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 financial analysis are based on hard information, especially on time series modeling. Despite of using only hard information, some literature incorporates soft text information to predict financial risk. Moreover, sentiment analysis, a technique to make an assessment of the sentiments expressed at soft 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 various 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 sentiment lexicon. 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 Full-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 Psychosocial Dictionary) might missclassify 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 the original sentiment lexicon.

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, VB, RB, IN) are used as seeds to expand the original sentiment lexicon. The 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 $D = \{d_1, d_2, \ldots, d_N\}$, in which each $d_i \in \mathbb{R}^p$ 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 $\sigma_i$). This prediction problem can be defined as follows: $\sigma_i = f(d_i; w)$. The goal is to learn a p-dimensional vector $w$ from the training data $T = \{d_i, y_i\}$, where $c_i \in \mathbb{R}_+$, $y_i \in \mathbb{R}_+$. In this paper, we adopt the Support Vector Regression (SVR) for training such a regression model.

Ranking Task

Instead of estimating 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 defining 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 $Exp\rightarrow R$, such that the rank order of the set of companies is specified by the real value that the model $f$ takes. Specifically, $f(d_i) > f(d_j)$ 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, $\sigma_i(t)$, (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) U.S. 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−S1H 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−S1H), NLTK is applied to attach the POS tag to each word in the training texts; we attach the POS tag to each word with an underscore notation (e.g., default_VB). For both Exp−S1H and Exp−SY, we use the top-20 expanded words for each word (e.g., Table 3) to construct expanded keyword lists. In total, for Exp−S1H, the expanded list contains 9,282 unique words and for Exp−SY, the list has 13,534 unique words.

Table 4: Performance of Regression

<table>
<thead>
<tr>
<th>Word</th>
<th>Cosine Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>uncured</td>
<td>0.7357</td>
</tr>
<tr>
<td>default</td>
<td>0.7357</td>
</tr>
</tbody>
</table>

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 features, TF, TFIDF, and LOGF, 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.

Table 5: Performance of Ranking

<table>
<thead>
<tr>
<th>Word</th>
<th>Cosine Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>uncur</td>
<td>0.5129</td>
</tr>
<tr>
<td>breach</td>
<td>0.5129</td>
</tr>
</tbody>
</table>

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 example is from the “In Each of the following...” section):

“uncur” (stemmed), for the word “default” in Table 3. Two pieces of sentences are listed as follows (the example is from the “In Each of the following...” section):

“...and Discussions

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 © Ming-Feng Tsai and Chuan-Ju Wang 2014