Financial Report Analysis for Risk Prediction

Prof. Chuan-Ju Wang

Department of Computer Science
University of Taipei
Joint work with Prof. Ming-Feng Tsai

November 21, 2014
Department of Computer Science and Engineering
Yuan Ze University
Outline

1. Introduction
2. Related Work
3. Methodology
4. Experiments
5. Conclusions and Future Work
In finance, there are typically two kinds of information:

1. **Soft information**: text, including opinions, ideas, and market commentary.
2. **Hard information**: numbers, such as financial measures and historical prices.

Financial field: Predict risk by GARCH model.¹ (hard information)

This paper aims to incorporate soft information to study financial risk among companies.

---

Financial risk is the amount of chance that a chosen investment instrument (e.g., stock) will lead to a loss.

In finance, volatility is a common empirical measure of risk.

This paper explores the soft textual information in financial reports.

In an attempt to analyze the financial risk of a set of companies based on the volatilities of their stock returns.
Our Approaches

- Two analytic techniques are adopted:
  - Regression approach: Predict the stock return volatilities.
  - Ranking approach:\(^2\) Rank the companies in line with their *relative risk levels*.

- The texts are the annual SEC\(^3\)-mandated financial reports.

---

\(^2\)An earlier version of this work was presented in ECIR’13.
\(^3\)Securities and Exchange Commission
Financial Sentiment Analysis

- We also apply sentiment analysis on the tasks of risk prediction.\(^4\)
- A financial-specific sentiment lexicon is adopted for analysis.
- Experimental results: Both the regression and ranking models trained on the finance-specific sentiment lexicon only can obtain comparable performance to those trained on the original texts.

\(^4\) An earlier version of this work was presented in IJCNLP’13.
Related Work: Text Mining in Finance

Most financial studies related to risk analysis are based on hard numerical information, especially on time series modeling.\(^5\)

Studies conducted on mining financial reports or news:

- Lin et al. (2008) propose a method to predict short-term stock price movements via combining both qualitative and quantitative features of financial reports.\(^6\)
- Schumaker and Chen (2009) examine a predictive machine learning approach for financial news articles analysis using several different textual representations: bag of words, noun phrases, and named entities.\(^7\)
- Kogan et al. (2009) applied a regression approach to predict stock return volatilities of companies via their financial reports.\(^8\)

---

\(^5\) E.g., Armano et al. (2005), Christoffersen et al. (2000), and Hung (2009).

\(^6\) Published in *ACM Transactions on Management Information Systems*.

\(^7\) Published in *ACM Transactions on Information Systems*.

\(^8\) Published in the Proceedings of NAACL’09.
Several studies using textual analysis to examine the sentiment of numerous news items, articles, financial reports, and tweets about public companies.\(^9\)

For most sentiment analysis algorithms, the sentiment lexicon is the most important resource.\(^{10}\)

However, a general purpose sentiment lexicon might mis-classify common words in financial text.

Loughran and McDonald (2011)\(^{11}\) provide a financial specific sentiment lexicon.

---

\(^9\) E.g., Garcia (2013) and Price et al. (2012).

\(^{10}\) Feldman. (2013), Techniques and applications for sentiment analysis. *Communications of the ACM.*

\(^{11}\) Published in *Journal of Finance.*
Risk Proxy: Stock Return Volatility

- Stock Return
  \[ R_t = \frac{(S_t - S_{t-1})}{S_{t-1}} \]

- Stock Return Volatility
  
  A common risk metric measured by the standard deviation of returns over a period of time.

  \[ v_{[t-n,t]} = \sqrt{\frac{\sum_{i=t-n}^{t} (R_i - \bar{R})^2}{n}} \]
  , where \( \bar{R} = \sum_{i=t-n}^{t} \frac{R_i}{(n+1)} \).
Risk-Level Splitting Mechanism

- Classify the volatilities of \( n \) stocks into \( 2\ell + 1 \) risk levels.
- The distribution over \( \ln(\nu) \) across companies tends to have a bell shape.
- For example, with \( \ell = 2 \), there are 5 risk levels (i.e., 0, 1, 2, 3, 4):

\[
   r = \begin{cases} 
   0 & \text{if } \ln(\nu) \in (-\infty, m - 2s], \\
   1 & \text{if } \ln(\nu) \in (m - 2s, m - s], \\
   2 & \text{if } \ln(\nu) \in (m - s, m + s), \\
   3 & \text{if } \ln(\nu) \in [m + s, m + 2s), \\
   4 & \text{if } \ln(\nu) \in [m + 2s, \infty), 
   \end{cases}
\] (1)

where \( m \) is the sample mean and \( s \) is the sample standard deviation of the logarithm of volatilities of \( n \) stocks (denoted as \( \ln(\nu) \)).\(^{12}\)

\(^{12}\)We take the logarithm of volatilities as it is standard in finance.
The words have different meaning between finance-specified lexicon and general-purpose lexicon.

- Almost three-fourths of the words in the 10-K financial reports from year 1994 to 2008, which are identified as negative by the widely used Harvard Psychosociological Dictionary, are typically not considered negative in financial contexts.\textsuperscript{13}
- Example: vice: evil, defect; vice: secondary

\textsuperscript{13}Loughran and McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. Journal of Finance.
## Six Finance-Specific Lexicons

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

- Predict target: **Stock return volatility** (regression) and **relative risk levels** (ranking)

- **Features**
  - **Soft textual information**: All words or financial sentiment words
  - **Hard numerical information**: The twelve months before the report volatility for each company
Regression

- A collection of financial reports $D = \{d_1, d_2, \ldots, d_n\}$, in which each $d_i \in \mathbb{R}^p$ is associated with a company $c_i$.
- We seek to predict the company’s volatility $v_i$.
- Such a prediction can be defined by a parameterized function $f$ as follows:
  \[
  \hat{v}_i = f(d_i; w). 
  \]  
  (2)
- The goal is to learn a $p$-dimensional vector $w$ from the training data $T = \{(d_i, v_i) | d_i \in \mathbb{R}^p, v_i \in \mathbb{R}\}$.
- Support Vector Regression (SVR) is adopted in the paper.
Our goal is to rank companies by using their financial reports according to the volatilities of stock returns.

Split the volatilities of company stock returns within a year into different risk levels.

Given a collection of financial reports $D$, we aim to rank the companies via a ranking model $f : \mathbb{R}^p \rightarrow \mathbb{R}$.

- $f(d_i) > f(d_j)$ means that the model asserts that $c_i \succ c_j$ (the company $c_i$ is more risky than $c_j$).

Ranking SVM is adopted in the paper.
Corpora: The 10-K Corpus

- Section 7 “management's discussion and analysis of financial conditions and results of operations” (MD&A)
- The Sarbanes-Oxley Act of 2002:¹⁴ Explain the drastic increase in length during the 2002-2003 period

<table>
<thead>
<tr>
<th>Year</th>
<th># of Documents</th>
<th># of Unique Terms</th>
<th># of Terms</th>
<th># Terms/Doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>1,406</td>
<td>19,613</td>
<td>2.7M</td>
<td>1,934</td>
</tr>
<tr>
<td>1997</td>
<td>2,260</td>
<td>26,039</td>
<td>4.7M</td>
<td>2,059</td>
</tr>
<tr>
<td>1998</td>
<td>2,461</td>
<td>29,020</td>
<td>5.9M</td>
<td>2,395</td>
</tr>
<tr>
<td>1999</td>
<td>2,524</td>
<td>30,359</td>
<td>7.2M</td>
<td>2,855</td>
</tr>
<tr>
<td>2000</td>
<td>2,424</td>
<td>30,312</td>
<td>6.7M</td>
<td>2,766</td>
</tr>
<tr>
<td>2001</td>
<td>2,596</td>
<td>32,292</td>
<td>7.7M</td>
<td>2,956</td>
</tr>
<tr>
<td>2002</td>
<td>2,845</td>
<td>38,692</td>
<td>11.4M</td>
<td>3,993</td>
</tr>
<tr>
<td>2003</td>
<td>3,611</td>
<td>48,513</td>
<td>17.7M</td>
<td>4,894</td>
</tr>
<tr>
<td>2004</td>
<td>3,558</td>
<td>50,674</td>
<td>19.5M</td>
<td>5,469</td>
</tr>
<tr>
<td>2005</td>
<td>3,474</td>
<td>53,388</td>
<td>21.0M</td>
<td>6,046</td>
</tr>
<tr>
<td>2006</td>
<td>3,306</td>
<td>51,147</td>
<td>19.4M</td>
<td>5,871</td>
</tr>
</tbody>
</table>

¹⁴http://en.wikipedia.org/wiki/SarbanesOxley_Act
## 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>

Table: Statistics of the Financial Lexicon. Some words occur in more than one word lists, so the number of unique stemmed sentiment words is 1,546 rather than 1,664.
Feature Representation

- We use the TFIDF, LOG1P\(^{15}\) to represent the text information of documents.

\[
\text{TFIDF}(t, d) = \frac{\text{TC}(t, d)}{|d| \times \log(|D|/|d \in D : t \in d|)} = \text{TF}(t, d) \times \text{IDF}(t, d) = \log(1 + \text{TC}(t, d))
\]

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

\(^{15}\)Kogan et al. (2009), Predicting risk from financial reports with regression. In *Proceedings of NAACL*. 
For the ranking task, we categorize the companies of each year into **5 risk levels**.

![Risk Level Distributions (1996-2006)](image)
Experimental Settings (2)

- Consecutive 5 years’ historical financial reports: Train the models

- The trained models are tested by the following year.

- Example:
  - Test set: The 2001 financial reports.
## 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>Average</th>
</tr>
</thead>
<tbody>
<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>Regression (LOG1P+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>0.17470</td>
<td>0.16002</td>
<td>0.18734</td>
<td>0.14421</td>
<td>0.13647</td>
<td>0.14638</td>
<td>0.15086</td>
</tr>
<tr>
<td>ALL</td>
<td>0.18082</td>
<td>0.17175</td>
<td>0.17175</td>
<td>0.12879</td>
<td>0.13038</td>
<td>0.14287</td>
<td>0.15436</td>
</tr>
<tr>
<td>SEN</td>
<td>0.18506</td>
<td>0.16367</td>
<td><strong>0.15795</strong></td>
<td>0.12822</td>
<td>0.13029</td>
<td>0.13998</td>
<td>0.15086</td>
</tr>
<tr>
<td>Kendall's Tau</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>0.62455</td>
<td>0.61973</td>
<td><strong>0.60755</strong></td>
<td>0.58616</td>
<td>0.59990</td>
<td>0.58248</td>
<td>0.60339</td>
</tr>
<tr>
<td>ALL</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>0.60162</td>
</tr>
<tr>
<td>SEN</td>
<td><strong>0.63349</strong></td>
<td>0.62280</td>
<td>0.60527</td>
<td>0.59017</td>
<td><strong>0.60273</strong></td>
<td>0.58287</td>
<td><strong>0.60622</strong>*</td>
</tr>
<tr>
<td>Spearman's Rho</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>0.65486</td>
<td>0.65001</td>
<td><strong>0.63874</strong></td>
<td>0.61548</td>
<td>0.62857</td>
<td>0.60942</td>
<td>0.63284</td>
</tr>
<tr>
<td>ALL</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>0.63141</td>
</tr>
<tr>
<td>SEN</td>
<td><strong>0.66397</strong></td>
<td>0.65303</td>
<td>0.63646</td>
<td>0.61953</td>
<td><strong>0.63133</strong></td>
<td><strong>0.60999</strong></td>
<td><strong>0.63572</strong>*</td>
</tr>
</tbody>
</table>
Analysis: Regression and Ranking

Figure: Number of Occurrences of the Top 10 Weighted Terms Learned.
Financial Sentiment Terms Analysis

Figure: Highly-Weighted Terms Learned from the 6 Ranking Models of Using Original Texts (ORG) and Only Sentiment Words (SEN).
The term “amend” from the Fin-Lit list:

(from AGO, 2006 Form 10-K)
On March 22, 2005, we amended the term loan agreements to, among other reasons, lower the borrowing rate by 25 basis points from LIBOR plus 2.00% to LIBOR plus 1.75%.

- In finance, the *amend* usually means “to change by some formal processes.”
- This top-ranked term suggests that companies amending their policies frequently are associated with relative high risk.
The term “deficit” from the Fin-Neg list:

(from AXS-One Inc., 2006 Form 10-K)
At December 31, 2005, we had cash and cash equivalents of $3.6 million and a working capital deficit of $3.6 million which included $8.2 million of deferred revenue. The increase of the working capital deficit from $3.3 million at December 31, 2004 is primarily the result of a decrease in cash and decreased accounts receivable offset partially by a decrease in deferred revenue.
Summary of the Experiments

- The words learned from the ranking models are much more consistent than those from the regression ones.
- Using only sentiment words as the training data:
  - Generate better performance than using the original texts
  - Provide a way to understand the relations between financial risk and financial sentiment information.
This paper incorporates soft information to study financial risk among companies.

1. Introduce the ranking approach.
2. Attest the importance of the financial sentiment words on risk prediction.
3. Provide us more insights and understanding into the impact of financial soft textual information, especially financial sentiments, on companies’ future risk analysis.
Future Work

- Use hard information to modify the constraints of Ranking SVM, rather than simply treating the information as features.
  - The hard information usually dominates the learned model.
  - It may cause the difficulty in analyzing soft information in the learned model.

- Apply the continuous vector representations of words for discovering keywords from a financial sentiment lexicon.\textsuperscript{16}

- Build a Chinese financial sentiment lexicon.

\textsuperscript{16}An earlier version of this work was presented in EMNLP’14.