"The Cat is Out of the Bag" Explainable Risk Ranking with Financial Reports

Ting-Wei Lin,[1] Ruei-Yao Sun,[1] Hsuan-Ling Chang,[2] Chuan-Ju Wang,[3] Ming-Feng Tsai,[1]
[1]Department of Computer Science, National Chengchi University, Taiwan
[2]Department of Finance, National Taiwan University, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
{jacky841114, a0113130}@gmail.com, D05723004@ntu.edu.tw, cjwang@citi.sinica.edu.tw, mftsai@nccu.edu.tw

Abstract

We propose an eXplainable Risk Ranking (XRR) model that uses multilevel encoders and attention mechanisms to analyze financial risks among companies. In specific, the proposed method utilizes the textual information in financial reports to rank the relative risks among companies and locate top high-risk companies; moreover, via attention mechanisms, XRR enables to highlight the critical words and sentences within financial reports that are most likely to influence financial risk and thus boasts better model interpretability. Experimental results evaluated on 10-K financial reports show that XRR significantly outperforms several baselines, yielding up to 7.4% improvement in terms of two ranking correlation metrics. Furthermore, in our experiments, the model interpretability is evaluated by using finance-specific sentiment lexicons at word level and an annotated reference list at the sentence level to examine the learned attention models.

Introduction

Most finance literature on risk analysis has focused on quantitative approaches (Fama and French 1993; Toma and Dedua 2014; Saini and Bates 1984; Aikman et al. 2011). One of the most important works (Fama and French 1993) discovered that the size of a company and its book-to-market ratio are the key factors to financial risk; outside of these two key factors, other factors that may as well affect financial risk are still uncertain. With the progress in text analytics, there have been many studies trying to uncover other potential risk factors by exploiting alternative textual information (e.g., news, reviews, and financial reports) to analyze financial risk (Ding et al. 2015; Rekabsaz et al. 2017; Nopp and Hanbury 2015).

Due to the noise within finance documents and the information gap between texts and financial numerical measures, it is difficult to predict the exact finance quantities (e.g., stock return and volatility) and to extract useful information and relations directly by using textual information. Thus, the work in (Tsai and Wang 2016) proposes using ranking-based methods for analyzing financial risk with the use of textual information and shows that ranking-based methods are more suitable than regression-based methods for such an analytic task. However, the work in (Tsai and Wang 2016) and other pioneering studies such as (Kogan et al. 2009; Schumaker and Chen 2009) mainly use simple and handcrafted features to describe financial documents, like bags-of-words, noun phrases, and named entities. Thus, these approaches are difficult to model complex structures or semantics in texts, which limits their potential and usage scenarios.

In recent years, deep neural networks such as CNN (LeCun et al. 1998), GRU (Chung et al. 2014), and BERT (Devlin et al. 2018) have demonstrated promising results across NLP tasks such as document classification and sentiment analysis (Dos Santos and Gatti 2014; Akhtar et al. 2017). The advancements are due to the superiority of these techniques in learning semantically meaningful representations. Although such deep learning approaches can extract the latent features from texts, most of these models are not interpretable, which is however a vital ingredient in models for finance applications. To some extent, attention mechanisms alleviate the interpretability problem, as attention layers explicitly weight the components' representations. Thus, it is often undoubted that attention mechanisms can identify information that models find important.

To advance the state of the art, we propose an eXplainable Risk Ranking model (XRR) to capture key information from financial reports and investigate related financial risks. Specifically, XRR is a deep neural network model incorporating multilevel explainable structures and learning to rank techniques for ranking relative risks defined by post-event return volatility (Loughran and McDonald 2011) among companies. To build the XRR model, we first design a multilevel explainable structure to model the complex structures within financial texts by using sequence encoders based on bidirectional gated recurrent units (GRUs) at both the word and sentence levels. At each level, the attention mechanism is leveraged to make the model explainable. Moreover, unlike many previous hierarchical deep neural network architectures, which are mainly on classification tasks (Ding et al. 2015; Luo et al. 2018), XRR ranks the relative risks among companies and locates top high-risk companies. To enable this, we propose a pairwise ranking loss based on a siamese network with two parallel multilevel explainable structures. In addition, we propose using the post-event return volatility as the proxy of financial risk because it excludes the effect of several important macro-economic

factors and is effective for monitoring the event effect on the change of stock prices (Loughran and McDonald 2011; Tsai, Wang, and Chien 2016).

We conduct comprehensive experiments using a large collection of 10-K financial reports from 1996 to 2013, consisting of 39,083 reports in total. The results show that the proposed XRR significantly outperforms other baselines in terms of all evaluation metrics. For robustness, we also conduct a comparison on different financial risk proxies and conduct several financial analyses to verify our results. Moreover, we conduct evaluation and discussion by using external finance-specific sentiment lexicons and an annotated reference list at the sentence level to examine the learned financial sentiment texts with high attention scores and the corresponding financial risks. In this evaluation, XRR exhibits a stronger retrieval power compared to the baselines and provides more insightful understanding into the impact of the financial texts on companies' future risks. In summary, XRR advances the state of the art in the following four dimensions.

1. We propose a deep neural network architecture for risk ranking with financial reports, allowing for modeling financial texts with more complex structures than those traditional non-neural models.
2. With the multilevel attention mechanism, the proposed model is explainable at both the word and sentence levels, the ability of which is essential for finance applications.
3. We propose using the post-event return volatility as a risk proxy for such text analytic tasks, and our experiments also attest the appropriateness of the proxy for the tasks.
4. We conduct extensive experiments and analyses on a large collection of financial reports, the results of which attest the effectiveness of the proposed method in terms of both ranking performance and interpretability.

Methodology

We first formulate the risk ranking problem, and then provide an brief description of the post-event return volatility. Finally, we describe the proposed XRR model in detail.

Definitions and Problem Formulation

We rank the companies along with their relative financial risks with the use of companies' associated textual information via a pairwise ranking model. Note that we here use the post-event return volatility as a proxy of financial risk for each company. Following the work in (Tsai and Wang 2016), we slot the volatilities within a year into several risk levels; thus, each company c_i corresponds to a risk level $v_i \in \mathbb{Z}$. Given a collection of financial reports $\mathcal{D}$, we generate a set of pairs of financial reports $\{(d_\ell, d_j) | d_\ell, d_j \in \mathcal{D}\}$, each element in which corresponds to a pair of financial reports for two companies c_ℓ and c_j. We thus have the pairwise risk model $f : \mathbb{R}^p \to \mathbb{R}$ for comparison between companies c_ℓ and c_j such that

$$E(d_\ell, d_j) = \mathbb{1}_{\{v_\ell > v_j\}}, \tag{1}$$

where v_i denotes the risk level of company c_i and p denotes the dimension of the representation of a report d_i. Note that the rank order of the set of companies is specified by the real score that the model f takes. In particular, $f(\mathbf{d}_\ell) > f(\mathbf{d}_j)$ is taken to mean that the model asserts that $c_\ell \succ c_j$, where $\mathbf{d}_i \in \mathbb{R}^p$ denotes the representation of report d_i and $c_\ell \succ c_j$ means that c_ℓ is ranked higher than c_j; that is, the company c_ℓ is riskier than c_j.

Post-event Return Volatility

Post-event volatility has been widely used as a proxy of financial risk in finance research, especially in the case of event study (Ito, Lyons, and Melvin 1998). In contrast to the naive stock return volatility, which is defined as the standard deviation of the daily stock returns over a certain period, post-event volatility calculation takes into account macro-economic factors; thus, such a measure excludes the effect of these macro-economic factors and is effective for monitoring the event's effect on the change of stock prices. As a result, for event study, it is considered a more suitable risk proxy than the naive stock return volatility, though many data mining works adopt the naive stock return volatility to conduct the analysis. Note that in the above context, "event" refers to the filing of a financial report.

Following the definition in (Loughran and McDonald 2011; Tsai, Wang, and Chien 2016), we define the post-event return volatility as the root-mean-square error from a Fama and French three-factor model (Fama and French 1993) for days [6, 252] after the event and at least 60 daily observations. Then, we focus on modeling the effect on the post-event return volatility of a company after its report filing. For comparison purposes, we also include the results of naive stock return volatility in the Experiments section.

Multilevel Explanation Structure

Inspired by several hierarchical language networks (Yang et al. 2016; Hu et al. 2018; Ding et al. 2015), we construct XRR, our pairwise risk ranking model, using a multilevel structure to represent pairs of financial reports. The structure is mainly made of a word-level embedding matrix and two major components at both word and sentence levels: the GRU sequence encoder and the multilevel attention mechanism (see Figure 1).

Embedding Matrix Given the set of word vocabulary $\mathcal{W}$, we embed each word $w \in \mathcal{W}$ into a real-valued vector x through a embedding matrix $W_e \in \mathbb{R}^{|\mathcal{W}| \times m}$, where m is the dimension of word vectors.

GRU Sequence Encoder Given a report $d \in \mathcal{D}$ with L sentences $\{s_1, s_2, \dots s_L\}$, s_t denotes the embedded representation of the t-th sentence. In each report, the t-th sentence consists of l words $\{w_{t1}, w_{t2}, \dots, w_{tl}\}$, where $w_{ti} \in \mathcal{W}$. To encode both sentences and documents, we adopt bidirectional GRUs at both the word and sentence level, respectively, which leverage past and future information to better utilize contextual finance information. Generally speaking, in the sentence encoder, for the ℓ-th word in the t-th sentence, $w_{t\ell}$, with its corresponding word embedding $x_{t\ell}$ from W_e, the word can be depicted by concatenating the forward hidden state $\overrightarrow{h}_{t\ell}$ and the backward one $\overleftarrow{h}_{t\ell}$ of the GRU encoders; that is, the annotation of the ℓ-th word in the t-th sentence

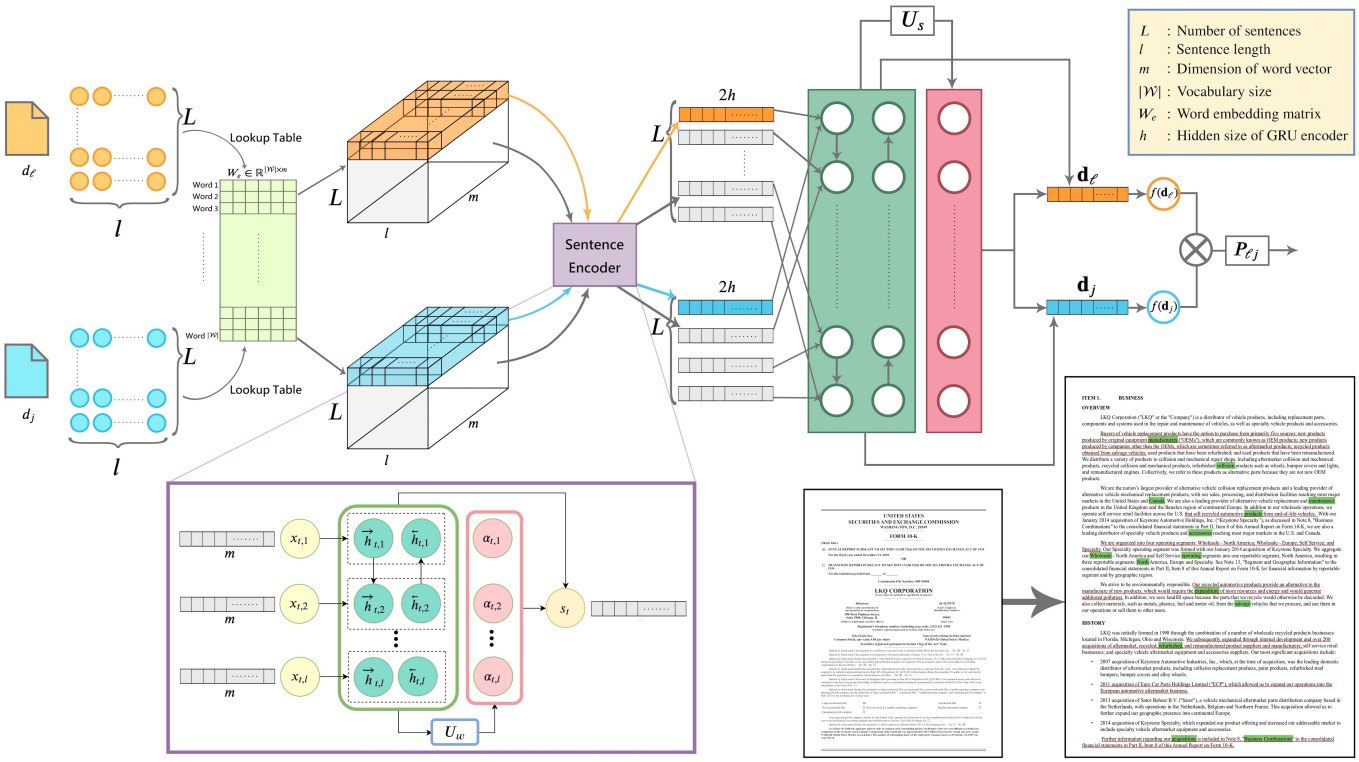


Figure 1: XRR network structure

becomes

$$h_{t\ell} = \overrightarrow{h}_{t\ell} \oplus \overleftarrow{h}_{t\ell} = \overrightarrow{\text{GRU}}(x_{t\ell}) \oplus \overleftarrow{\text{GRU}}(x_{t\ell}),$$

for $\ell = 1, 2, \dots l$, where $\overrightarrow{h}_{t\ell}, \overleftarrow{h}_{t\ell} \in \mathbb{R}^h$, $\oplus$ denotes the concatenation operator, and h refers to the hidden size of a GRU encoder. Then, we have $h_{t\ell} \in \mathbb{R}^{2h}$ and $H_w = (h_{t1}, \cdots, h_{tl}) \in \mathbb{R}^{l \times 2h}$.

Following the same process, in the document encoder, the t-th sentence is represented by concatenating the forward hidden state $\overrightarrow{h}_t$ and the backward one $\overleftarrow{h}_t$, i.e.,

$$h_t = \overrightarrow{h}_t \oplus \overleftarrow{h}_t = \overrightarrow{\text{GRU}}(s_t) \oplus \overleftarrow{\text{GRU}}(s_t),$$

for $t = 1, 2, \dots L$, where $\overrightarrow{h}_t, \overleftarrow{h}_t \in \mathbb{R}^h$. Then we have $h_t \in \mathbb{R}^{2h}$ and $H_s = (h_1, \cdots, h_L) \in \mathbb{R}^{L \times 2h}$.

Multilevel Attention Mechanism To provide fine-grained explainable results, the proposed XRR involves one level of attention at the word level and one at the sentence level; these pay more or less attention to individual words and sentences and capture influential texts in financial reports with respect to financial risks. Specifically, for the t-th sentence, we feed each word annotation $h_{t\ell}$ through a fully-connected layer to yield $u_{t\ell}$ as the hidden representation of $h_{t\ell}$, after which the attention mechanism measures the importance of the hidden representation $u_{t\ell}$ with a word level context vector U_w and obtains a normalized importance weight $\alpha_{t\ell}$ through a softmax function. After that, we compute the sentence vector s_t as a weighted sum of the word annotations. Mathematically speaking, we have

$$u_{t\ell} = \tanh\left(W_w h_{t\ell} + b_w\right), \ \ell = 1, 2, \dots l$$

$$\alpha_{t\ell} = \frac{\exp(u_{t\ell}^\top U_w)}{\sum_{i=1}^{l} \exp(u_{ti}^\top U_w)}, \ \ell = 1, 2, \dots l$$

$$s_t = \sum_{\ell=1}^{l} \alpha_{t\ell} h_{t\ell},$$

where $W_w \in \mathbb{R}^{a \times 2h}$, $b_w \in \mathbb{R}^a$, and $U_w \in \mathbb{R}^a$.

Similar to the above procedure, we feed the hidden representation of each sentence annotation h_t by using a single-layer perceptron to get u_t, which is associated with a normalized importance weight α_t via a sentence level context vector U_s, i.e.,

$$u_t = \tanh\left(W_s h_t + b_s\right), \ t = 1, 2, \dots L$$

$$\alpha_t = \frac{\exp(u_t^\top U_s)}{\sum_{i=1}^{L} \exp(u_i^\top U_s)}, \ t = 1, 2, \dots L$$

where $W_s \in \mathbb{R}^{a \times 2h}$, $b_w \in \mathbb{R}^a$, and $U_s \in \mathbb{R}^a$.

Finally, with the weight vector α_t for $t = 1, \cdots, L$, the representation of each report $d_i \in \mathcal{D}$, $\mathbf{d}_i$, is computed as a weighted sum of the sentence annotations as

$$\mathbf{d}_i = \sum_{t=1}^{L} \alpha_t h_t. \tag{2}$$

Pairwise Deep Ranking

We use a pairwise approach to rank the financial reports according to their financial risk levels. To this end, we build a

pair of multilevel structures described in the previous subsection, with the weights shared across both sides of the structures, as illustrated in Figure 1. Given a pair of financial reports (d_ℓ, d_j), where the company associated with d_ℓ is riskier than that with d_j according to their risk levels, the goal of the ranking model $f(\cdot)$ is to generate a higher score for d_ℓ. Denote $\Psi = \{(d_\ell, d_j) \mid E(d_\ell, d_j) = 1\}$ as the set of all "positive" pairs, each element in which is fed into two separate but identical hierarchical structures. Our goal is to learn a score function $f(\cdot)$ that satisfies

$$f(\mathbf{d}_\ell) > f(\mathbf{d}_j), \forall (d_\ell, d_j) \in \Psi, \tag{3}$$

where $\mathbf{d}_i$ denotes the dense representation of report d_i obtained from Eq. (2). Note that in practice, we implement a siamese network for $f(\cdot)$ that adopts the same weights while working in tandem on two different input vectors to compute comparable output vectors.

To obtain an overall risk ranking for all companies (reports), we adopt a standard RankNet (Burges et al. 2005) loss layer to learn a posterior probability distribution $P_{\ell j}$ that is *close* to the target probability $E(d_\ell, d_j)$ defined in Eq. (1) for each pair (d_ℓ, d_j), where

$$P_{\ell j} = \frac{\exp\left(f(\mathbf{d}_\ell) - f(\mathbf{d}_j)\right)}{1 + \exp\left(f(\mathbf{d}_\ell) - f(\mathbf{d}_j)\right)}.$$

A natural choice for measuring the closeness between two probability distributions is binary cross-entropy; thus we have the objective function to be minimized as

$$\min - \sum_{(d_\ell, d_j) \in \Psi} \left(E(d_\ell, d_j) \log P_{\ell j} + \left(1 - E(d_\ell, d_j)\right) \log\left(1 - P_{\ell j}\right)\right).$$

Experiments

Data Description

We conducted experiments on a large collection of 10-K reports from year 1996 to year 2013 provided by (Tsai, Wang, and Chien 2016), which are annual reports required by the Securities and Exchange Commission (SEC) providing comprehensive overviews of companies' business and financial conditions and which include audited financial statements. Specifically, following previous studies in (Kogan et al. 2009; Tsai and Wang 2016; Tsai, Wang, and Chien 2016; Buehlmaier and Whited 2018), we used only Section 7 "Management's Discussion and Analysis of Financial Conditions and Results of Operations" (MD&A) in the experiments as it contains the most important forward-looking statements about the companies.

Experimental Settings

We first split the post-event return volatilities of companies within a year into five different risk levels[1] and generated a set of pairs of financial reports based on the relative difference of levels among the companies. Due to the huge numbers of document pairs, we sampled 3,000 pairs to train the model in each epoch; moreover, we differentiated the pair sampling probabilities based on their degree of proximity to the testing year; that is, pairs closer to the testing year were given a higher sampling probability. In addition, the dimension of the word vector, m, depended on the pre-trained word embedding models used, the hidden size of the GRU (h) is set to 100, and the attention size (a) is set to 100. The maximum number of words in sentences (l) and that of sentences in documents (L) were set to 150 and 70, respectively. The values of the model hyperparameters for the compared method were decided using a grid search over different settings; we used the combination that led to the best performance.

Pre-trained Word Embedding

We evaluated different word embedding models to construct the pre-trained word embedding matrix W_e.

1. **Fin-Word2Vec** (Tsai, Wang, and Chien 2016) denotes vectors pre-trained via Word2Vec with a skip-gram model trained on the 10-K Corpus (39083 reports from 18 years); each word is represented as a 300-dimensional vector.
2. **BERT-Large, Uncased** (Devlin et al. 2018) contains 24-layer, 1024-hidden, 16 heads, and 340M parameters; each word in a document is represented by a 1024-dimensional vector, and only the word embedding is used in our model.[2]
3. **GloVe** (Pennington, Socher, and Manning 2014) representations are 300-dimensional word vectors[3] trained on 840 billion tokens of Common Crawl data.

In the following experiments, we denote each word embedding model with the first character of its name (i.e., F, B, G) with parentheses, e.g., XRR (B) for XRR with BERT-Large.

Compared Methods

We compare the proposed XRR with several baseline models including a ranking-based and two multi-class classification models.

1. **RankSVM**[4] is used in (Tsai and Wang 2016), which adopts ranking SVM with TF-IDF of words as features, where IDF is computed from the documents in a single year as the document frequency of a specific word may vary across different years.
2. **FastText** is proposed by (Grave et al. 2017), a simple and efficient baseline for document classification.
3. **HAN** is proposed by (Yang et al. 2016), adopting hierarchical networks with attention mechanisms for document classification. For HAN, we used GloVe as the pre-trained word embedding and sorted the companies using the probabilities of the high-risk class in the softmax layer.

[1] We here split the volatilities based on 30-th, 60-th, 80-th, and 90-th percentiles, yielding the average numbers of the five categories per year as 702, 702, 467, 234, and 234, respectively.

[2] Note that in BERT models, words in different sentences (or documents) are associated with different representations; to reflect this, we treat words in different documents as different words.

[3] https://nlp.stanford.edu/projects/glove/

[4] The linear kernel was adopted with $C = 1$; all other parameters were left at the default SVMRank values.

Metric	Method	Model \ Test year	2001	2002	2003	...	2010	2011	2012	2013	Average
τ	Classification	Fasttext	0.475	0.388	0.401	...	0.449	0.460	0.452	0.463	0.426
		HAN	0.527	0.474	0.582	...	0.557	0.569	0.590	0.593	0.535
	Ranking	RankSVM	0.549	0.521	0.525	...	0.589	0.592	0.593	0.591	0.547
		XRR (G)	0.536	0.501	0.502	...	0.580	0.607	0.623	0.607	0.547
		XRR (B)	0.541	0.525	0.518	...	0.591	0.616	0.632	0.625	0.559
		XRR (F)	**0.570**	**0.541**	**0.553**	...	**0.605**	**0.616**	**0.637**	**0.629**	**0.573***
ρ	Classification	Fasttext	0.589	0.493	0.506	...	0.573	0.583	0.568	0.585	0.540
		HAN	0.648	0.587	0.599	...	0.690	0.702	0.720	0.727	0.661
	Ranking	RankSVM	0.685	0.657	0.661	...	0.733	0.733	0.731	0.732	0.686
		XRR (G)	0.671	0.632	0.636	...	0.720	0.750	0.762	0.748	0.684
		XRR (B)	0.675	0.659	0.657	...	0.732	0.756	0.772	0.766	0.697
		XRR (F)	**0.702**	**0.675**	**0.691**	...	**0.749**	**0.760**	**0.773**	**0.768**	**0.711***

Notation $*$ denotes significance compared to the best baseline under a permutation test with $p < 0.05$.

Table 1: Performance comparison

Evaluation Metrics

To evaluate the performance of our model, we first adopted two common rank correlation metrics: Spearman's Rho (ρ) (Myers, Well, and Lorch 2003) and Kendall's Tau (τ) (G. 1938). Given two ranked lists $X = \{x_1, x_2, \dots x_n\}$ and $Y = \{y_1, y_2, \dots y_n\}$, the formula of these two metrics are defined as

$$\rho = 1 - \frac{6\sum(x_\ell - y_\ell)^2}{n(n^2-1)}$$

$$\tau = \frac{\text{\#concordant pairs} - \text{\#discordant pairs}}{n(n^2-1)}$$

For the measure of Kendall's Tau, any pair of observations (x_ℓ, y_ℓ) and (x_j, y_j) is concordant if the ranks for both elements agree; that is, if both $x_\ell \succ x_j$ and $y_\ell \succ y_j$ or if both $x_j \succ x_\ell$ and $y_j \succ y_\ell$. In contrast, it is discordant if $x_\ell \succ x_j$ and $y_j \succ y_\ell$ or if $x_j \succ x_\ell$ and $y_\ell \succ y_j$. If $x_\ell = x_j$ or $y_\ell = y_j$, the pair is neither concordant nor discordant.

Experimental Results

Table 1 tabulates the experimental results, in which all reports from the five-year period preceding the testing year are used as the training data. For example, the reports from 1996 to 2000 constitute the training data, and the trained model is tested on the reports of year 2001. The boldface number in the table denotes the best result among all methods per test year. As shown in the table, the proposed XRR reveals the strong correlations in terms of the two metrics between the predicted financial risk levels and the actual levels. We attribute the superior performance of XRR to the following observations: 1) The RankSVM and XRR ranking-based methods successfully identify relative risks between each financial document pair and yield better performance than the two classification models; 2) XRR models a much more complex structure of representations of financial texts than the traditional bag-of-words model, yielding better performance than RankSVM+TF-IDF.

In addition, we compare the proposed XRR using different pre-trained word embeddings. The results show that XRR (F), the model with Fin-Word2Vec, yields consistently better performance than those with GloVe or BERT. A closer look at the results shows that although XRR with BERT yields better results than that with GloVe, the model using a domain-specific word embedding, i.e., XRR (F), still achieves the best performance among the three. This demonstrates that a high-quality, domain-specific word embedding is also an important factor for such a task.[5]

On the other hand, while correctly ranking all reports along with their financial risk is important, financial scholars and practitioners may care more about locating the most risky companies. To examine this type of performance,[6] we further use the concepts of precision@K and recall in information retrieval as our evaluation metrics, where we use the realized post-event volatilities to rank the companies in each year and treat the top-K companies as our ground truth when calculating precision. In addition, in terms of recall, we take the companies with the highest risk levels as the ground truth. As shown in Figure 2, our method outperforms both RankSVM and HAN in terms of these two metrics, indicating that the proposed XRR is more effective at locating high-risk companies than the other two methods. Note that in the following subsections, we use the results of XRR (F), the best model, for further analyses and interpretability discussion. We also omit the notation denoting the pre-trained word embedding, i.e., "(F)", to simplify the notation.

Fine-grained Analysis

We here conduct a fined-grained analysis to further investigate the performance of companies associated with different risk levels. To do so, we first equally split the companies within a year into five different risk levels according to their realized post-event return volatilities; we then calculate the

[5] Due to resource limitations, we could not train a domain-specific BERT model; however, we speculate that using a domain-specific BERT would yield further improvements.

[6] We omit the comparison to Fasttext here as its performance in Table 1 distances it from the other three models.

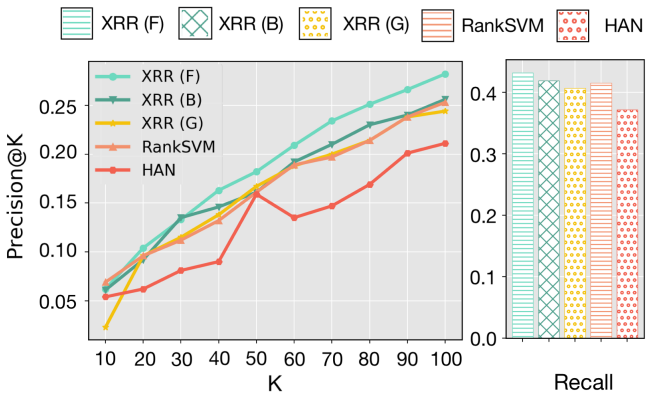


Figure 2: Evaluation on high-risk companies

Variables	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Firm Size	8.5052	7.8410	6.9821	6.1892	5.7281

Table 2: Firm size analysis

τ and ρ correlation metrics for companies in each rank. As shown in the heat map in Figure 3, where the color denotes the correlation, the proposed model yields better performance for companies with higher financial risk, which shows that the model effectively locates high-risk companies, thus making our approach useful in practice.

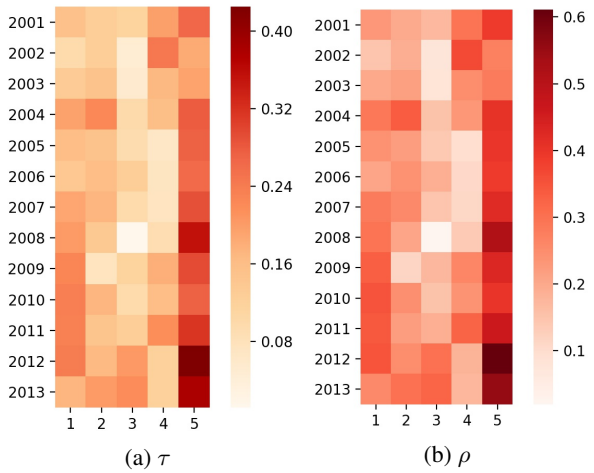
Also, we investigate the relation between the predicted risk levels and the average firm size[7] of the companies at each risk level. According to (Fama and French 1993), smaller firms are typically associated with higher financial risk than larger ones. To examine the rationality of our prediction, we equally split the firms based on our predicted scores in each year into five risk levels and calculate the average firm size separately in each of the five groups. Table 2 shows that the predicted high-risk companies (Rank 5) are on average small in terms of their firm size, which indicates that our model learned from textual information from financial reports yields findings consistent with the literature in finance.

Different Risk Measure Analysis

To demonstrate the suitability of using post-event return volatility as our risk proxy, we compare its performance with the naive stock volatility in Figure 4. The definition of the naive stock volatility is the standard deviation of stock returns[8] over a certain period. Following the setting in (Tsai, Wang, and Chien 2016), we choose daily stock returns for 12 months after the report filing date to calculate the naive stock return volatility. In Figure 4 we observe that the correlations between the predicted risk scores and post-event volatilities are much higher than those between the predicted scores and the naive stock return volatility. This is because the naive stock return volatility is a noisy risk proxy for pure textual

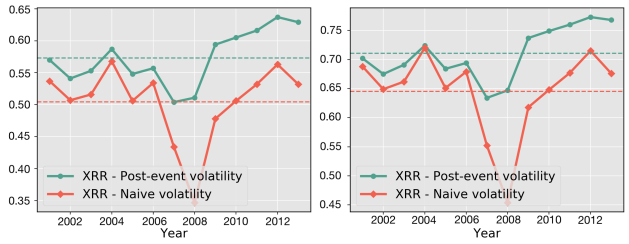
[7] The firm size is defined as the logarithm of the sum of all current and long-term assets held by a company (in million dollars).

[8] The stock return is the appreciation in the price plus any dividends paid, divided by the original price of the stock.



(a) τ (b) ρ

Figure 3: Fine-grained correlation analysis



(a) τ (b) ρ

Figure 4: Comparison of different volatility measures

analysis, as it does not exclude other macro-economic or human behavior risk, making it difficult for models to capture the relation between text and risk. One obvious case in year 2008, the well-known financial crisis, shows that the naive stock return volatility was drastically affected by the market, causing its lowest correlation of the whole sample period.

Discussions on Interpretability

Financial Sentiment Terms Analysis

We evaluate the word attention mechanism of XRR and HAN by using the finance-specific sentiment lexicon (FL) proposed by (Loughran and McDonald 2011), which consists of the following six word lists:[9]

1. **Fin-Neg**: negative business terminologies (e.g., deficit)
2. **Modal**: words expressing different levels of confidence (e.g., could, might).
3. **Fin-Pos**: positive business terminologies (e.g., profit)
4. **Fin-Unc**: words denoting uncertainty, with emphasis on the general notion of imprecision rather than exclusively focusing on risk (e.g., appear, doubt).
5. **Fin-Con**: words denoting constraining, a factor that restricts the amount or quality of investment options (e.g., prevent, limit).
6. **Fin-Lit**: words reflecting a propensity for legal contest or, per our label, litigiousness (e.g., amend, forbear).

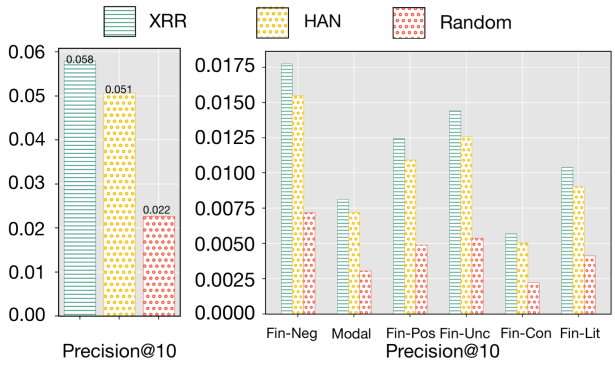


Figure 5: Word attention analysis

We first rank the terms in each sentence according to their learned attention weights and use the top-10 terms to conduct the evaluation. The left panel in Figure 5 plots the precision@10 for each method, for which the terms in the union of the six word lists are considered as the ground truth. Observe that compared to the other two methods, XRR captures more terms listed in the lexicon; note that Random denotes the methods that randomly select 10 terms from each sentence. In addition, in the right panel of Figure 5, we conduct a finer analysis by treating the words in each word list as the ground truth. An interesting finding is that XRR locates more negative words in **Fin-Neg** than the other two methods. Previous literature shows that negative terms are usually highly correlated with financial risk (Loughran and McDonald 2011; Tsai and Wang 2016). For instance, *deficit* usually means "an excess of liabilities over assets, of losses over profits, or of expenditure over income in finance;" it is clear that a company's report that is highly associated with *deficit* usually implies higher future risk. This finding shows that the proposed model is consistent with many previous findings and highlights negative financial words more than other models.

Financial Sentiment Sentences Analysis

We further use an annotated list at the sentence level to analyze the results of sentence-level attention mechanisms in XRR. The reference list contains 2,432 sentences labeled as risk-related ones. In particular, there are 1,539 high risk-related sentences and 896 low risk-related ones, each of which is selected from the MD&A sections of the used 10-K dataset.[10]

For evaluation, we treat the 1,539 high risk-related sentences in financial reports as our ground truth. In each financial report containing at least one high-risk labeled sentence, we rank all of the sentences according to their learned attention weights and use the top-10 sentences to conduct the evaluation in terms of precision and recall. As shown in Figure 6, the XRR model is generally capable of highlighting more risky sentences in terms of both metrics; note that the dotted lines in the figure denote the average performance over different years. These results again demonstrate that the

[9] https://sraf.nd.edu/textual-analysis/resources/

[10] The list will be publicly available upon publication.

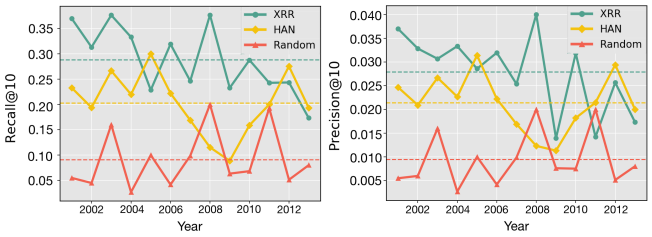


Figure 6: Sentence attention analysis

sentence-level attention weights of XRR reveal a stronger and a more straightforward relation between texts and financial risk than other models.

Furthermore, we provide two example sentences that are associated with high attention scores in Figure 7, where that in (a) is in the annotated list and its attention weight is four times the average attention weight of sentences in the reports associated with the highest risk level. Also, our model also identifies a non-labeled sentence (b) as a high weighted sentence in which the terms "redeem" and "loss" are both associated with negative effects for the company and might bring uncertainty and risk in the future. Such results demonstrate that the XRR model effectively finds the important parts within a document. Therefore, considering financial scholars and practitioners' concerns about risky information in financial reports, these examples indicate that our model spotlights texts that are highly correlated to high risk in financial reports and effectively provides the important parts within a document as a brief summary thereof.

(a)	**From Chromcraft Revington, Inc., From 10-K** Forward-looking statements are not guarantees of performance or outcomes and are subject to certain risks and uncertainties that could cause actual results or outcomes to differ materially from those reported, expected or anticipated as of the day of this report.
(b)	**From Timberland BanCorp, Inc., 2008 Form 10-K** In June 2008, the Company redeemed its $ 29.1 million investment in the AMF family of mutual funds for the underlying securities and cash, and recorded a loss of $2.8 million.

Figure 7: Examples of sentence attention

Conclusion

In this paper, we propose XRR to rank companies to keep them in line with their relative risk levels specified by their post-event volatilities, in which the textual information in financial reports is leveraged to make the prediction. Experimental results on a real-world financial report dataset demonstrate that our approach exhibits a stronger ranking power compared to the baselines. Furthermore, the evaluation on interpretability also attests the effectiveness of our model for providing explainable results.

References

Aikman, D.; Alessandri, P.; Eklund, B.; Gai, P.; Kapadia, S.; Martin, E.; Mora, N.; Sterne, G.; Willison, M.; et al. 2011. Funding liquidity risk in a quantitative model of systemic stability. *Central Banking, Analysis, and Economic Policies Book Series* 15:371–410.

Akhtar, M. S.; Kumar, A.; Ghosal, D.; Ekbal, A.; and Bhattacharyya, P. 2017. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proc. EMNLP*, 540–546.

Buehlmaier, M. M., and Whited, T. M. 2018. Are financial constraints priced? evidence from textual analysis. *The Review of Financial Studies* 31(7):2693–2728.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. N. 2005. Learning to rank using gradient descent. In *Proc. ICML*, 89–96.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, 4171–4186.

Ding, X.; Zhang, Y.; Liu, T.; and Duan, J. 2015. Deep learning for event-driven stock prediction. In *Proc. IJCAI*, 2327–2333.

Dos Santos, C., and Gatti, M. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proc. COLING*, 69–78.

Fama, E. F., and French, K. R. 1993. Common risk factors in the returns on stocks and bonds. *Journal of financial economics* 33(1):3–56.

G., K. M. 1938. A new measure of rank correlation. *Biometrika* 30(1/2):81–93.

Grave, E.; Mikolov, T.; Joulin, A.; and Bojanowski, P. 2017. Bag of tricks for efficient text classification. In *Proc. EACL*, 427–431.

Hu, Z.; Liu, W.; Bian, J.; Liu, X.; and Liu, T.-Y. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proc. WSDM*, 261–269.

Ito, T.; Lyons, R. K.; and Melvin, M. T. 1998. Is there private information in the fx market? the tokyo experiment. *The Journal of Finance* 53(3):1111–1130.

Kogan, S.; Levin, D.; Routledge, B. R.; Sagi, J. S.; and Smith, N. A. 2009. Predicting risk from financial reports with regression. In *Proc. NAACL*, 272–280.

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Loughran, T., and McDonald, B. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance* 30(1):81–93.

Luo, L.; Ao, X.; Pan, F.; Wang, J.; Zhao, T.; Yu, N.; and He, Q. 2018. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *Proc. IJCAI*, 4244–4250.

Myers, J. L.; Well, A.; and Lorch, R. F. 2003. Research design and statistical analysis. *Lawrence Erlbaum* 30.

Nopp, C., and Hanbury, A. 2015. Detecting risks in the banking system by sentiment analysis. In *Proc. EMNLP*, 591–600.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Proc. EMNLP*, 1532–1543.

Rekabsaz, N.; Lupu, M.; Baklanov, A.; Hanbury, A.; Dür, A.; and Anderson, L. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. *arXiv preprint arXiv:1702.01978*.

Saini, K. G., and Bates, P. S. 1984. A survey of the quantitative approaches to country risk analysis. *Journal of Banking & Finance* 8(2):341–356.

Schumaker, R. P., and Chen, H. 2009. Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)* 27(2):12.

Toma, A., and Dedua, S. 2014. Quantitative techniques for financial risk assessment: A comparative approach using different risk measures and estimation methods. *Procedia Economics and Finance* 8:712–719.

Tsai, M.-F., and Wang, C.-J. 2016. On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research* 257(1):243–250.

Tsai, M.-F.; Wang, C.-J.; and Chien, P.-C. 2016. Discovering finance keywords via continuous-space language models. *ACM Transactions on Management Information Systems (TMIS)* 7(3):7.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *Proc. NAACL*, 1480–1489.