Financial Keyword Expansion via Continuous Word Vector Representations

Ming-Feng Tsai
Department of Computer Science, National Chengchi University, Taiwan

Chuan-Ju Wang
Department of Computer Science, University of Taipei, Taiwan

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.

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: 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. Despite of using only hard information, some literature incorporates soft textual information to predict financial risk. 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.

Continuous vector space models 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, because of their superiority of capturing syntactic and semantic regularities in language. In this paper, we apply the Continuous Bag-of-Words (CBOW) model 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-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.

Keyword Expansion via 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.

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.

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 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.

Financial Risk Prediction

Regression Task

Given a collection of financial reports $D = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_n\}$, in which each $\boldsymbol{d}_i \in \mathbb{R}^p$ 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: $v_i = f(\boldsymbol{d}_i; \boldsymbol{w})$. The goal is to learn a p-dimensional vector $\boldsymbol{w}$ from the training data $T = \{(\boldsymbol{d}_i, v_i) \mid \boldsymbol{d}_i \in \mathbb{R}^p, v_i \in \mathbb{R}\}$. In this paper, we adopt the Support Vector Regression (SVR) for training such a regression model.

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 model f takes. Specifically, $f(\boldsymbol{d}_i) > f(\boldsymbol{d}_j)$ means that the model asserts that $c_i > c_j$, where $c_i > c_j$ means that c_i is ranked higher than c_j ; that is, the company c_i is more risky than c_j. This paper adopts Ranking SVM.

Experiments and Discussions

Dataset and Preprocessings

In the experiments, we use the 10-K corpus to conduct our financial risk prediction tasks. All documents and the 6 financial sentiment word lists are stemmed by the Porter stemmer, and some stop words are also removed. For financial risk prediction, the twelve months after the report volatility for each company, $v^{+(12)}$, (which measures the future risk for each company) is treated as the ground truth, 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.

Keyword Expansion

In our experiments, Section 7 (Management Discussion and Analysis) in 10-K corpus is used as training texts for the tool (word2vec) 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 is applied to attach the POS tag to each word in the training texts; we attach the POS tag to a word with an underscore notation (e.g., default_`VB`). 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.

[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

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."

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 ***d***, three word features, `TF`, `TFIDF` and `LOG1P`, are used.

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.

Discussions

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

> · · · terminate the agreement upon certain events of bankruptcy or insolvency or upon an **uncured** breach by the Company of certain covenants · · ·
>
> · · · 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.

Conclusions and Future Work

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.

EMNLP 2014