Interpretable Operational Risk Classification with Semi-Supervised Variational Autoencoder

Fan Zhou[1], **Shengming Zhang**[1], **Yi Yang**[2]
[1]University of Electronic Science and Technology of China.
[2]Hong Kong University of Science and Technology.
fan.zhou@uestc.edu.cn, shmizhang@gmail.com, imyiyang@ust.hk

Abstract

Operational risk management is one of the biggest challenges nowadays faced by financial institutions. There are several major challenges of building a text classification system for automatic operational risk prediction, including imbalanced labeled/unlabeled data and lacking interpretability. To tackle these challenges, we present a semi-supervised text classification framework that integrates multi-head attention mechanism with Semi-supervised variational inference for Operational Risk Classification (SemiORC). We empirically evaluate the framework on a real-world dataset. The results demonstrate that our method can better utilize unlabeled data and learn visually interpretable document representations. SemiORC also outperforms other baseline methods on operational risk classification.

1 Introduction

In the decade since the global financial crisis, banks and regulators have become increasingly alert to operational risks (OR). However, the banks still struggle to deal with operational risks effectively (Hoffman, 2002). It is reported that major banks global wide have suffered nearly $210 billion in operational risk losses since 2011 [1]. Operational risks refer to the risks of loss due to errors, breaches, interruptions or damages caused by people, internal processes, systems or external events (Coleman, 2010). One of the daily jobs of risk officers is screening potential operational risks from a massive amount of online news outlets. Therefore, there is an urgent need for financial organizations to use artificial intelligence methods for OR prediction.

While this task can be easily formulated as a classic document classification problem, there are at least two challenges in designing such an intelligent OR prediction system. First, acquiring labels from risk officers is time-costly, and there is no standard labeled dataset for this task. Second, providing explanations is critical for OR prediction as risk officers cannot solely rely on prediction outcomes for subsequent decision making. Therefore, these practical issues call for an interpretable semi-supervised text classification framework for OR prediction. However, little prior literature has specifically studied these issues in one framework.

To tackle the above-mentioned practical challenges, we propose a semi-supervised text classification model based on the semi-supervised variational autoencoder (SemiVAE) (Kingma et al., 2014) and multi-head attention mechanism (Vaswani et al., 2017) for OR prediction task. SemiVAE allows effective learning of latent representation from both labeled and unlabeled data, and multi-head attention mechanism produces the direct visualization of informative words associated with multi-label predictive outcomes. Learning the model parameters is effective and scalable under the variational inference method.

This paper contributes to the burgeoning body of research on using NLP techniques in key financial applications. For example, the prior study leverages the textual features in firm annual reports to predict a firm's stock price volatility using firm annual reports (Kogan et al., 2009) and earnings announcement transcripts (Qin and Yang, 2019). Other researches make use of news articles and social media data to predict financial markets variables, such as stock return, firm performance, default prediction and market sentiment (Tetlock, 2007; Schumaker and Chen, 2009; Ding et al., 2015; Luo et al., 2018). It is worth emphasizing that the pre-requisites of using NLP in key financial applications are effective and transparent. In many cases, it requires extensive domain expertise to annotate the variable of interests. Moreover,

[1] https://www.bain.com/insights/how-banks-can-manage-operational-risk/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 846–852

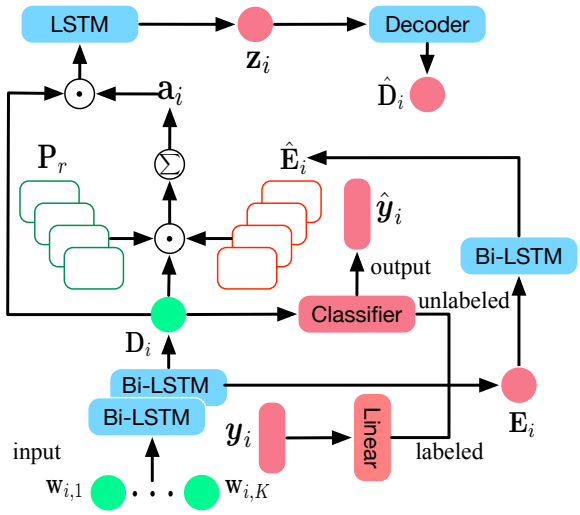

Figure 1: The framework of the proposed SemiORC. $\odot$ denotes the matrix multiplication. In the training process, the predicted labels $\hat{\boldsymbol{y}}_i$ and observed labels $\boldsymbol{y}_i$ are used to compute the classification loss for labeled data.

the black-box model does not meet the needs for actionable managerial insights. Thus, we hope that this work, which aims at addressing common issues in financial NLP system, provides valuable design guidance for financial applications with a significant societal and economic impact.

2 The Proposed Method

We now proceed with the details of our model SemiORC, and the overall architecture is shown in Figure 1. In a nutshell, SemiORC consists of an encoder, a decoder and a semi-supervised classifier. Specifically, the encoder network combines the document representation and label embedding to learn latent variables of words. The decoder is used to generate document representation based on these latent variables. We model the semi-supervised classifier by the LSTM, the fully-connected layer, and the softmax function.

Problem Definition. Let $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ be a set of finance documents with labeled $\mathcal{D}_l$ and unlabeled data $\mathcal{D}_u$. Each labeled document $D_i \in \mathcal{D}_l$ is associated with a number of operational risks $\boldsymbol{y}_i (\subseteq \boldsymbol{y})$, where $\boldsymbol{y} = \{y_1, y_2, \cdots, y_R\}$ is a set of R risk labels (e.g., Data Privacy and Bank Prosecution, etc.). We consider operational risk classification (ORC) problem that labels the unlabeled documents with possible operational risks, i.e., $D_i (\in \mathcal{D}_u) \rightarrow \hat{\boldsymbol{y}}_i (\subseteq \boldsymbol{y})$.

Document Representation. In SemiORC, we employ a Bidirectional LSTM (Bi-LSTM) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) as the basic content learning module. Let D_i be the i-th document with K words and $w_{i,k}$ denotes the one-hot representation of the k-th word. We first embed the k-th word into low-dimensional vectors using an embedding matrix $\mathbf{M}$: $\mathbf{w}_{i,k} = w_{i,k}\mathbf{M}$, where $\mathbf{w}_{i,k} \in \mathbb{R}^d$ and d is the dimension of word embedding. Then, we use the two-layer Bi-LSTMs as the document encoder to obtain the representation of k-th word by concatenating the forward and backward hidden states of the second Bi-LSTM layer:

$$\overrightarrow{\mathbf{h}}_k = \overrightarrow{\text{LSTM}}(\overrightarrow{\mathbf{h}}_{k-1}, \mathbf{w}_{i,k}), \quad (1)$$

$$\overleftarrow{\mathbf{h}}_k = \overleftarrow{\text{LSTM}}(\overleftarrow{\mathbf{h}}_{k+1}, \mathbf{w}_{i,k}), \quad (2)$$

where $\mathbf{h}_k = [\overrightarrow{\mathbf{h}}_k, \overleftarrow{\mathbf{h}}_k] \in \mathbb{R}^{2d}$. Then, we can obtain the i-th document representation $\mathbf{D}_i \in \mathbb{R}^{K \times 2d}$ by concatenating all words' representation in this document. Meanwhile, we get two final states from two directions of the second Bi-LSTM layer: hidden state $\mathbf{f}_i \in \mathbb{R}^{2 \times d}$ and cell state $\mathbf{m}_i \in \mathbb{R}^{2 \times d}$.

Label Embedding. In order to efficiently leverage risk label information, we propose a useful way to encode labels into low dimensional vectors in the training process. We first get label embedding matrix $\mathbf{E}_i$ as follows:

$$\mathbf{E}_i = \begin{cases} \text{Linear}(\boldsymbol{y}_i), & \text{if } D_i \in \mathcal{D}_l \\ \text{Classifier}(\mathbf{D}_i), & \text{if } D_i \in \mathcal{D}_u \end{cases} \quad (3)$$

where $\mathbf{E}_i \in \mathbb{R}^{d \times L_i}$ and L_i is the number of $\boldsymbol{y}_i$. $\boldsymbol{y}_i$ are the observed operational risks of i-th document, and the Linear is a fully-connected layer. The Classifier is a semi-supervised classifier, which can predict risk labels and learn the corresponding label embedding based on both labeled and unlabeled document representation. Inspired by prior work (Rai et al., 2015; Yang et al., 2018; Wang et al., 2018), we incorporate two final states $\mathbf{f}_i$ and $\mathbf{m}_i$ into label embedding $\mathbf{E}_i$ through another Bi-LSTM, which is beneficial to learn the specific label embedding of i-th document:

$$\hat{\mathbf{E}}_i = \text{Bi-LSTM}(\mathbf{E}_i, (\mathbf{f}_i, \mathbf{m}_i)), \quad (4)$$

where $\hat{\mathbf{E}}_i \in \mathbb{R}^{2d \times L_i}$.

Multi-head Attention. The document vector usually involves rich semantics in multiple semantic spaces. However, the traditional attention mechanisms only focus on a specific semantic space of document representation to learn the weights of

words, which ignores the influence of other semantic spaces. In our work, we utilize the multi-head attention mechanism (Vaswani et al., 2017; Tao et al., 2018; Huang et al., 2019) to learn the weights of all words for the corresponding labels in each document. We first project document representation $\mathbf{D}_i$ and label embedding matrix $\hat{\mathbf{E}}_i$ to h different semantic spaces through different learnable projection matrices. Then, we learn the weight matrices of words for the labels from these semantic spaces:

$$\mathbf{D}_i^{(r)} = \mathbf{D}_i \cdot \mathbf{P}_r, \quad \hat{\mathbf{E}}_i^{(r)} = \mathbf{P}_r^\top \cdot \hat{\mathbf{E}}_i, \tag{5}$$

$$\mathbf{a}_i^{(r)} = \text{softmax}(\mathbf{D}_i^{(r)} \cdot \hat{\mathbf{E}}_i^{(r)}), r = 1, \cdots, h \tag{6}$$

where $\mathbf{P}_r \in \mathbb{R}^{2d \times (2d/h)}$ is the r-th projection matrix, $\hat{\mathbf{D}}_i^{(r)} \in \mathbb{R}^{K \times (2d/h)}$, and $\hat{\mathbf{E}}_i^{(r)} \in \mathbb{R}^{(2d/h) \times L_i}$. $\mathbf{a}_i^{(r)} \in \mathbb{R}^{K \times L_i}$ denotes the weight matrix of words for the corresponding labels at the r-th semantic spaces. Besides, $\mathbf{a}_i = \frac{1}{h}\sum_{r=1}^{h} \mathbf{a}_i^{(r)}$ is the average accumulated weight matrix of words. Subsequently, we can learn latent variables of words from the document representation through the LSTM network. Inspired by prior work (Xu et al., 2017), we combine the label embedding and the latent variables to generate the document representation through the Decoder:

$$\mathbf{z}_i = \text{LSTM}[\text{sigmoid}(\mathbf{a}_i \cdot \mathbf{a}_i^\top) \cdot \mathbf{D}_i], \tag{7}$$

$$\hat{\mathbf{D}}_i = \text{Decoder}[\mathbf{z}_i + \tanh(\text{Linear}(\mathbf{a}_i \cdot \hat{\mathbf{E}}_i^\top))], \tag{8}$$

where $\mathbf{z}_i \in \mathbb{R}^{K \times (d/2)}$, $\hat{\mathbf{D}}_i \in \mathbb{R}^{K \times 2d}$, and the Linear is another fully-connected layer. The sigmoid and tanh are two activation functions. We model the Decoder by the LSTM network.

Leveraging Unlabeled Financial Documents. Various machine learning models, including SVM (Cesa-Bianchi et al., 2006), representation learning (Dai and Le, 2015), and adversarial training (Miyato et al., 2017), have been used to solve the semi-supervised text classification. Recently, VAE-based methods have been successfully used in semi-supervised learning and utilize unlabeled data to model the generating process of underlying data (Kingma and Welling, 2014; Miao et al., 2016; Xie and Ma, 2019; Gururangan et al., 2019). In addition, previous work (Xu et al., 2017) proposes to incorporate labels into the decoder RNN for better text classification performance.

In this work, we use the semi-supervised variational autoencoder (SemiVAE) (Kingma et al., 2014; Yang et al., 2019) to exploit these data, which provides an efficient way to approximate the posterior distribution of latent variables by deriving a lower bound for the marginal likelihood of the observed data (a.k.a. ELBO). More specifically, we assume a latent variable $\mathbf{z}$ for generating the representation of finance document, whose true posterior distribution $p(\mathbf{z}|\mathcal{D})$ is usually too complicated to have an analytical form. We alternatively resort to the distribution in an exponential family to approximate the true posterior: $q(\mathbf{z}|\mathcal{D}) \sim p(\mathbf{z}|\mathcal{D})$. The ELBO on the marginal likelihood of the finance documents is as follows:

$$\begin{aligned}\log & p_\theta(\mathcal{D}) \geq \log p_\theta(\mathcal{D}) - \text{KL}[q_\phi(\mathbf{z}|\mathcal{D}) \parallel p_\theta(\mathbf{z}|\mathcal{D})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{D})}[\log p_\theta(\mathcal{D}|\mathbf{z})] - \text{KL}\left[q_\phi(\mathbf{z}|\mathcal{D})||p_\theta(\mathbf{z})\right],\end{aligned} \tag{9}$$

where $q_\phi(\mathbf{z}|\mathcal{D})$ is an approximation to the true posterior $p_\theta(\mathbf{z}|\mathcal{D})$. Since the objective is to minimize the KL divergence between $q_\phi(\mathbf{z}|\mathcal{D})$ and the true distribution $p_\theta(\mathbf{z}|\mathcal{D})$ – we can alternatively maximize ELBO $\mathcal{L}(\mathcal{D})$ of $\log p(\mathcal{D})$.

Our model consists of three components: an encoder network $q_\phi(\mathbf{z}_i|\mathbf{D}_i, \boldsymbol{y}_i)$, the decoder network $p_\theta(\hat{\mathbf{D}}_i|\boldsymbol{y}_i, \mathbf{z}_i)$, and a semi-supervised classifier $q_\phi(\hat{\boldsymbol{y}}_i|\mathbf{D}_i)$. For each labeled finance data $D_i \in \mathcal{D}_l$ and its corresponding observed risk labels $\boldsymbol{y}_i \subseteq \boldsymbol{y}$, the ELBO $\mathcal{L}(\mathcal{D}_l)$ with corresponding latent variable $\mathbf{z}$ is as follows:

$$\begin{aligned}\log p_\theta(\mathbf{D}_i, \boldsymbol{y}_i) \geq & \mathbb{E}_{q_\phi(\mathbf{z}_i|\mathbf{D}_i, \boldsymbol{y}_i)}[\log p_\theta(\hat{\mathbf{D}}_i|\boldsymbol{y}_i, \mathbf{z}_i)] \\ & + \log p_\theta(\boldsymbol{y}_i) - \text{KL}\left[q_\phi(\mathbf{z}_i|\mathbf{D}_i, \boldsymbol{y}_i)||p(\mathbf{z}_i)\right] \\ & = -\mathcal{L}(\mathcal{D}_l),\end{aligned} \tag{10}$$

where $\text{KL}\left[q_\phi(\mathbf{z}_i|\mathbf{D}_i, \boldsymbol{y}_i)||p(\mathbf{z}_i)\right]$ is the KL divergence between the latent posterior $q_\phi(\mathbf{z}_i|\mathbf{D}_i, \boldsymbol{y}_i)$ and the prior distribution $p(\mathbf{z}_i)$ that should be minimized. Note that we utilize the KL cost annealing method (Bowman et al., 2016; Sønderby et al., 2016) to smooth the training process by gradually increasing the weight β of KL cost from 0 to 1.

In the case of each unlabeled document $D_i \in \mathcal{D}_u$, the corresponding risks $\hat{\boldsymbol{y}}_i$ are predicted by performing posterior inference with a probabilistic classifier $q_\phi(\hat{\boldsymbol{y}}_i|\mathbf{D}_i)$. We now have the following ELBO $\mathcal{L}(\mathcal{D}_u)$, by considering possible risks $\hat{\boldsymbol{y}}_i$ as another latent variable:

$$\begin{aligned}\log p_\theta(\mathbf{D}_i) \geq & \sum_{\boldsymbol{y}_i} q_\phi(\hat{\boldsymbol{y}}_i|\mathbf{D}_i)(-\mathcal{L}(\mathcal{D}_l)) + \\ & \mathcal{H}(q_\phi(\hat{\boldsymbol{y}}_i|\mathbf{D}_i)) = -\mathcal{L}(\mathcal{D}_u),\end{aligned} \tag{11}$$

The ELBO $\mathcal{L}(\mathcal{D})$ on the marginal likelihood for the entire dataset is as follows:

$$\mathcal{L}(\mathcal{D}) = \sum_{\mathcal{D}_l} \mathcal{L}(\mathcal{D}_l) + \sum_{\mathcal{D}_u} \mathcal{L}(\mathcal{D}_u) + \alpha \mathbb{E}_{(\mathbf{D}_i, \boldsymbol{y}_i) \in \mathcal{D}_l}[-\log q_\phi(\hat{\boldsymbol{y}}_i | \mathbf{D}_i)]. \quad (12)$$

where the last term denotes an additional classification loss of classifier $q_\phi(\hat{\boldsymbol{y}}_i|\mathbf{D}_i)$ when learning from the labeled data with a weight controlling hyper-parameter α.

3 Experiments

Label	Risk type
(0)	Non relevant
(1)	Internal Fraud
(2)	Bank Disruption & System Failure
(3)	External Fraud
(4)	Employment Practices & HR
(5)	Compliance & Regulation
(6)	Clients & Market Practices
(7)	Data Privacy

Table 1: Operational Risk Categories. Numbers in the parentheses are the category index used in experiments.

Data Description. Our proprietary dataset combines a set of 5,483 financial news articles, collected by a risk management team (with a focus on Asian-Pacific region) in an international banking organization. The financial news articles are collected from several online mainstream financial news outlets during Feb 1, 2019, to Mar 1, 2019. The news outlets include government agency such as the Association of Certified Financial Crime Specialists (ACFCS) and news agency such as The Edge Markets and Japan Times. We remove noise data (e.g., inserted advertising and specific symbol) of all finance documents. There are eight Operational Risk categories in Tabel 1, as defined in Basel Accords. The details of our dataset are as follows: 730 labeled documents; 4,753 unlabeled documents; the average number of risk labels and words for documents are 2.1 and 453, respectively.

Baselines. We consider the following baselines. **Logistic Regression** is a vanilla supervised classification baseline. It only leverages labeled documents to build a text classifier and predict risk categories. We also consider the following three semi-supervised learning baselines. **Transductive SVM (TSVM)** (Joachims, 1999) is a widely used semi-supervised method that extends SVMs with the goal that there are a few unlabeled data near the margin as possible. **Semi-supervised Variational Autoencoder (SemiVAE)** (Kingma et al., 2014) proposes to utilize a deep generative model to exploit unlabeled data. Our model **SemiORC** uses SemiVAE as one key component. **Semi-supervised Sequential Variational Autoencoder (SSVAE)** (Xu et al., 2017) proposes to use a modified version of LSTM as the decoder and is the state-of-the-art semi-supervised model for text classification. However, none of the above baselines can highlight keywords that are informative to prediction outcomes, since they are black-box semi-supervised learning models. Lastly, we consider one ablation baseline **ORC**, which is a supervised version of our SemiORC. It ignores unlabeled data for modeling document representation.

Evaluation Metrics. We follow the standard evaluation metrics of multi-label classification, including hamming loss, accuracy and micro-F1 score. **Hamming-loss** (Schapire and Singer, 1999) calculates the average Hamming distance between true labels and predicted labels. **Accuracy** computes the subset accuracy between true labels and predicted labels. **Micro-F1** (Manning et al., 2008) returns a weighted average of precision and recall, which is computed from true positives, false negatives, and false positives.

Experimental Setting. Our model SemiORC is implemented with Tensorflow on a machine with NVIDIA GeForce GTX 1080Ti. Specifically, we optimize the training process of the model using Adam optimizer (Kingma and Ba, 2015) and dropout regularization (Srivastava et al., 2014; Gal and Ghahramani, 2016). We set the number of projection matrices and the dimension of word embedding as $h = 4$ and $d = 64$. The learning rate and weight parameter α are empirically tuned to 0.001 and 2, respectively. The dropout rate is scaled from 0.3 to 0.7. For Logistic Regression and TSVM, we both use doc2vec (Le and Mikolov, 2014) to learn the finance document representation. Additionally, we leverage the scikit-learn (Pedregosa et al., 2011) to build two text classifiers to predict the corresponding risk labels. For SemiVAE and SSVAE, we model the encoders, the decoders, and the classifiers by the LSTM networks.

Experiment Results. We perform 10 runs of 10-fold cross-validation on the dataset for each method. Table 2 reports the overall classification performance on three metrics. We can see that SemiORC achieves the best classification per-

Method	Hamming-loss	Accuracy	Micro-F1 score
Logistic Regression	0.156(±0.030)	0.406(±0.026)	0.510(±0.031)
TSVM	0.135(±0.027)	0.392(±0.037)	0.493(±0.036)
SemiVAE	0.106(±0.024)	0.417(±0.031)	0.595(±0.020)
SSVAE	0.097(±0.019)	0.457(±0.026)	0.621(±0.022)
ORC	0.105(±0.013)	0.443(±0.022)	0.601(±0.028)
SemiORC	**0.084**(±0.018)	**0.529**(±0.020)	**0.651**(±0.023)

Table 2: Overall operational risk classification results. The bold number indicates statistically significant over the second-best results at 5% level under a one-tailed t-test ($p < 0.05$). Standard deviation is reported in the parentheses. For Hamming-loss, the smaller number indicates better performance. For Accuracy and Micro-F1 score, the larger number is better.

Example documents	1	2	3	4	5	6	7
A group of companies run by them were involved in a conflict of interest							
Company B failed to update expired software certificates...							
Cyber criminals typically look for loopholes of vulnerable systems in...							
Retailers acknowledge vulnerabilities of ... and recognize the need for encryption							
Some personal data had been compromised because of the cyber intrusion							
Banks had experienced data breaches because their systems were under attacks							
Bank employees helped taxpayers open bank accounts to assist them in tax evasion							
The Anti-Money Laundering Act mandates that each citizen links identity number							

Table 3: Examples of financial documents where the keywords are highlighted. Each row is an example document, and darker color indicates higher attention weight. Note that only words with the largest attention weights in the sentence are colored for better illustration. Right columns are indexes of each operational risk category, where the color density indicates the predicted probability that the left document is belonging to the category.

formance in all three metrics. Compared with SSVAE, SemiORC improves the Hamming-loss by 13.4% (0.097 vs. 0.084), Accuracy by 15.7% (0.457 vs. 0.529), Micro-F1 score by 4.8% (0.621 vs. 0.651). Compared with the pioneer semi-supervised learning model SemiVAE, SemiORC improves the Hamming-loss by 20.7% (0.106 vs. 0.094), Accuracy by 21.1% (0.417 vs. 0.529), and Micro-F1 score by 24.3% (0.493 vs. 0.651). The key difference between SemiORC and SSVAE or SemiVAE is that we leverage the multi-head attention mechanism to learn the weights of informative words which better encodes labeled and unlabeled documents. Moreover, we can conclude that utilizing unlabeled data can significantly improve model performance (ORC vs. SemiORC). Considering that the current risk management team in the bank only utilizes labeled data, this improvement is quite significant and should be emphasized.

Transparent Operational Risk Prediction. In financial institutions, risk officers are strictly required to comply with regulations and be responsible for any decisions that they make. Therefore, in order for the operation risk prediction system to be useful, it calls for transparency in the text classification system. SemiORC highlights keywords that are informative to each predicted risk type, as shown in Table 3. Take the last document "the anti-money laundering act mandates that each citizen links identify number" for example. It is predicted to be multiple labels (category 5, 6 and 7). By examining the highlighted keywords, we can see word "anti-money" has the highest attention weight under category 6 while "identity" has the highest attention weight under category 7. In other words, each predicted label is associated with a set of label-related keywords, which provides a visual explanation of why a financial news article is assigned to a specific risk category. The label-dependent attention words allow risk officers to screen out the news articles efficiently and to assess the operational risk categories accurately.

4 Conclusion

To conclude, in this paper, we work on a significant practical problem in the financial industry: operational risk prediction. We design a text classification framework with the multi-head attention mechanism and SemiVAE. In sum, our framework aims to address two common issues in the financial industry: lacking labeled data and the need for transparency in prediction outcomes.

Acknowledgement

This work was supported by the National Natural Science of China under Grant No.61602097 and No.61472064.

References

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. 2006. Hierarchical classification: combining bayes with svm. In *Proceedings of the 23rd international conference on Machine learning*, pages 177–184. ACM.

Rodney Coleman. 2010. Operational risk. *Wiley Encyclopedia of Operations Research and Management Science*.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*, pages 2327–2333.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems*, pages 1019–1027.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. 2019. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 5880–5894.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Douglas G Hoffman. 2002. *Managing operational risk: 20 firmwide best practice strategies*, volume 109. John Wiley & Sons.

Po-Yao Huang, Xiaojun Chang, and Alexander G. Hauptmann. 2019. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1461–1467.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, 2015, Conference Track Proceedings*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, 2014, Conference Track Proceedings*.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

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 Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1188–1196.

Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. 2018. Beyond polarity: interpretable financial sentiment analysis with hierarchical query-driven attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4244–4250.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of the 33nd International Conference on Machine Learning*, pages 1727–1736.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, Conference Track Proceedings*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *ACL*, pages 390–401.

Piyush Rai, Changwei Hu, Ricardo Henao, and Lawrence Carin. 2015. Large-scale bayesian multi-label learning via topic-based label embeddings. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, pages 3222–3230.

Robert E. Schapire and Yoram Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Mach. Learn.*, 37(3):297–336.

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

Mike Schuster and K K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. 2016. How to train deep variational autoencoders and probabilistic ladder networks. *CoRR*, abs/1602.02282.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Chongyang Tao, Shen Gao, Mingyue Shang, Wei Wu, Dongyan Zhao, and Rui Yan. 2018. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4418–4424.

Paul C. Tetlock. 2007. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 2321–2331.

Zhongbin Xie and Shuai Ma. 2019. Dual-view variational autoencoders for semi-supervised text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 5306–5312.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3358–3364.

Kaijia Yang, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Exploiting noisy data in distant supervision relation classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,Volume 1 (Long and Short Papers)*, pages 3216–3225.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.