RiskFinder: A Sentence-level Risk Detector for Financial Reports

Yu-Wen Liu[†], **Liang-Chih Liu**[*], **Chuan-Ju Wang**[‡], **Ming-Feng Tsai**[†^]
[†]Dept. of Computer Science, National Chengchi University
[*]Dept. of Information and Finance Management, National Taipei University of Technology
[‡]Research Center of Information Technology Innovation, Academia Sinica
[^]MOST Joint Research Center for AI Technology and All Vista Healthcare
g10435@cs.nccu.edu.tw, lcliu@ntut.edu.tw,
cjwang@citi.sinica.edu.tw, mftsai@nccu.edu.tw

Abstract

This paper presents a web-based information system, RiskFinder, for facilitating the analyses of soft and hard information in financial reports. In particular, the system broadens the analyses from the word level to sentence level, which makes the system useful for practitioner communities and unprecedented among financial academics. The proposed system has four main components: 1) a Form 10-K risk-sentiment dataset, consisting of a set of risk-labeled financial sentences and pre-trained sentence embeddings; 2) metadata, including basic information on each company that published the Form 10-K financial report as well as several relevant financial measures; 3) an interface that highlights risk-related sentences in the financial reports based on the latest sentence embedding techniques; 4) a visualization of financial time-series data for a corresponding company. This paper also conducts some case studies to showcase that the system can be of great help in capturing valuable insight within large amounts of textual information. The system is now online available at https://cfda.csie.org/RiskFinder/.

1 Introduction

A great deal of mass media outlets such as newspapers and magazines, or financial reports required by authorities such as the SEC[1]-mandated Form-10Q and Form-10K, play an important role in disseminating information to participants in financial markets. The spread of this information may quickly or slowly influence the sentiment of market participants and thus reshape their perspectives on economic numbers, such as stock prices and interest rate levels. This information comes in two types: soft information, usually referring to textual information such as opinions and market commentary; and hard information, that is, numerical information such as historical time series of stock prices.

Due to the strong relation between the textual information and numerical measures, there has been a growing body of studies in the fields of finance and data science that adopt the techniques of natural language processing (NLP) and machine learning to examine the interaction between these two types of information (e.g., Kogan et al., 2009; Tsai et al., 2016; Tsai and Wang, 2017; Rekabsaz et al., 2017). For example, Loughran and McDonald (2011) and Jegadeesh and Wu (2013) investigate how the disclosures of finance sentiment or risk keywords in SEC-mandated financial reports affect investor expectations about a company's future stock prices. Moreover, Kogan et al. (2009) and Tsai and Wang (2017) exploit sentiment analysis of 10-K reports for financial risk analysis. Furthermore, in Liu et al. (2016), a web-based information system, FIN10K, is proposed for financial report analysis and visualization.

However, these studies and systems all focus on word-level analyses, which likely yield biased results or explanations because the usage of words in finance context, especially in financial reports, is usually complicated and sometimes includes tricks to hide original sentiment (Liu et al., 2016). One of the prominent examples is the negations of positive words to frame negative statements (Loughran and McDonald, 2016); moreover, we also find that the word "offset" is usually used to hide negative information with positive sentiment words. Therefore, to advance the understanding of financial textual information, this paper further constructs an information system based on sentence-level analysis to assist practitioners to capture more precise and meaningful insight within large amounts of textual information in finance.

There have been several studies proposed to produce distributed representations of words, such as

[1] Securities and Exchange Commission

Proceedings of NAACL-HLT 2018: Demonstrations, pages 81–85
New Orleans, Louisiana, June 2 - 4, 2018. ©2018 Association for Computational Linguistics

word2vec (Mikolov et al., 2013). In recent years, there have also been several studies that extend the proportion from word level to sentence, paragraph, or even document level, such as doc2vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017), and Siamese-CBOW (Kenter et al., 2016). Following the fruitful progress of these techniques of word and sentence embeddings, this paper presents a web-based information system, RiskFinder, that broadens the content analysis from the word level to sentence level for financial reports.

The proposed system contains four main parts: 1) a Form 10-K risk-sentiment dataset, consisting of a set of risk-labeled financial sentences and pre-trained sentence embeddings; 2) metadata that summarizes the basic information about each financial report; 3) an interface that highlights risk-related sentences in the financial reports; 4) a visualization of financial time-series data associated with the corresponding financial reports. In the proposed system, we use the 10-K Corpus (Liu et al., 2016) which contains 40,708 financial reports from year 1996 to 2013. In addition to the 10-K corpus, we also construct a set of labeled financial sentences with respect to financial risk by involving 8 financial specialists including accountants and financial analysts to ensure the quality of the labeling. With the labeled sentences and the large collection of financial reports, we apply FastText (Bojanowski et al., 2017) and Siamese-CBOW (Kenter et al., 2016) to sentence-level textual analysis. Due to the superior performance of FastText, the system highlights high risk sentences in those reports via using FastText. For comparison purposes, the system shows the numbers of these highlighted high-risk sentences for the selected report, and we at the same time display the time-aligned relevant quantitative information such as the historical stock prices of the selected company to visualize its financial risk, which considerably facilitates case studies in the field of finance and accounting. Finally, we publish the pre-trained sentence vectors trained on the 10-K corpus and our constructed labeled sentences.

2 System Description

Figure 1 illustrates the user interface of the proposed RiskFinder system. In the system, there are 4 major components for risk detection, the details of which are provided in the following subsections.

2.1 Form 10-K Risk-sentiment Dataset

The first component is the collection of the Form 10-K risk-sentiment dataset. This work provides the financial risk-sentiment dataset that consists of two types of data: a set of risk-labeled financial sentences and the pre-trained sentence embeddings. There are in total 432 labeled financial sentences in the dataset, which were selected from the MD&A[2] sections of the 10-K corpus. When selecting the candidate sentences, we first used the six financial sentiment lexicons proposed by Loughran and McDonald (2011) to filter sentences and randomly chose 24 sentences per year for annotation, yielding in total 432 sentences for 18 years (1996 to 2013). To construct the risk-labeled dataset, eight financial specialists including accountants, financial analysts and consultants participated in the annotation task to ensure the quality of the labeling. In the annotation process, each of the candidate sentences was labeled by three different annotators, and then the rule of majority was used to determine the degree of risk of the sentence. A total of 138 sentences out of the 432 sentences are identified as high-risk sentences, and Cronbach's alpha, which is regarded as an indicator to determine the internal consistency, showed the reliability of 0.784. In addition, the dataset also includes the pre-trained sentence vectors, each of which is a 100-dimension real-valued vector, trained by FastText and Siamese-CBOW on the 10-K corpus.[3]

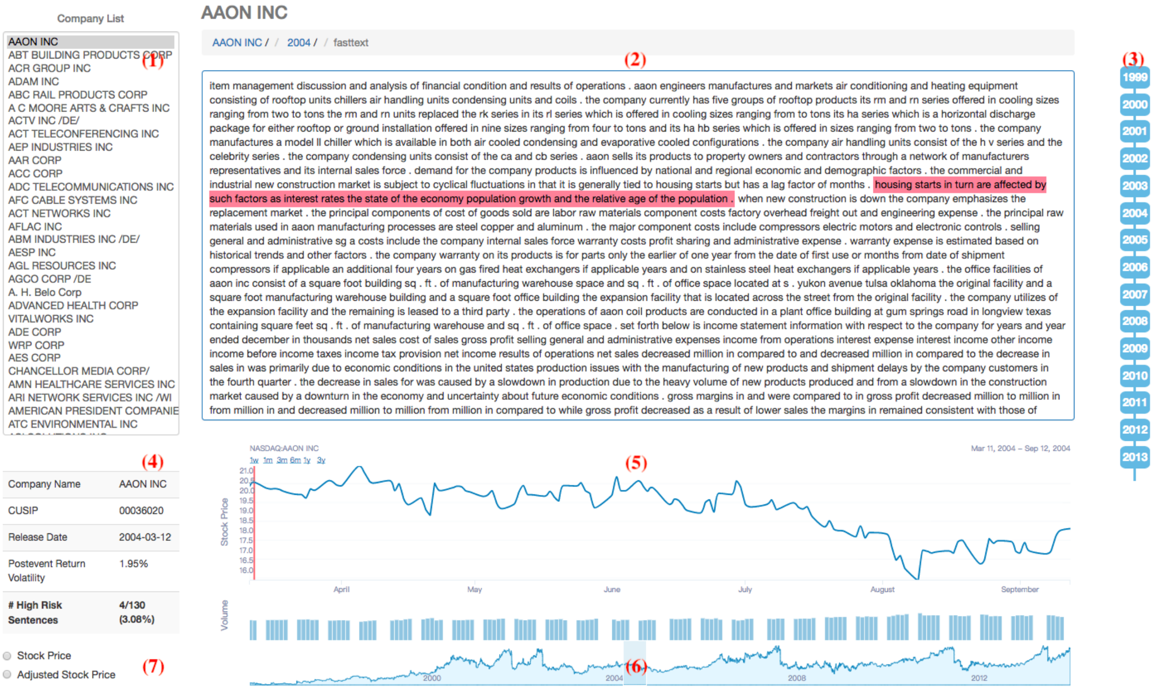
2.2 The Metadata

The second component of the proposed system is the metadata of the companies in the Form 10-K corpus. A user can first select the company in the list (1) (e.g., AEP INDUSTRIES INC in Figure 1). Following the selection, the MD&A section of the report and all fiscal years of the chosen company are then automatically loaded in the window (2) and timeline (3), respectively. A user can select their interesting fiscal year in timeline (3). Then, our system simultaneously displays in table (4) the following metadata which summarizes the basic information of the corresponding company:

1. the company name;
2. the company's CUSIP number, which facilitates the retrieval of information associated

[2] Management Discussion and Analysis of Financial Condition and Results of Operations

[3] The dataset is available at `https://cfda.csie.org/RiskFinder/` upon publication.



Company List

AAON INC
ABT BUILDING PRODUCTS CORP
ACR GROUP INC
ADAM INC
ABC RAIL PRODUCTS CORP
A C MOORE ARTS & CRAFTS INC
ACTV INC /DE/
ACT TELECONFERENCING INC
AEP INDUSTRIES INC
AAR CORP
ACC CORP
ADC TELECOMMUNICATIONS INC
AFC CABLE SYSTEMS INC
ACT NETWORKS INC
AFLAC INC
ABM INDUSTRIES INC /DE/
AESP INC
AGL RESOURCES INC
AGCO CORP /DE
A. H. Belo Corp
ADVANCED HEALTH CORP
VITALWORKS INC
ADE CORP
WRP CORP
AES CORP
CHANCELLOR MEDIA CORP/
AMN HEALTHCARE SERVICES INC
ARI NETWORK SERVICES INC /WI
AMERICAN PRESIDENT COMPANIE
ATC ENVIRONMENTAL INC

AAON INC

AAON INC / / 2004 / / fasttext

item management discussion and analysis of financial condition and results of operations . aaon engineers manufactures and markets air conditioning and heating equipment consisting of rooftop units chillers air handling units condensing units and coils . the company currently has five groups of rooftop products its rm and rn series offered in cooling sizes ranging from two to tons the rm and rn units replaced the rk series in its rl series which is offered in cooling sizes ranging from to tons its ha series which is a horizontal discharge package for either rooftop or ground installation offered in nine sizes ranging from four to tons and its ha hb series which is offered in sizes ranging from two to tons . the company manufactures a model ll chiller which is available in both air cooled condensing and evaporative cooled configurations . the company air handling units consist of the h v series and the celebrity series . the company condensing units consist of the ca and cb series . aaon sells its products to property owners and contractors through a network of manufacturers representatives and its internal sales force . demand for the company products is influenced by national and regional economic and demographic factors . the commercial and industrial new construction market is subject to cyclical fluctuations in that it is generally tied to housing starts but has a lag factor of months . housing starts in turn are affected by such factors as interest rates the state of the economy population growth and the relative age of the population . when new construction is down the company emphasizes the replacement market . the principal components of cost of goods sold are labor raw materials component costs factory overhead freight out and engineering expense . the principal raw materials used in aaon manufacturing processes are steel copper and aluminum . the major component costs include compressors electric motors and electronic controls . selling general and administrative sg a costs include the company internal sales force warranty costs profit sharing and administrative expense . warranty expense is estimated based on historical trends and other factors . the company warranty on its products is for parts only the earlier of one year from the date of first use or months from date of shipment compressors if applicable an additional four years on gas fired heat exchangers if applicable years and on stainless steel heat exchangers if applicable years . the office facilities of aaon inc consist of a square foot building sq . ft . of manufacturing warehouse space and sq . ft . of office space located at s . yukon avenue tulsa oklahoma the original facility and a square foot manufacturing warehouse building and a square foot office building the expansion facility that is located across the street from the original facility . the company utilizes of the expansion facility and the remaining is leased to a third party . the operations of aaon coil products are conducted in a plant office building at gum springs road in longview texas containing square feet sq . ft . of manufacturing warehouse and sq . ft . of office space . set forth below is income statement information with respect to the company for years and year ended december in thousands net sales cost of sales gross profit selling general and administrative expenses income from operations interest expense interest income other income income before income taxes income tax provision net income results of operations net sales decreased million in compared to and decreased million in compared to the decrease in sales in was primarily due to economic conditions in the united states production issues with the manufacturing of new products and shipment delays by the company customers in the fourth quarter . the decrease in sales for was caused by a slowdown in production due to the heavy volume of new products produced and from a slowdown in the construction market caused by a downturn in the economy and uncertainty about future economic conditions . gross margins in and were compared to in gross profit decreased million to million in from million in and decreased million to million from million in compared to while gross profit decreased as a result of lower sales the margins in remained consistent with those of

Company Name	AAON INC
CUSIP	00036020
Release Date	2004-03-12
Postevent Return Volatility	1.95%
# High Risk Sentences	**4/130 (3.08%)**

○ Stock Price
○ Adjusted Stock Price

Figure 1: The user interface of the RiskFinder system

	Siamese-CBOW	FastText
Accuracy	0.656	0.813
Precision	0.776	0.774
Recall	0.438	0.621
F1-measure	0.558	0.684

Table 1: Performance for risk sentence classification

with this company from other widely used databases (e.g., Compustat and CRSP[4]);

3. the report release date;
4. the annualized post-event return volatilities;[5]
5. the number and the percentage of highlighted risk-related sentences in this report.

2.3 Risk-related Sentence Detection

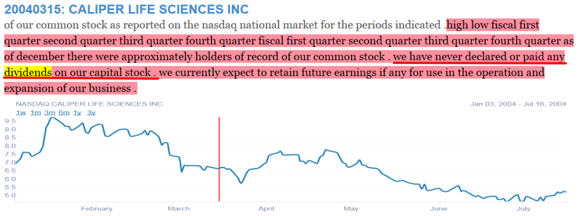
To examine a company's financial risk given the textual information in its financial reports on Form 10-K, the proposed system focuses on the sentence-level investigation into the MD&A section, which is considered a major part where the firm managements are most likely to reveal information (Loughran and McDonald, 2011). To obtain representative sentence representations, we adopt two sentence embedding techniques to conduct sentence-level textual analyses: FastText, a linear classifier that enables efficient sentence classification and yields performance on par with other deep learning classifiers, and Siamese-CBOW, a neural network architecture that obtains word embeddings, directly optimized for sentence representations. Since Siamese-CBOW, the second approach, produces only sentence embedding (by simply averaging the word embedding of each word in a sentence), we then employ a logistic regression to construct a binary classifier to detect high-risk sentences. Table 1 tabulates the performance tested on the 87 sentences in terms of accuracy, precision, recall, and F1-measure with 4-fold cross-validation. Note that the 432 sentences are split into the training data and the testing data with 345 and 87 sentences, respectively, where the proportions of high-risk sentences are kept the same in training and testing sets. Due to the significant superior performance of FastText, we adopt it to highlight high-risk sentences in the reports in the proposed system. Furthermore, the number and percentage of highlighted high-risk sentences in each report are summarized in table (4) of Figure 1.

2.4 Visualization for Financial Measures

Our system attempts to facilitate the analysis of textual information and capture more insight into the financial risk associated with the announcement

[4] CRSP is the abbreviation for the Center for Research in Security Prices.

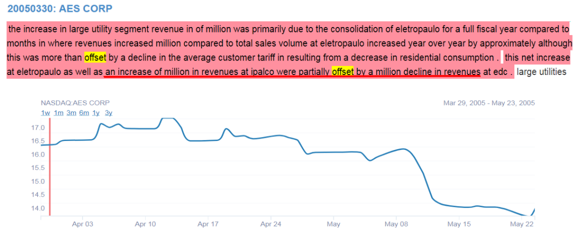
[5] The postevent return volatility is the root-mean square error from a Fama-French three-factor model (Loughran and McDonald, 2011; Tsai et al., 2016).



20040315: CALIPER LIFE SCIENCES INC

of our common stock as reported on the nasdaq national market for the periods indicated .high low fiscal first quarter second quarter third quarter fourth quarter fiscal first quarter second quarter third quarter fourth quarter as of december there were approximately holders of record of our common stock . we have never declared or paid any dividends on our capital stock . we currently expect to retain future earnings if any for use in the operation and expansion of our business .

Figure 2: A positive word in a high-risk sentence



20050330: AES CORP

the increase in large utility segment revenue in of million was primarily due to the consolidation of eletropaulo for a full fiscal year compared to months in where revenues increased million compared to total sales volume at eletropaulo increased year over year by approximately although this was more than offset by a decline in the average customer tariff in resulting from a decrease in residential consumption . this net increase at eletropaulo as well as an increase of million in revenues at ipalco were partially offset by a million decline in revenues at edc . large utilities

Figure 3: The tone transition word

of each financial report. Therefore, in addition to highlight the high-risk sentences, our system also displays time-aligned quantitative information for comparison purposes, such as historical prices and trading volumes of the selected company's stock, as shown in the chart (5) of Figure 1. In particular, the release date of the report is highlighted through a red vertical line in the chart; users can adjust the window (6) to show the corresponding quantitative information for a certain period. Note that two types of historical stock prices are provided in the proposed system: the original stock prices and those adjusted due to stock splits and dividend payouts; the mode can be altered through (7).

3 Case Study

Framing negative or high-risk information through negations of positive financial sentiment words is prevalent among the drafters of financial corpora. Such type of expression could alleviate the adverse effect on market participants' sentiment, but that makes the identification of high-risk statements far more complicated than the discussions in extant finance and accounting literature. From word-level to sentence-level analysis, our proposed RiskFinder system can effectively recognize high-risk information padded with positive words. The positive word "dividend" in stock markets is a notable example; Figure 2 shows that Caliper Life Science Inc.'s stock prices trend downward after the announcement of its year 2004 financial report stating that "we have never declared or paid any dividends on our capital stock . . . ." In a more complex form of the expressions with the positive words "favorable," the system identifies the information "we cannot assure you that these markets will continue to grow or remain favorable to the life science industry . . . " as a high-risk sentence. Moreover, our system also detects the tendency that the drafters of 10-K financial reports use the word "offset" to hide high-risk information behind the low-risk one. For instance, Figure 3 shows that AES CORP.'s stock prices trend downward within 2 months after the announcement of its year 2005 financial report stating that ". . . an increase of million in revenues at ipalco were partially offset by a million decline in revenues . . . ." These examples show that our system can greatly facilitate case studies and be of great help for practitioners in capturing meaningful insight within large amounts of textual information in finance from the sentence-level point of view.

4 Conclusions

In this paper, we introduce RiskFinder to broaden the textual analysis of financial reports from the word level to sentence level. This extension is essential for the analytics of financial reports because the sentences are usually long and complicated, and high-risk information is sometimes hided with positive sentiment words. Built on financial reports in the 10-K Corpus with a set of risk-labeled sentences, this proposed system applies FastText to automatically identify high-risk sentences in those reports. For comparison purposes, the system even displays the time-aligned relevant quantitative information to visualize the financial risk associated with the announcement of each report.

This work is a preliminary study, the purpose of which is to demonstrate the importance of sentence-level analysis and the integration of soft and hard information in finance. Therefore, in the future, we will continue to extend the system, with an emphasis on incorporating more state-of-the-art learning algorithms or developing new algorithms to better detect risk-related sentences. In addition, one of our future work is to obtain more labeled sentences with respect to financial risk or even different financial aspects, which should further enhance the performance and usability of the proposed system.

Acknowledgments

This research was partially supported by the Ministry of Science and Technology of Taiwan under the grants MOST 106-3114-E-009-011 and 107-2634-F-002-011.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Narasimhan Jegadeesh and Di Wu. 2013. Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese cbow: Optimizing word embeddings for sentence representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '16, pages 941–951.

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 the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 272–280.

Yu-Wen Liu, Liang-Chih Liu, Chuan-Ju Wang, and Ming-Feng Tsai. 2016. Fin10k: A web-based information system for financial report analysis and visualization. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2441–2444.

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

Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS '13, pages 3111–3119.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Dür, and Linda Anderson. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '17, pages 1712–1721.

Ming-Feng Tsai and Chuan-Ju Wang. 2017. On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1):243–250.

Ming-Feng Tsai, Chuan-Ju Wang, and Po-Chuan Chien. 2016. Discovering finance keywords via continuous-space language models. *ACM Transactions on Management Information Systems*, 7(3):7:1–7:17.