L_d = \{v_1, v_2, v_3, e_{1,2}, e_{1,3}, e_{2,3}\}$
• E.g., $e_{1,2}$ is “AAPL_goog” if $v_1$ is “AAPL”, and $v_2$ is “goog”.
• Further apply the Labeled-LDA on this labeled data set.

The Semantic Stock Network (SSN)

• The stock network is $V = \{v_1, v_2, v_3\}$.
• $e_{1,2}$ is the edge that connect $v_1$ and $v_2$.

Stock Market Prediction

• Two-dimensional $\{x_t, y_t\}$ vector autoregression model (VAR)
• Regress $y_t$ on $x_t$ using least square regression in R.
• $y_t = \sum_{i=1}^T (\theta_i x_{t-i}) \ast (\theta_{i} y_{t-i}) + \epsilon_t$
• Experiment with different window sizes and lags.
• Evaluate prediction accuracy of Price ($1/1$) movement.

Figure 3. An example stock network.

Figure 4. Tweet label design.

Figure 5. Prediction on Sapple on lag 2. (x-axis is training window size, y-axis is the accuracy.)

<table>
<thead>
<tr>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>$aapl$</td>
</tr>
<tr>
<td>$goog$</td>
</tr>
<tr>
<td>$dis$</td>
</tr>
<tr>
<td>$csco$</td>
</tr>
<tr>
<td>$amzn$</td>
</tr>
<tr>
<td>$ebay$</td>
</tr>
<tr>
<td>$tgt$</td>
</tr>
<tr>
<td>$wmt$</td>
</tr>
<tr>
<td>$qcom$</td>
</tr>
<tr>
<td>$intc$</td>
</tr>
</tbody>
</table>

Table 1. Average and best (in parentheses) prediction accuracies (over window sizes of {15, 60}) of some other cases with different covariates, cell of $dis(0.96)$ means “$dis$” takes the maximum price correlation strength of 0.96 with “$goog$” (similar for others in column CSN). The best performances are highlighted in bold.

Conclusion & Future Work

• SSN is robust to find stock pairs with real-world relationship.
• Sentiment based approaches perform better than all price based ones. Furthermore, sentiment of the neighbors in SSN performs best in general.
• The business of offline companies like Target Corp. ($tgt$) and Wal-Mart Stores Inc. ($wmt$) are highly affected by online business like $amzn$.
• Future Work:
  • Fully exploit the network power.
  • Connect social media text to financial reports.