Financial Keyword Expansion via Continuous Word Vector Representations

Ming-Feng Tsai
Department of Computer Science &
Program in Digital Content and Technology
National Chengchi University
Taipei 116, Taiwan
mftsai@nccu.edu.tw

Chuan-Ju Wang
Department of Computer Science
University of Taipei
Taipei 100, Taiwan
cjwang@utaipei.edu.tw

Abstract

This paper proposes to apply the continuous vector representations of words for discovering keywords from a financial sentiment lexicon. In order to capture more keywords, we also incorporate syntactic information into the Continuous Bag-of-Words (CBOW) model. Experimental results on a task of financial risk prediction using the discovered keywords demonstrate that the proposed approach is good at predicting financial risk.

1 Introduction

In the present environment with a great deal of information, how to discover useful insights for decision making is becoming increasingly important. In finance, there are typically two kinds of information (Petersen, 2004): *soft information* usually refers to text, including opinions, ideas, and market commentary, whereas *hard information* is recorded as numbers, such as financial measures and historical prices. Most financial studies related to risk analysis are based on hard information, especially on time series modeling (Christoffersen and Diebold, 2000; Lee and Tong, 2011; Wu et al., 2014; Yümlü et al., 2005). Despite of using only hard information, some literature incorporates soft textual information to predict financial risk (Kogan et al., 2009; Leidner and Schilder, 2010; Tsai and Wang, 2013). Moreover, sentiment analysis, a technique to make an assessment of the sentiments expressed in various information, has also been applied to analyze the soft textual information in financial news, reports, and social media data (Devitt and Ahmad, 2007; Loughran and McDonald, 2011; Wang et al., 2013).

Continuous vector space models (Bengio et al., 2003; Schwenk, 2007; Mikolov et al., 2010) are neural network language models, in which words are represented as high dimensional real valued vectors. These representations have recently demonstrated promising results across variety of tasks (Schwenk, 2007; Collobert and Weston, 2008; Glorot et al., 2011; Socher et al., 2011; Weston et al., 2011), because of their superiority of capturing syntactic and semantic regularities in language. In this paper, we apply the Continuous Bag-of-Words (CBOW) model (Mikolov et al., 2013) on the soft textual information in financial reports for discovering keywords via financial sentiments. In specific, we use the continuous vector representations of words to find out similar terms based on their contexts. Additionally, we propose a straightforward approach to incorporate syntactic information into the CBOW model for better locating similarly meaningful or highly correlated words. To the best of our knowledge, this is the first work to incorporate more syntactic information by adding Part-Of-Speech (POS) tags to the words before training the CBOW model.

In our experiments, the corpora are the annual SEC[1]-mandated financial reports, and there are 3,911 financial sentiment keywords for expansion. In order to verify the effectiveness of the expanded keywords, we then conduct two prediction tasks, including regression and ranking. Observed from our experimental results, for the regression and ranking tasks, the models trained on the expanded keywords are consistently better than those trained the original sentiment keywords only. In addition, for comparison, we conduct experiments with random keyword expansion as baselines. According to the experimental results, the expansion either with or without syntactic information outperforms the baselines. The results suggest that the CBOW model is effective at expanding keywords for financial risk prediction.

[1] Securities and Exchange Commission

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1453–1458,
October 25-29, 2014, Doha, Qatar.

2 Keyword Expansion via Financial Sentiment Lexicon

2.1 Financial Sentiment Lexicon

A sentiment lexicon is the most important resource for sentiment analysis. Loughran and McDonald (2011) states that a general purpose sentiment lexicon (e.g., the Harvard Psychosociological Dictionary) might misclassify common words in financial texts. Therefore, in this paper, we use a finance-specific lexicon that consists of the 6 word lists provided by (Loughran and McDonald, 2011) as seeds to expand keywords. The six lists are negative (Fin-Neg), positive (Fin-Pos), uncertainty (Fin-Unc), litigious (Fin-Lit), strong modal words (MW-Strong), and weak modal words (MW-Weak).[2]

2.2 Simple Keyword Expansion

With the financial sentiment lexicon, we first use a collection of financial reports as the training texts to learn continuous vector representations of words. Then, each word in the sentiment lexicon is used as a seed to obtain the words with the highest n cosine distances (called the top-n words for the word) via the learned word vector representations. Finally, we construct an expanded keyword list from the top-n words for each word.

2.3 Keyword Expansion with Syntactic Information

For the expansion considering syntactic information, we attach the POS tag to each word in the training texts first. Then, the words in the sentiment lexicon with 4 major POS tags (i.e., `JJ`, `NN`, `VB`, `RB`) are used as seeds to expand. The rest of steps is similar to that in Section 2.2.

The reason of considering POS tags for expansion is that, in general, a word with different POS tags may result in different lists of top-n words. Table 1 shows the top-5 words for the word "default" with different POS tags (noun and adjective). Note that none of the words in the two lists overlaps.

default (`NN`)		default (`JJ`)	
Word	Cosine Distance	Word	Cosine Distance
default (v.)	0.63665	nonconform (v.)	0.63462
unwaiv (v.)	0.63466	subprim (v.)	0.62404
uncur (v.)	0.62285	chattel (n.)	0.61510
trigger (n.)	0.60080	foreclos (adj.)	0.61397
unmatur (v.)	0.58208	unguarante (v.)	0.60559

Table 1: Top-5 Words for the word "default."

3 Financial Risk Prediction

3.1 The Risk Measure: Volatility

Volatility is a measure for variation of prices of a stock over a period of time. Let S_t be the price of a stock at time t. Holding the stock from time $t-1$ to time t would lead to a simple return: $R_t = S_t/S_{t-1} - 1$ (Tsay, 2005). The volatility of returns for a stock from time $t-n$ to t can thus be defined as follows:

$$v_{[t-n,t]} = \sqrt{\frac{\sum_{i=t-n}^{t} (R_i - \bar{R})^2}{n}}, \qquad (1)$$

where $\bar{R} = \sum_{i=t-n}^{t} R_i/(n+1)$.

3.2 Regression Task

Given a collection of financial reports $D = \{\mathbf{d_1}, \mathbf{d_2}, \ldots, \mathbf{d_n}\}$, in which each $\mathbf{d_i} \in \mathbb{R}^p$ and is associated with a company c_i, we aim to predict the future risk of each company c_i (which is characterized by its volatility v_i). This prediction problem can be defined as follows:

$$\hat{v}_i = f(\mathbf{d_i}; \mathbf{w}). \qquad (2)$$

The goal is to learn a p-dimensional vector $\mathbf{w}$ from the training data $T = \{(\mathbf{d}_i, v_i) | \mathbf{d}_i \in \mathbb{R}^p, v_i \in \mathbb{R}\}$. In this paper, we adopt the Support Vector Regression (SVR) (Drucker et al., 1997) for training such a regression model. More details about SVR can be found in (Schölkopf and Smola, 2001).

3.3 Ranking Task

Instead of predicting the volatility of each company in the regression task, the ranking task aims to rank companies according to their risk via the textual information in their financial reports. We first split the volatilities of company stock returns within a year into different risk levels by the mechanism provided in (Tsai and Wang, 2013). The risk levels can be considered as the relative difference of risk among the companies.

After obtaining the relative risk levels of the companies, the ranking task can be defined as follows: Given a collection of financial reports D, we aim to rank the companies via a ranking model $f : \mathbb{R}^p \rightarrow \mathbb{R}$ such that the rank order of the set of companies is specified by the real value that the

[2] `http://www.nd.edu/~mcdonald/Word_Lists.html`.

model f takes. Specifically, $f(\mathbf{d_i}) > f(\mathbf{d_j})$ means that the model asserts that $c_i \succ c_j$, where $c_i \succ c_j$ means that c_i is ranked higher than c_j; that is, the company c_i is more risky than c_j. For this task, this paper adopts Ranking SVM (Joachims, 2006).

4 Experiments

4.1 Dataset and Preprocessings

In the experiments, we use the 10-K corpus (Kogan et al., 2009) to conduct our financial risk prediction tasks. All documents and the 6 financial sentiment word lists are stemmed by the Porter stemmer (Porter, 1980), and some stop words are also removed.

For financial risk prediction, the ground truth, the twelve months after the report volatility for each company, $v^{+(12)}$, (which measures the future risk for each company) can be calculated by Equation (1), where the stock prices can be obtained from the Center for Research in Security Prices (CRSP) US Stocks Database. In addition, to obtain the relative risks among companies used in the ranking task, we follow (Tsai and Wang, 2013) to split the companies of each year into 5 risk levels.

4.2 Keyword Expansion

In our experiments, Section 7 (Management Discussion and Analysis) in 10-K corpus is used as training texts for the tool (word2vec[3]) to learn the continuous vector representations of words.

For the simple expansion (denoted as EXP−SIM hereafter), we use the total 1,667 stemmed sentiment words as seeds to obtain the expanded keywords via the learned word vector representations. For the expansion considering syntactic information (denoted as EXP−SYN), NLTK[4] is applied to attach the POS tag[5] to each word in the training texts; we attach the POS tag to a word with an underscore notation (e.g., default_VB). For simplicity, we combine some POS tags to one tag via the tag replacement; for example, the tags JJR (adjective, comparative) and JJS (adjective, superlative) are replaced to JJ (adjective). The detailed replacement rules are tabulated in Table 2. Words from the sentiment lexicon with the four types of POS tags (i.e., JJ, NN, VB, RB) are consider as the seeds to expand the keywords. For both EXP−SIM and EXP−SYN, we use the top-20 expanded words for each word (e.g., Table 3) to construct expanded keyword lists. In total, for EXP−SIM, the expanded list contains 9,282 unique words and for EXP−SYN, the list has 13,534 unique words.

[3] https://code.google.com/p/word2vec/

[4] http://www.nltk.org/

[5] The most common POS tag scheme, the Penn Treebank POS Tags, is adopt in the paper.

After Replacement	Before Replacement
JJ	JJ, JJR, JJS
NN	NN, NNS, NNP, NNPS
PRP	PRP, PRP$
RB	RB, RBR, RBS
VB	VB, VBD, VBG, VBN, VBP, VBZ
WP	WP, WP$

Table 2: Tag Replacement Rules.

Word	Cosine Distance	Word	Cosine Distance
uncur	0.569498	event	0.466834
indentur	0.565450	lender	0.459995
waiv	0.563656	forbear	0.456556
trigger	0.559936	represent	0.450631
cure	0.539999	breach	0.446851
nonpay	0.538445	noncompli	0.431490
unmatur	0.525251	gecc	0.430712
unwaiv	0.510359	customari	0.424447
insolv	0.488534	waiver	0.419338
occurr	0.471123	prepay	0.418969

Table 3: Top-20 (Stemmed) Words for the Word "default."

4.3 Word Features

In the experiments, the bag-of-words model is adopted and three word features are used to represent the 10-K reports in the experiments. Given a document $\mathbf{d}$, three word features (i.e., TF, TFIDF and LOG1P) are calculated as follows:

- $\texttt{TF}(t, \mathbf{d}) = \texttt{TC}(t, \mathbf{d})/|\mathbf{d}|$,
- $\texttt{TFIDF}(t, \mathbf{d}) = \texttt{TF}(t, \mathbf{d}) \times \texttt{IDF}(t, \mathbf{d}) = \texttt{TC}(t, \mathbf{d})/|\mathbf{d}| \times \log(|D|/|\mathbf{d} \in D : t \in \mathbf{d}|)$,
- $\texttt{LOG1P} = \log(1 + \texttt{TC}(t, \mathbf{d}))$,

where $\texttt{TC}(t, \mathbf{d})$ denotes the term count of t in $\mathbf{d}$, $|\mathbf{d}|$ is the length of document $\mathbf{d}$, and D denotes the set of all documents in each year.

4.4 Experimental Results

Tables 4 and 5 tabulate the experimental results of regression and ranking, respectively, in which the training data is composed of the financial reports in a five-year period, and the following year is the test data. For example, the reports from year 1996 to 2000 constitute a training data, and the learned model is tested on the reports of year 2001.

[TFIDF] Year	SEN	(Baseline) EXP−RAN	EXP−SIM	EXP−SYN	SEN	(Baseline) EXP−RAN	EXP−SIM	EXP−SYN
	Kendall's Tau (Kendall, 1938).				Spearman's Rho (Myers et al., 2003)			
2001	0.4384	0.4574	0.4952	**0.5049**	0.4701	0.4889	0.5266	**0.5375**
2002	0.4421	0.4706	0.4881	**0.4944**	0.4719	0.5007	0.5187	**0.5256**
2003	0.4414	0.4706	**0.5105**	0.5006	0.4716	0.5015	**0.5418**	0.5318
2004	0.4051	0.4551	0.4750	**0.4961**	0.4335	0.4842	0.5043	**0.5255**
2005	0.3856	0.4482	0.5126	**0.5294**	0.4117	0.4757	0.5418	**0.5579**
2006	0.3784	0.4385	0.4588	**0.4867**	0.4029	0.4641	0.4847	**0.5129**

Table 5: Performance of Ranking.

[LOGP] Year	SEN	(Baseline) EXP−RAN	EXP−SIM	EXP−SYN
	Mean Squared Error			
2001	0.2526	0.2360	0.2195	**0.2148**
2002	0.2858	0.2649	0.2433	**0.2381**
2003	0.2667	0.2512	0.2320	**0.2350**
2004	0.2345	0.2140	0.1902	**0.1872**
2005	0.2241	0.2014	0.1754	**0.1682**
2006	0.2256	0.2072	0.1889	**0.1825**

Table 4: Performance of Regression

In the tables, SEN denotes the experiments trained on the words from the original financial sentiment lexicon. Despite of the experiments trained on EXP−SIM and EXP−SYN, we also conduct experiments with random keyword expansion (called EXP−RAN); for the comparison purpose, we keep the number of words in the randomly expanded word list the same as that in EXP−SYN. Note that the randomly expanded list contains all sentiment words and the rest of words are randomly chosen from the vocabulary of the dataset. The columns with label EXP−RAN denote the results averaged from 20 randomly expanded word lists. The bold face numbers denote the best performance among the four word lists.

As shown in Tables 4 and 5, for both regression and ranking tasks, the models trained on expanded keywords (i.e., EXP−$\star$) are consistently better than those trained on the original sentiment keywords only.[6] Additionally, we treat the experiments with randomly expanded word list (EXP−RAN) as the baselines.[7] From the two tables, we observe that the expansion either with or without syntactic information outperforms the baselines. Note that, for the EXP−SIM, the number of words used for training the regression and ranking models is even less than that of EXP−RAN. The results suggest that the CBOW model is effective at expanding keywords for financial risk prediction. Furthermore, incorporating syntactic information into the CBOW model can even enhance the performance for the tasks of financial risk prediction.

[6]Due to the page limits, only the results trained on features LOGP for regression and TFIDF for ranking are reported, but the performance for models trained on features TF, TFIFG, and LOGP is very consistent.

[7]The results for EXP−SYN are all significant better than the baseline with $p < 0.05$.

4.5 Discussions

Below we provide the original texts from 10-K reports that contain the top 1 expanded word, "uncur" (stemmed), for "default" in Table 3. Two pieces of sentences are listed as follows (the company Investment Technology Group, 1997):

> $\cdots$ terminate the agreement upon certain events of bankruptcy or insolvency or upon an **uncured** breach by the Company of certain covenants $\cdots$

> $\cdots$ any termination of the license agreement resulting from an **uncured** default would have a material adverse effect on the Company's results of operations.

From the above examples, the expanded word "uncur" has similar meaning to "default," which confirms the capability of our method of capturing similarly meaningful or highly correlated words.

5 Conclusions

This paper applies the continuous bag-of-words model on the textual information in financial reports for expanding keywords from a financial sentiment lexicon. Additionally, we propose a simple but novel approach to incorporate syntactic information into the continuous bag-of-words model for capturing more similarly meaningful or highly correlated keywords. The experimental results for the risk prediction problem show that the expansion either with or without syntactic information outperforms the baselines. As a direction for further

research, it is interesting and important to provide more analysis on the expanded words via the continuous vector representations of words.

Acknowledgments

This research was partially supported by the National Science Council of Taiwan under the grants NSC 102-2420-H-004-052-MY2, 102-2221-E-004-006, and 102-2221-E-845-002-MY3.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Peter F Christoffersen and Francis X Diebold. 2000. How relevant is volatility forecasting for financial risk management? *Review of Economics and Statistics*, 82(1):12–22.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, IMCL '08, pages 160–167.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 984–991.

Harris Drucker, Chris JC Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9:155–161.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pages 513–520.

Thorsten Joachims. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 217–226.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 272–280.

Yi-Shian Lee and Lee-Ing Tong. 2011. Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge-Based Systems*, 24(1):66–72.

Jochen L. Leidner and Frank Schilder. 2010. Hunting for the black swan: risk mining from text. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 54–59.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Jerome L Myers, Arnold D Well, and Robert F Lorch Jr. 2003. *Research design and statistical analysis*. Routledge.

Mitchell A Petersen. 2004. Information: Hard and soft. Technical report, Northwestern University.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

Bernhard Schölkopf and Alexander J Smola. 2001. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech & Language*, 21(3):492–518.

Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pages 129–136.

Ming-Feng Tsai and Chuan-Ju Wang. 2013. Risk ranking from financial reports. In *Advances in Information Retrieval*, pages 804–807. Springer.

Ruey S Tsay. 2005. *Analysis of financial time series*. Wiley.

Chuan-Ju Wang, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang. 2013. Financial sentiment analysis for risk prediction. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, IJCNLP '13, pages 802–808.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of the Twenty-Second international joint conference on Artificial*

Intelligence-Volume Volume Three, pages 2764–2770.

Desheng Dash Wu, Shu-Heng Chen, and David L Olson. 2014. Business intelligence in risk management: Some recent progresses. *Information Sciences*, 256:1–7.

Serdar Yümlü, Fikret S Gürgen, and Nesrin Okay. 2005. A comparison of global, recurrent and smoothed-piecewise neural models for istanbul stock exchange (ise) prediction. *Pattern Recognition Letters*, 26(13):2093–2103.