On the Importance of Text Analysis for Stock Price Prediction

Heeyoung Lee (Stanford),
Mihai Surdeanu (University of Arizona),
Bill MacCartney (Google),
Dan Jurafsky (Stanford)
Motivation

• Many people are interested in the financial market
• A vast amount of text information online
• Hard to monitor in real time for individual investors
Financial reports list events that impact company performance

- “8-K” financial reports for companies
- US companies required to file an 8-K upon events like layoffs, mergers, officer changes

On November 15, 2011, the Board of Directors (the “Board”) of Apple Inc. (the “Company”) appointed Robert A. Iger to the Board. Mr. Iger will serve on the Audit and Finance Committee of the Board.
The problem

• Can we predict stock price movement after the 8-K report is released?
  • Since 8-K generally released after market close, can we predict what happens when market opens?
• Modeled as a three-class classification problem:
  **UP:** $\Delta > +1\%$
  **DOWN:** $\Delta < -1\%$
  **STAY:** $-1\% < \Delta < +1\%$
• Various controls and normalizations (normalize by change in S&P 500 index)
Our question: How is the text of 8-K reports linked with stock movements?

- Does language in 8-K reports offer additional information?
  - What kind of information?

- How long does the effect persist in the marketplace?
Related Work

- Xie et al. (2013): Tree representations of information in news
- Bollen et al. (2010): Twitter mood for market movement
- Bar-Haim et al. (2011): Identifying better expert investors
- Leinweber and Sisk (2011): Effect of news and the time needed to process the news in event-driven trading
- Kogan et al. (2009): A method that predicts risk based on financial reports
- Engelberg (2008): Linguistic information has more long-term predictability

- Only few previous results show improvements from textual information on predicting the impact of financial events on top of quantitative features
Corpus

• 8-K financial reports for S&P 500 companies
  • Between 2002-2012 (2009-2010 dev, 2011-2012 test)
  • 13,671 documents, 28M words
  • 8-K reports are mostly released before the market open or after the market closed

• Financial Annotations:
  • Daily history of stock prices
    • The (normalized) difference in the company’s stock price before and after the report is released
  • Earnings Per Share (EPS)
    • Reported EPS (from 8-K reports)
  • Consensus EPS (the estimation of analysts)
### Baseline Financial Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Earnings surprise</strong></td>
<td>The gap between the actual and expected earnings per share</td>
</tr>
<tr>
<td><strong>Recent stock price changes</strong></td>
<td>1 week, 1 month, 1 quarter, 1 year</td>
</tr>
<tr>
<td><strong>Volatility S&amp;P 500 index</strong></td>
<td>Roughly represents the expected movement of S&amp;P 500 index over the following 30 days</td>
</tr>
<tr>
<td><strong>Event category features</strong></td>
<td>Event types in 8-K reports (merger, bankruptcy, ...)</td>
</tr>
</tbody>
</table>
Linguistic Features

• Unigram features

... On a GAAP basis, the Company reported a net loss of $356 million ... We remain intensely focused on helping ...

{loss:1, basis:1, remain:2 ... }
Linguistic Features

• Unigrams are too sparse!

\{loss:1, basis:1, remain:2, dimension:0, clandestine:0, ... \}

• Projected textual features into a smaller-dimension latent space by non-negative matrix factorization
  • Unigrams are projected into 50, 100, 200 dimension vectors
  • Benefits: captures latent meaning, faster training
Linguistic Features

loss

{1, 0, 0, ...} \( \in \mathbb{R}^n \)

{0.11, -0.23, 1.76, ...} \( \in \mathbb{R}^n \)

basis

{0, 1, 0, ...} \( \in \mathbb{R}^n \)

{-0.34, 0.61, -0.7, ...} \( \in \mathbb{R}^n \)

\( n = 50, 100, 200 \)

{loss:1, basis: 1, remain:2 ...} \( \in \mathbb{R}^n \)

{1, 1, 2, ...} \( \in \mathbb{R}^n \)

{0.2391, 0.7293, -1.2301, ...} \( \in \mathbb{R}^n \)
Example

Visa inc 8-K report released after Oct 29 market close

Visa inc $\Delta$ -0.2%

S&P 500 $\Delta$ +1%

Normalized $\Delta$

Oct 29 Close Oct 30 Open

Label: DOWN

Oct 29 Close Oct 30 Open

-1.2%
Example

- **Earnings surprise** = \((0.58-0.56)/0.56*100 = 3.57\%\)
- **Event type**: Results of Operations and Financial Condition

*From 8-K report*

. . . On a GAAP basis, the Company reported a net loss of $356 million . . . We remain intensely focused on helping ...

- **Unigrams**: \{loss: 1, basis: 1, ...\}
- **NMF vectorized**: \{0.2391, 0.7293, -1.2301, ... \}
Classifier

- Random forest with 2000 trees
- Tuned on the development set (The size of random subset of features when training)
- PMI based feature selection
  - Among top 5000 features, keep features occurred at least 10 times throughout the training data -> 2319 unigram features
# Five Models

- Financial baseline 1
- Financial baseline 2
- Explicit text
- Text projected into latent space
- Everything (ensemble)

<table>
<thead>
<tr>
<th>Feature</th>
<th>B1</th>
<th>B2</th>
<th>Uni</th>
<th>NMF</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earnings surprise</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Recent movements</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Volatility index</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Event category</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Unigrams</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NMF vector</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random guess</td>
<td>33.3</td>
</tr>
<tr>
<td>Majority class</td>
<td>34.9</td>
</tr>
<tr>
<td>Baseline1</td>
<td>49.4</td>
</tr>
<tr>
<td>Baseline2</td>
<td>50.1</td>
</tr>
<tr>
<td>Unigram model</td>
<td>54.4</td>
</tr>
<tr>
<td>NMF 50</td>
<td>54.7</td>
</tr>
<tr>
<td>NMF 100</td>
<td>55.4</td>
</tr>
<tr>
<td>NMF 200</td>
<td>55.3</td>
</tr>
<tr>
<td>Ensemble</td>
<td>55.5</td>
</tr>
</tbody>
</table>

- Financial baseline 1: using only earnings surprise
- Financial baseline 2: using numeric & event category features to see financial baseline
- Ensemble: combined unigram and 3 NMF models
Temporal Aspect Model

- Investigates the predictive power of text as we move farther away from the event
- Same as before, it predicts UP/DOWN/STAY
Results:
Temporal Aspect Model

* BE: before event

- The effect of linguistic features diminishes quickly
Results:
Temporal Aspect Model

- Engelberg’s (2008):
  - Soft information (textual information) requires a higher processing cost and takes longer to affect the market
Results: Temporal Aspect Model

• Why the opposite result?

• Our data: company report (short term)
• Their data: news article (longer term)

• News articles reflect not only new information, but also the perspectives and opinions of third parties
• Hypothesis: the market is highly sensitive to company reports in the short term, but more sensitive to third party perspectives in the longer term
Positive & Negative words from Unigrams

**Positive**: Increase, growth, new, strong, forward, well, grow, product, future, we

**Negative**: charge, loss, lower, decline, reduce, down, adjust, regulation, offset, reduction, while
Error Analysis

- Correct label: DOWN
- Unigram: DOWN, baseline2: UP

... despite lower than expected growth, the more rapid decline in demand for FedEx Express services, particularly from Asia, outpaced our ability to reduce operating costs.

... 

- lower, decline, and reduce are in our top features
Error Analysis (2)

• Correct label: DOWN
• Unigram: STAY, baseline2: DOWN

"We are very pleased with our fourth quarter and full year results, as well as ...

• Unigram model incorrectly predicts STAY even though the company has unsatisfactory earnings per share
Example Decision Path

- Abbott Laboratories
- Apr 15, 2009 report
- True label: DOWN
- Decision: DOWN

... This resulted in a $230 million **decline** in Depakote sales in the first quarter ...

Earning surprise

... increase

decline

... 1 Quarter Δ

... DOWN
Negative Results

- Sentiment features
  - SentiWordNet: domain difference. e.g., *growth* is objective (or slightly negative)
  - Sentiment lexicons from Jegadeesh and Wu (2013): The list is small and 77% of lexicons are already in our feature set.

- Bigrams and Word clustering features
- Other classifiers (logistic regression or multilayer perceptron)
Limitations

• Do not trade using this yet 😊
• Predicts movement but not magnitude (how big the change will be)
• It ignores several important real-world factors:
  • Transaction costs: bid-ask spreads
  • Slippage: the tendency of large trading programs to move the market
  • Borrowing costs for short positions
• Does not parse numbers and their meaning
Conclusions

• New corpus of 8-K financial reports, annotated with financial metrics
• Text helps to predict stock price movement,
  • short-term movement is sensitive to events in company reports
  • longer-term movement may be sensitive to third-party perspectives in news
• Domain is important when using sentiment lexicons
  • Using all unigrams was better than using published lexicons
• Methodology
  • Dimensionality reduction of word vectors helps
  • Random forests worked well to combine lexical and numeric features