Financial Sentiment Analysis for Risk Prediction

Tse Liu
Department of Computer Science
National Chengchi University

Joint work with Prof. Chuan-Ju Wang,
Prof. Ming-Feng Tsai and Chin-Ting Chang

IJCNLNP 2013, October 16, 2013
Outline

1. Introduction
2. Methodology
3. Experiments
4. Conclusion
Introduction

- Financial field: Predict risk by GARCH model.\(^1\)

- Kogan used the bag-of-words model to bring the text information into prediction.\(^2\)

- Sentiment analysis is the task of finding the attitudes of authors about specific objects.

- In finance, the sentiments can be used to reflect the correlations with other financial measures, such as stock returns and volatilities.

---


\(^2\)Kogan et al. (2009). Predicting risk from financial reports with regression. *In NAACL '09.*
This paper attempts to use the finance-specific sentiment lexicon to model the relations between sentiment information and financial risk.

- Predict target: Financial risk (stock return volatility).
- Features: Text information, financial information (stock return volatility).

We formulate the problem as Regression and Ranking prediction tasks:

- Regression predict target: Volatility of companies.
- Ranking predict target: Relative risk level of companies.
Methodology

Risk Proxy: Stock Return Volatility

- Stock Return
  
  \[
  \text{Total Stock Return} = \frac{(S_1 - S_0)}{S_0}
  \]

- Volatility
  
  - In finance, volatility is a common risk metric measured by the standard deviation of a stock returns over a period of time.

- Stock Return Volatility
  
  - Let \( S_t \) be the price of a stock at time \( t \).

  \[
  V_{[t-n,t]} = \sqrt{\frac{\sum_{i=t-n}^{t} (R_i - \overline{R})^2}{n}}, \text{ where } \overline{R} = \sum_{i=t-n}^{t} \frac{R_i}{(n+1)}.
  \]
For most sentiment analysis algorithms, the sentiment lexicon is the most important resource.\(^3\)

The words have different meaning between finance lexicon and general-purpose lexicon.

\(^3\)Feldman. (2013), Techniques and applications for sentiment analysis. *Communications of the ACM*
### Six Finance-Specific Lexicons

<table>
<thead>
<tr>
<th>Class</th>
<th>Meaning</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fin-Neg</strong></td>
<td>Negative business terminologies</td>
<td>deficit, delist</td>
</tr>
<tr>
<td><strong>Fin-Pos</strong></td>
<td>Positive business terminologies</td>
<td>profit, integr</td>
</tr>
<tr>
<td><strong>Fin-Unc</strong></td>
<td>Words denoting uncertainty</td>
<td>doubt</td>
</tr>
<tr>
<td><strong>Fin-Lit</strong></td>
<td>Propensity for legal contest</td>
<td>amend, forbear</td>
</tr>
<tr>
<td><strong>MW-Strong</strong></td>
<td>Strong levels of confidence</td>
<td>must, best</td>
</tr>
<tr>
<td><strong>MW-Weak</strong></td>
<td>Weak levels of confidence</td>
<td>may, perhaps</td>
</tr>
</tbody>
</table>

---

Methodology

Problem Formulation

- Predict target: Stock return volatility.

- Features
  - Text information: Financial sentiment words (finance-specific lexicons).
  - Financial information: The twelve months before the report volatility for each company.

- Predict target: Financial risk (stock return volatility).
Methodology

Regression and Ranking

Regression:

\[
\min_{w} V(w) = \frac{1}{2} < w, w > + \frac{C}{n} \sum_{i=1}^{n} \max(|v_i - f(d_i; w)| - \epsilon, 0)
\]

Ranking:

Ranking solves the same optimization problem as regression, but the difference is that ranking focuses on the pair-wised ranking orders.
Corpora and Dictionary

- **The 10-K Corpus**

- **Six Finance-Specific Lexicons**
  - Fin-Neg
  - Fin-Pos
  - Fin-Unc
  - Fin-Lit
  - MW-Strong
  - MW-Weak
## Statistics of the Financial Lexicon

<table>
<thead>
<tr>
<th>Dictionary</th>
<th># of Words</th>
<th># of Stemmed Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fin-Neg</td>
<td>2,349</td>
<td>918</td>
</tr>
<tr>
<td>Fin-Pos</td>
<td>354</td>
<td>151</td>
</tr>
<tr>
<td>Fin-Unc</td>
<td>291</td>
<td>127</td>
</tr>
<tr>
<td>Fin-Lit</td>
<td>871</td>
<td>443</td>
</tr>
<tr>
<td>MW-Strong</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>MW-Weak</td>
<td>27</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,911</strong></td>
<td><strong>1,664</strong></td>
</tr>
</tbody>
</table>
We use the **TFIDF**, **LOG1P** to represent the text information of documents.

\[
TFIDF(t, d) = TF(t, d) \times IDF(t, d) = \frac{TC(t, d)}{|d| \times \log(|D|/\sum_{d \in D: t \in d})}
\]

\[
LOG1P = \log(1 + TC(t, d))
\]

In addition to the finance-specific lexicon, we add the twelve months before the report volatility for each company.
Experiments

Experimental Setting

- We use every 5 years historical financial reports to train the models.
  - The trained models are tested by the following year.
- Example:
  - Test set: The 2001 financial reports.
## Corpora statistic

<table>
<thead>
<tr>
<th>Year</th>
<th>Words</th>
<th>Documents</th>
<th>Words/Doc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>5.58M</td>
<td>1,406</td>
<td>3,969</td>
</tr>
<tr>
<td>1997</td>
<td>9.52M</td>
<td>2,260</td>
<td>4,213</td>
</tr>
<tr>
<td>1998</td>
<td>12.06M</td>
<td>2,461</td>
<td>4,902</td>
</tr>
<tr>
<td>1999</td>
<td>14.77M</td>
<td>2,524</td>
<td>5,852</td>
</tr>
<tr>
<td>2000</td>
<td>13.67M</td>
<td>2,424</td>
<td>5,639</td>
</tr>
<tr>
<td>2001</td>
<td>15.64M</td>
<td>2,596</td>
<td>6,025</td>
</tr>
<tr>
<td><strong>2002</strong></td>
<td><strong>23.04M</strong></td>
<td><strong>2,845</strong></td>
<td><strong>8,100</strong></td>
</tr>
<tr>
<td>2003</td>
<td>35.78M</td>
<td>3,611</td>
<td>9,910</td>
</tr>
<tr>
<td>2004</td>
<td>39.38M</td>
<td>3,558</td>
<td>11,069</td>
</tr>
<tr>
<td>2005</td>
<td>42.39M</td>
<td>3,474</td>
<td>12,204</td>
</tr>
<tr>
<td>2006</td>
<td>39.23M</td>
<td>3,306</td>
<td>11,867</td>
</tr>
</tbody>
</table>

---

\(^6\)The Sarbanes-Oxley Act of 2002.
## Experimental Results

<table>
<thead>
<tr>
<th>Task (Features)</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>Mirco-avg</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(LOG1P+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORG</td>
<td><strong>0.18082</strong></td>
<td>0.17175</td>
<td>0.17157</td>
<td>0.12879</td>
<td>0.13038</td>
<td>0.14287</td>
<td>0.15271</td>
</tr>
<tr>
<td>SEN</td>
<td>0.18506</td>
<td><strong>0.16367</strong></td>
<td><strong>0.15795</strong></td>
<td><strong>0.12822</strong></td>
<td><strong>0.13029</strong></td>
<td><strong>0.13998</strong></td>
<td><strong>0.14894</strong></td>
</tr>
<tr>
<td><strong>Mean Squared Error</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Kendall’s Tau</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORG</td>
<td>0.62173</td>
<td><strong>0.63626</strong></td>
<td>0.58528</td>
<td><strong>0.59350</strong></td>
<td>0.59651</td>
<td>0.57641</td>
<td>0.59965</td>
</tr>
<tr>
<td>SEN</td>
<td><strong>0.63349</strong></td>
<td>0.62280</td>
<td><strong>0.60527</strong></td>
<td>0.59017</td>
<td><strong>0.60273</strong></td>
<td><strong>0.58287</strong></td>
<td><strong>0.60458</strong></td>
</tr>
<tr>
<td><strong>Spearman’s Rho</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORG</td>
<td>0.65271</td>
<td><strong>0.66692</strong></td>
<td>0.61662</td>
<td><strong>0.62317</strong></td>
<td>0.62531</td>
<td>0.60371</td>
<td>0.62939</td>
</tr>
<tr>
<td>SEN</td>
<td><strong>0.66397</strong></td>
<td>0.65303</td>
<td><strong>0.63646</strong></td>
<td>0.61953</td>
<td><strong>0.63133</strong></td>
<td><strong>0.60999</strong></td>
<td><strong>0.63403</strong></td>
</tr>
</tbody>
</table>

**Figure:** Experimental Results of Using Original Texts and Only Sentiment Words.
Experiments

Analysis: Regression and Ranking

Figure: Number of Occurrences of the Top 10 Weighted Terms Learned.
Experiments

Financial Sentiment Terms Analysis

Figure: Highly-Weighted Terms Learned from the 6 Ranking Models of Using Original Texts (ORG) and Only Sentiment Words (SEN).

Financial Sentiment Analysis for Risk Prediction
This paper identifies the importance of sentiment words in financial reports associated with financial risk.

The experimental results show that the models trained on sentiment words can result in comparable performance to those on origin texts.

The learned models also suggest strong correlations between financial sentiment words and the risk of companies.

As a result, these findings provide us more insight into the impact of financial sentiment words on companies’ future risk.