European Journal of Operational Research 000 (2016) 1–8

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

On the risk prediction and analysis of soft information in finance reports

Ming-Feng Tsai[a], Chuan-Ju Wang[b,*]

[a] Department of Computer Science and Program in Digital Content and Technology, National Chengchi University, No. 64, Sec. 2, Zhinan Rd., Taipei 116, Taiwan
[b] Research Center for Information Technology Innovation, Academia Sinica, No. 128 Academia Road, Sec. 2, Taipei 115, Taiwan

ARTICLE INFO

Article history:
Received 29 September 2015
Accepted 29 June 2016
Available online xxx

Keywords:
Finance
Risk prediction
Text mining
Sentiment analysis

ABSTRACT

We attempt in this paper to utilize soft information in financial reports to analyze financial risk among companies. Specifically, on the basis of the text information in financial reports, which is the so-called soft information, we apply analytical techniques to study relations between texts and financial risk. Furthermore, we conduct a study on financial sentiment analysis by using a finance-specific sentiment lexicon to examine the relations between financial sentiment words and financial risk. A large collection of financial reports published annually by publicly-traded companies is employed to conduct our experiments; moreover, two analytical techniques – regression and ranking methods – are applied to conduct these analyses. The experimental results show that, based on a bag-of-words model, using only financial sentiment words results in performance comparable to using the whole texts; this confirms the importance of financial sentiment words with respect to risk prediction. In addition to this performance comparison, via the learned models, we draw attention to some strong and interesting correlations between texts and financial risk. These valuable findings yield greater insight and understanding into the usefulness of soft information in financial reports and can be applied to a broad range of financial and accounting applications.

1. Introduction

The great amounts of data in today's environment make it more and more important to determine how to discover useful insights for improved decision-making. These discovered insights can result in the ability to take advantage of opportunities, minimize risks, and control costs. Big data analytics refers to techniques for exploring, discovering, and making data-driven decisions in the context of abundant data. These techniques include efforts toward using new analytic methods on either new data or data that has been combined in new ways.

Due to the prevalence of big data analytics, in recent years researchers have started to focus on analyzing new types of information. In finance, there are typically two kinds of information (Petersen, 2004): soft information, which usually refers to text, including opinions, ideas, and market commentary; and hard information, that is, numbers such as financial measures and historical prices. In contrast to previous works which use only hard information in the modeling of financial risk, in this paper we aim to incorporate soft information to study financial risk among companies.

Financial risk is the chance that a chosen investment instrument (e.g., stock) will lead to a loss. In finance, volatility is a common empirical measure of risk. Our main focus in this paper is to apply sentiment analysis to the task of risk prediction in an attempt to discover useful insights. In this study, we use a finance-specific sentiment lexicon to model the relations between sentiment information and financial risk; in specific, two analytic techniques are adopted: regression and ranking methods, and the texts are the annual SEC[1]-mandated financial reports. For the regression task, we attempt to predict stock return volatility via soft textual information. However, according to Kogan, Levin, Routledge, Sagi, and Smith (2009), it is considered difficult to thus predict real-world quantities using text information only; this is probably due to the huge amount of noise within text. Therefore, we propose solving this noise problem by using ranking techniques. Specifically, we first split the volatilities of company stock returns within a given year into several relative risk levels, and then we apply

* Corresponding author.
E-mail addresses: mftsai@nccu.edu.tw (M.-F. Tsai), cjwang@citi.sinica.edu.tw (C.-J. Wang).

[1] Securities and Exchange Commission.

http://dx.doi.org/10.1016/j.ejor.2016.06.069
0377-2217/

ranking techniques to rank the companies according to their relative risk levels. From the experimental results, we observe that, when trained on the finance-specific sentiment lexicon only, both regression and ranking models yield performance comparable to those trained on the original texts, even though the word dimension is reduced considerably, from hundreds of thousands to around only 1500. This indicates that finance-specific sentiments are the most crucial ingredients in financial reports. In addition, we also conduct analyses on the resultant models; this yields more insight and understanding into the impact of soft information in financial reports.

In addition to the proposed techniques, this paper also presents a web-based information system for financial report analysis and visualization to bridge the gap between technical results and useful interpretations.[2] With the system and our analyzed results, both academics and practitioners can more easily capture useful insights and understand the impact of soft information in financial reports. One potential application of the analyzed soft information is to help banks improve their credit-risk assessment, in particular their approach to qualitative assessment.[3] Moreover, practitioners such as fund managers can utilize the learned high-risk sentiment keywords to assist in designing their own investment strategies. For accounting research also, understanding the soft information in financial reports is a vital task, because the soft information can provide a very helpful context for understanding financial data and testing interesting economic hypotheses (Li, 2010). Therefore, it can be said that this study can be applied to a broad range of financial and accounting applications.

The remainder of this paper is organized as follows. In Section 2, we present related past work and outline our aims. We then describe in Section 3, how we accomplish our analysis: the definition of the risk measure, the mechanism of risk-level splitting, the financial sentiment lexicon, and the problem formulation. In Section 4, we present the details of our experimental settings and experimental results. In Section 5, we provide discussion and analysis, after which we conclude the paper.

2. Related work

In finance, there are typically two kinds of information: soft and hard information (Petersen, 2004). Soft information usually refers to textual information, including opinions, ideas, and market commentary, and hard information refers to numerical information such as historical time series of stock prices. Most financial studies related to risk analysis are based on hard numerical information, especially time series modeling (e.g., Armano, Marchesi, & Murru, 2005; Bodyanskiy & Popov, 2006; Christoffersen & Diebold, 2000; Chu, Santoni, & Liu, 1996; Dash, Hanumara, & Kajiji, 2003; Fu, 2011; Hung, 2009; Laih, 2014; Lee & Tong, 2011; Wu, Chen, & Olson, 2014; Yümlü, Gürgen, & Okay, 2005; Wong, Xia, & Chu, 2010). In natural language processing, some have used regression to predict continuous quantities. For instance, Mcauliffe and Blei (2007) predicted movie reviews and popularity from text via latent "topic" variables, and Lavrenko et al. (2000) used language models to analyze influences between text and time-series financial data (stock prices). In addition, in information retrieval, in recent years there have also been attempts to use learning-based methods to solve the text ranking problem (e.g., Burges et al., 2005; Freund, Iyer, Schapire, & Singer, 2003; Joachims, 2006), which has subsequently brought to the fore the topic of "learning to rank" in the fields of information retrieval and machine learning.

Some researchers have focused on mining financial reports or news (e.g., Balakrishnan, Qiu, & Srinivasan, 2010; Blasco, Corredor, Del Rio, & Santamaria, 2005; Groth & Muntermann, 2011; Huang & Li, 2011; Kogan et al., 2009; Leidner & Schilder, 2010; Lin, Lee, Kao, & Chen, 2008; Schumaker & Chen, 2009). Lin et al. (2008) used a weighting scheme to combine both qualitative and quantitative features of financial reports, and then proposed a method to predict short-term stock price movements. They used a hierarchical agglomerative clustering (HAC) method with K-means updating to improve the purity of the prototypes of financial reports, and then used the generated prototypes to predict stock price movements. Other research has focused on predicting risk from financial reports, for instance (Leidner & Schilder, 2010), in which the text mining was used to detect risks within a company, and then classify the detected risk into several types. The above two studies both used classification to mine financial reports. In 2009, Kogan et al. (2009) applied a regression approach to predict stock return volatilities of companies via their financial reports; specifically, the support vector regression (SVR) model was applied to mine the text information. Also, two state-of-the-art studies on textual information in MD&A disclosures have been conducted by Ball, Hoberg, and Maksimovic (2015), Frankel, Jennings, and Lee (2015); the first study points out that the content of the MD&A can be systematically adopted to explain the valuation of firms, whereas the second utilizes MD&A disclosures to predict current-year firm-level accruals via SVR.

Furthermore, following the explosion of sentiment information from social web sites, blogs, and online forums, sentiment analysis has emerged as a popular research area in computational linguistics (Mohammad & Turney, 2010; Narayanan, Liu, & Choudhary, 2009). In general, sentiment analysis attempts to determine author attitudes about given topics: this could include the author's judgments or evaluations, the author's emotional state when writing, or the author's intended emotional communication to readers. The growing importance of sentiment analysis applied to finance raises many research and practical issues, such as "Why is sentiment analysis important?" In finance, there have been several studies (e.g., Garcia, 2013; Loughran & McDonald, 2011; Price, Doran, Peterson, & Bliss, 2012) that used textual analysis to examine the sentiment of numerous news items, articles, financial reports, and tweets about public companies. For most sentiment analysis algorithms, the sentiment lexicon is the most important resource and has yielded improved results and analysis (Feldman, 2013). However, past works usually used general sentiment lexicons for analysis. As mentioned in Loughran and McDonald (2011), a general purpose sentiment lexicon can be prone to misclassify common words in financial texts; as shown in their work, almost three-fourths of the words in financial reports, which are identified as negative by the widely used Harvard Psychosociological Dictionary, are typically not considered negative in financial contexts.

In this paper we aim to apply the analytical techniques of regression and ranking methods to study the relations between texts and financial risk; moreover, we also conduct a study on sentiment analysis using a finance-specific sentiment lexicon. Via the experimental results, we attempt to identify interesting correlations between texts and financial risk in order to provide insights and understanding into the impact of soft information in financial reports.

3. Methodology

3.1. Stock return volatility

In finance, *volatility* is a common risk metric defined as the standard deviation of a stock's returns over a period of time. Historical volatilities can be derived from time series of past market

[2] The system is available at http://clip.csie.org/10K/.

[3] Please refer to http://www.mckinsey.com/business-functions/risk/our-insights/ratings-revisited-textual-analysis-for-better-risk-management for more details.

Please cite this article as: M.-F. Tsai, C.-J. Wang, On the risk prediction and analysis of soft information in finance reports, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2016.06.069