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

IJCNLP 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.

[1] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics.*

[2] Kogan et al. (2009). Predicting risk from financial reports with regression. *In NAACL '09.*

Introduction

- 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.

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 s returns over a period of time.
- Stock Return Volatility
  - Let S_t be the price of a stock at time t.

$$V_{[t-n,t]} = \sqrt{\frac{\sum_{i=t-n}^{t}(R_i - \overline{R})^2}{n}}, \textit{where } \overline{R} = \sum_{i=t-n}^{t} \frac{R_i}{(n+1)}.$$

Finance-specified Sentiment Lexicon

- 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[4]

Class	Meaning	Examples
Fin-Neg	Negative business terminologies	deficit, delist
Fin-Pos	Positive business terminologies	profit, integr
Fin-Unc	Words denoting uncertainty	doubt
Fin-Lit	Propensity for legal contest	amend, forbear
MW-Strong	Strong levels of confidence	must, best
MW-Weak	Weak levels of confidence	may, perhaps

[4] Loughran and McDonald. (2011), When is a liability not a liability? *The Journal of Finance.*

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).

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
  - An annual report required by the Securities and Exchange Commission (SEC) since 1996 to 2006.
- Six Finance-Specific Lexicons
  - Fin-Neg
  - Fin-Pos
  - Fin-Unc
  - Fin-Lit
  - MW-Strong
  - MW-Weak

Statistics of the Financial Lexicon

Dictionary	# of Words	# of Stemmed Words
Fin-Neg	2,349	918
Fin-Pos	354	151
Fin-Unc	291	127
Fin-Lit	871	443
MW-Strong	19	10
MW-Weak	27	15
Total	3,911	1,664

Feature Representation

- We use the *TFIDF*, *LOG1P*[5] to represent the text information of documents.

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

$$LOG1P = \log(1 + TC(t,\mathbf{d}))$$

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

[5] Kogan et al. (2009), Predicting risk from financial reports with regression. *In NAACL '09.*

Experimental Setting

- We use every 5 years historical financial reports to train the models
  - The trained models are tested by the following year.
- Example:
  - Training set: The 1996-2000 year financial reports.
  - Test set: The 2001 financial reports.

Corpora statistic

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

[6] The Sarbanes-Oxley Act of 2002.

Experimental Results

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

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

Analysis: Regression and Ranking

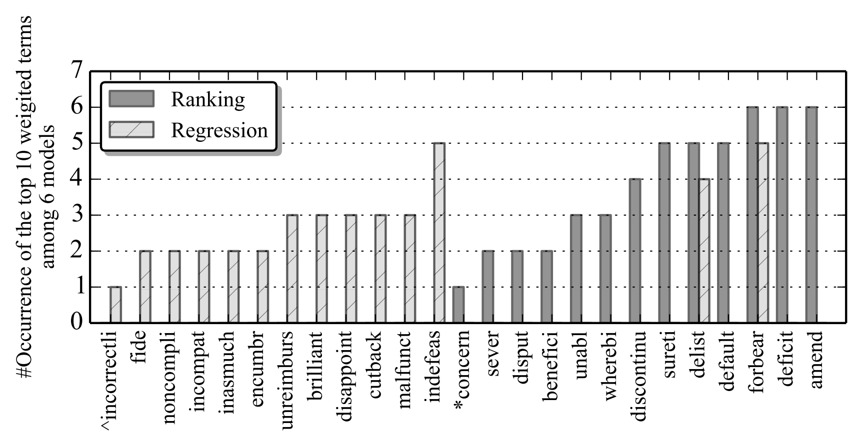

Figure : **Number of Occurrences of the Top 10 Weighted Terms Learned.**

Financial Sentiment Terms Analysis

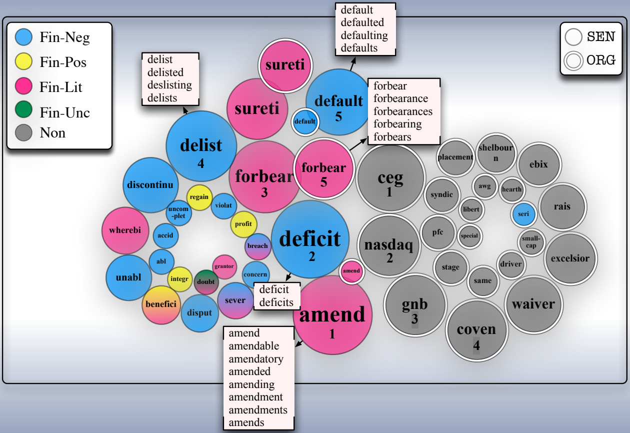

Figure : **Highly-Weighted Terms Learned from the 6 Ranking Models of Using Original Texts (ORG) and Only Sentiment Words (SEN).**

Conclusion

- 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.