Financial News Annotation by Weakly-Supervised Hierarchical Multi-label Learning

Hang Jiang[1] , **Zhongchen Miao**[1] , **Yuefeng Lin**[1] , **Chenyu Wang**[1] , **Mengjun Ni**[1] , **Jian Gao**[1] , **Jidong Lu**[1] , **Guangwei Shi**[1]

[1]Innovation Lab, Shanghai Financial Futures Information Technology Co., Ltd, Shanghai, China

{jianghang, miaozc, linyf, wangcy1, nimj, gaojian, lujd, shigw}@cffex.com.cn

Abstract

Financial news is an indispensable source for both investors and regulators to conduct research and investment decisions. To focus on specific areas of interest among the massive financial news, there is an urgent necessity of automatic financial news annotation, which faces two challenges: (1) supervised data scarcity for sub-divided financial fields; (2) the multifaceted nature of financial news. To address these challenges, we target the automatic financial news annotation problem as a weakly-supervised hierarchical multi-label classification. We propose a method that needs no manual labeled data, but a label hierarchy with one keyword for each leaf label as supervision. Our method consists of three components: word embedding with heterogeneous information, multi-label pseudo documents generation, and hierarchical multi-label classifier training. Experimental results on data from a well-known Chinese financial news website demonstrate the superiority of our proposed method over existing methods.

1 Introduction

To target information of concern among massive financial news quickly, there is a natural demand to search and analyze financial news based on topics. To cater for this, most financial news media adopt a manual annotation solution, which is too tedious to cope with rapidly growing financial news. Besides, manual annotation is not intelligent enough to meet the personalized needs of everyone. Therefore, to improve the searching efficiency and analysis accuracy of financial news, a critical step is automatic financial news annotation.

Indeed, the automatic financial news annotation is a classic problem of natural language processing (NLP), that is, text classification. Although related research keeps emerging, however, compared to those common scenarios of fully-supervised flat single-label text classification, our task faces two major challenges. First, supervised model training heavily relies on labeled data, while annotated corpus for each sub-divided financial field is cost expensive, considering the significant professional knowledge requirements for manual annotation. Second, a piece of financial news usually talks about multiple financial products and concepts from multiple levels and perspectives, but it is difficult to apply existing mature neural networks to multi-label and hierarchical text classification simultaneously.

In recognition of the challenges above, we propose a weakly-supervised hierarchical multi-label classification method for financial news. Our method is built upon deep neural networks, while it only requires a label hierarchy and one keyword for each leaf label as supervision, without any labeled data requirements. To leverage user-provided supervised keywords and semantic information in financial news, even though they are unlabeled, our method employs a two-step process of pre-training and self-training. During the pre-training process, we train a classifier with pseudo documents driven by user-provided keywords. Specifically, we model topic distribution for each category with user-provided keywords and generate multi-label pseudo documents from a bag-of-word model guided by the topic distribution. Self-training is a process of bootstrapping, using the predictions of unlabeled financial news as supervision to guide pre-training classifier fine-tuning iteratively. To ensure the effectiveness of self-training, a novel confidence enhancement mechanism is adopted. Besides, we include multi-modal signals of financial news into the word embedding process by heterogeneous information networks (HIN) [Sun and Han, 2012] encoding algorithm.

To summarize, we have the following contributions:

1. We propose a method of weakly-supervised hierarchical multi-label classification for financial news driven by user-provided keywords. With our proposed method, users do need to provide a label hierarchy with one keyword for each leaf label as the supervised source but not any manual labeled data.
2. To bridge the gap between low-cost weak supervision and expensive labeled data, we propose a multi-label pseudo documents generation module that almost reduces the annotation cost to zero.
3. In the hierarchical multi-label classification model training process, we transform the classification problem into a regression problem and introduce a novel confidence enhancement mechanism in the self-training process.
4. We demonstrate the superiority of our method over var-

ious baselines on a dataset from Cailianshe[1] (a well-known Chinese financial news website), conduct a thorough analysis of each component, and confirm the practical significance of hierarchical multi-label classification by an application.

2 Related Work

Financial text mining

As an important branch of fintech, financial text mining refers to obtaining valuable information from massive unstructured text data, which has attracted the attention of many researchers. The research object of text mining can be a company's financial report [Bai *et al.*, 2019], as well as self-media content such as Weibo (Chineses twitter) [Wang *et al.*, 2019]. The purpose of the research is also different, for example, studies [Sun *et al.*, 2016; Seong and Nam, 2019] analyze market prediction using financial news, and study [Kogan *et al.*, 2009] is dedicated to risk discovery. In our work, we take the financial news as the research object, and annotate each piece of news with multiple labels from a label hierarchy automatically.

Weakly-supervised text classification

Despite the maturity of adopting neural networks in supervised learning, the requirements for labeled data are extremely expensive and full of obstacles, so weakly-supervised learning emerges as the times require. Above all classic works, it can be roughly divided into two directions: extending the topic model in the semantic space by user-provided seed information [Chen *et al.*, 2015; Li *et al.*, 2016], and transforming weakly-supervised learning to full-supervised learning by generating pseudo documents [Zhang and He, 2013; Meng *et al.*, 2018].

Hierarchical text classification

Hierarchical classification is more complicated than flat one, considering the hierarchy of labels. A lot of research on applying SVM in hierarchical classification [Cai and Hofmann, 2004; Liu *et al.*, 2005] has been started from the first application of [Dumais and Chen, 2000]. Hierarchical dataless classification [Song and Roth, 2014] projects classes and documents into the same semantic space by retrieving Wikipedia concepts. [Meng *et al.*, 2019; Zhang *et al.*, 2019] is a continuation of the work in [Meng *et al.*, 2018], which solves the problem of hierarchical classification through a top-down integrated classification model. To our best knowledge, there is no hierarchical multi-label classification method based on weak supervision so far.

3 Problem Statement

We take the financial news annotation as a task of weakly-supervised hierarchical multi-label classification. Specifically, each piece of news can be assigned multiple labels, and each category can have more than one children categories but can only belong to at most one parent category.

To solve our task, we ask users to provide a tree-structured label hierarchy $\mathcal{T}$ and one keyword for each leaf label in $\mathcal{T}$. Then we propagate the user-provided keywords upwards from leaves to root in $\mathcal{T}$, that is, for each internal category, we aggregate keywords of its all descendant leaf classes as supervision.

Now we are ready to formulate the problem. Given a class hierarchy tree $\mathcal{T}$ with one keyword for each leaf class in $\mathcal{T}$, and news corpora $\mathcal{D} = \{D_1, D_2, ..., D_N\}$ as well. The weakly-supervised hierarchical multi-label classification task aims to assign the most likely labels set $C = \{C_{j_1}, C_{j_2}, ..., C_{j_n} | C_{j_i} \in \mathcal{T}\}$ to each $D_j \in \mathcal{D}$, where the number of assigned labels is arbitrary and C_{j_i} stays for classes at any level.

4 Methodology

The framework of our method is illustrated in Figure 1, which can be divided into three phases. Because the corpus we use is in Chinese, word segmentation is an essential step before classification. Considering the specificity of the financial corpus, we construct a financial segmentation vocabulary including financial entities, terminologies and English abbreviations by neologism discovery algorithm [Yao *et al.*, 2016].

4.1 Word embedding with heterogeneous information

Compared to plain textual data, financial news is a complex object composed of multi-modal signals, including news content, headline, medium, editor, and column. These signals are beneficial to topic classification, for example, editors are indicative because two pieces of news are more likely to share similar topic if they are supplied by the same editor as editors usually have stable specialty and viewpoints.

To learn d-dimensional vector representations for each word using such significant multi-modal signals in the corpus, we construct a HIN centered upon words [Zhang *et al.*, 2019]. Specifically, corresponds to heterogeneous information in financial news, we include seven types of nodes: news (N), columns (C), headlines (H), media (M), editors (E), words (W) and labels (L). In which, headlines (H) and words (W) are tokens segmented from title and content respectively. As a word-centric star schema is adopted, we add an edge between a word node and other nodes if they appear together, thus the weights of edges reflect their co-occurrence frequency.

Given a HIN following the above definition of nodes and edges, we can obtain word representations by learning nodes embeddings in this HIN. We use ESIM [Shang *et al.*, 2016], a typical HIN embedding algorithm, to learn nodes representations by restricting the random walk under the guidance of user-specified meta-paths. To guide the random walk, we need to specify meta-paths centered upon words and assign the weights by the importance of meta-path. In our method, we specify meta-paths as W–N–W, W–H–W, W–M–W, W–E–W, $W-C-W$ and W–L–W with empirical weights, modeling the multi-types of second-order proximity [Tang *et al.*, 2015] between words. Furthermore, we perform normalization $\boldsymbol{v}_w \leftarrow \boldsymbol{v}_w / ||\boldsymbol{v}_w||$ on embedding vector $\boldsymbol{v}_w$ for each word w.

[1]The website of Cailianshe: https://cls.cn

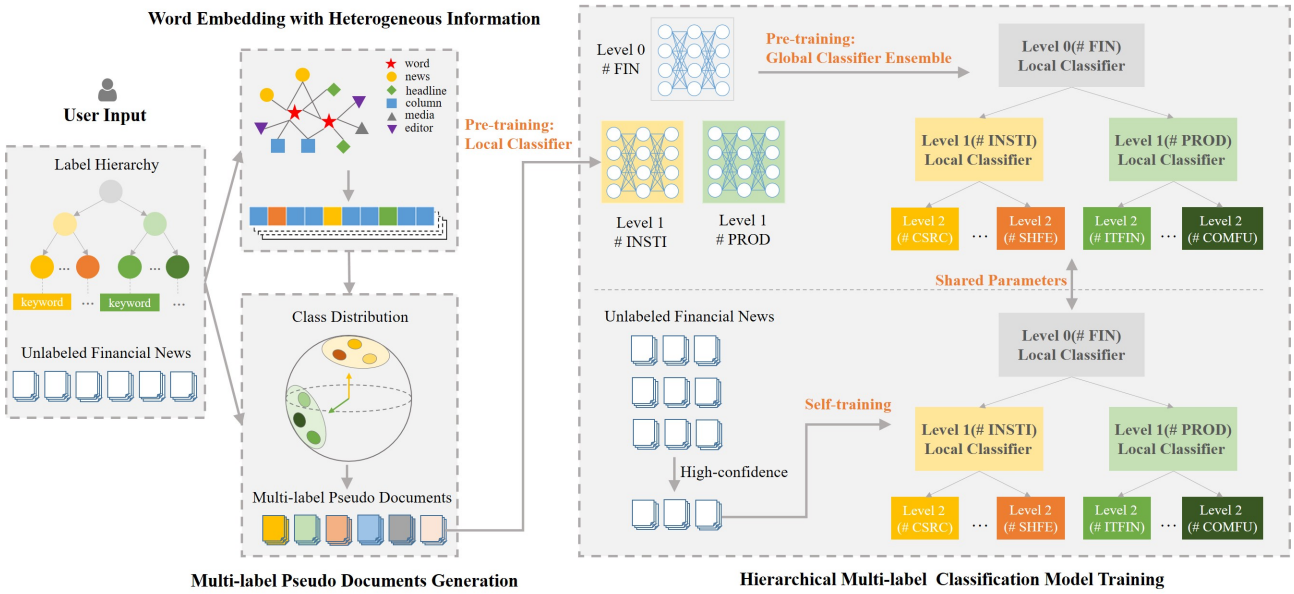

Figure 1: The framework of proposed method.

4.2 Multi-label pseudo documents generation

In this section, we first model class distribution in a semantic space with user-provided keywords, and then generate multi-label pseudo documents as supervised training data based on them.

Modeling class distribution

Assume that words and documents shared a uniform semantic space, so that we can leverage user-provided keywords to learn a class distribution [Meng *et al.*, 2018].

Specifically, we first take the inner product of two embedding vectors $\boldsymbol{v}_{w_1}^T \boldsymbol{v}_{w_2}$ as similarity measurement between two words w_1 and w_2 to retrieve top n nearest keywords set $K_j = \{w_{j0}, w_{j1}, ..., w_{jn}\}$ in semantic space for each class j based on user-provided keyword w_{j0}. Remind that we do not specify the parameter n above but terminate the keywords retrieving process when keyword sets of any two classes tend to intersect to ensure the absolute boundary between different classes. Then we fit the expanded keywords distribution $f(\boldsymbol{x}|C_j)$ to a mixture von Mises-Fisher (vMF) distributions [Banerjee *et al.*, 2005] to approximate class distribution for each class:

$$f(\boldsymbol{x}|C_j) = \sum_{h=1}^{m} \alpha_h f_h\left(\boldsymbol{x}|\boldsymbol{\mu}_h, \kappa_h\right) \tag{1}$$

where $f_h\left(\boldsymbol{x}|\boldsymbol{\mu}_h, \kappa_h\right)$, as a component in the mixture with a weight α_h, is the distribution of the h-th child of category C_j, m is equal to the number of C_j's children in the label hierarchy. In $f_h\left(\boldsymbol{x}|\boldsymbol{\mu}_h, \kappa_h\right)$, μ_h is the mean direction vectors and κ_h is the concentration parameter of the vMF distribution, which can be derived by Expectation Maximization (EM) [Banerjee *et al.*, 2005].

Pseudo documents generation

Given distribution for each class, we use a bag-of-words based language model to generate multi-label pseudo documents. We first sample l document vectors d_i from various class distribution $f(x|C)$ (l is not specific), and then build a vocabulary V_{d_i} that contains the top γ words closest to d_i in semantic space for each d_i. Given a vocabulary set $\mathcal{V}_d = \{V_{d_1}, V_{d_2}, ..., V_{d_l}\}$, we choose a number of words to generate pseudo document with probaliblity $p(w|D)$. Formally,

$$p\left(w|D\right) = \begin{cases} \beta p_B(w) & w \notin \mathcal{V}_d \\ \beta p_B(w) + (1-\beta) p_D(w) & w \in \mathcal{V}_d \end{cases} \tag{2}$$

where β is a "noisy" parameter to prevent overfitting, $p_B(w)$ is the background words distribution (i.e., word distribution in the entire corpus), $p_D(w)$ is the document-specific distribution, that is,

$$p_D(w) = \frac{1}{l} \sum_{i=1}^{l} \frac{\exp\left(d_i^T \boldsymbol{v}_w\right)}{\sum_{w' \in V_{d_i}} \exp\left(d_i^T \boldsymbol{v}_{w'}\right)} \tag{3}$$

where $\boldsymbol{v}_w$ is the embedding of word w. Meanwhile, pseudo labels need to be expressed. Suppose existing k document vectors d_i are generated from class j, then the label of class j of document D can be represented by,

$$label^*(D)_j = tanh(\sigma(\frac{k(1-\beta)}{l} + \frac{\beta}{m})) \tag{4}$$

where σ is a scale parameter to control the range of $label^*(D)_j$, and generally takes an empirical value.

Otherwise, if $\forall d_i$ is not generated from class j,

$$label^*(D)_j = \beta / m \tag{5}$$

Algorithm 1 Multi-label Pseudo Documents Generation

Input: Class distribution set $\{f(x|C_j)|_{j=1}^m\}$.
Parameter: number of probability distribution β to generate multi-label pseudo documents for each class; number of pseudo documents γ.
Output: A set of γ multi-label pseudo documents D^* and corresponding labels set $\mathcal{L}^*$

Initialize $D^* \leftarrow \emptyset$, $\mathcal{L}^* \leftarrow \emptyset$, $p \leftarrow \emptyset$;.
for $class\ index\ j\ from\ 1\ to\ m$ **do**
  for $probability\ distribution\ index\ i\ from\ 1\ to\ \beta$ **do**
    Sample document vector d_i from $f(x; C_j)$;
    Calculate probability distribution $p(w|d_i)$ based on Eq 2 // parameter $l = 1$ in Eq 2;
    $p \leftarrow p \cup p(w|d_i)$
  end for
end for
Sample γ probability distribution combinations from p
for $combination\ index\ i\ from\ 1\ to\ \gamma$ **do**
  $D_i^* \leftarrow\ empty\ string$
  Calculate probability distribution $p(w|D_i)$ based on Eq 2
  Sample $w_{ik} \sim p(w|D_i)$
  $D_i^* = D_i^* + w_{ik}$ //concatenate w_{ik} after D_i^*
  Calculate label $\mathcal{L}_i^*$ based on Eq 4 and Eq 5
  $D^* \leftarrow D^* \cup D_i^*$
  $\mathcal{L}^* \leftarrow \mathcal{L}^* \cup \mathcal{L}_i^*$
end for
return D^*, $\mathcal{L}^*$

where m is the number of children classes related to the local classifier.

Algorithm 1 shows the entire process for generating multi-label pseudo-documents.

4.3 Hierarchical multi-label classifier training

In this section, we pre-train CNN-based classifiers with pseudo documents and refine it with real unlabeled documents.

Pre-training with pseudo documents

Hierarchical classification model pre-training can be split into two parts: local classifier training for nodes and global classifier ensembling. We trained a neural classifier $M_L(\cdot)$ for each class with two or more children classes. $M_L(\cdot)$ has multi-scale convolutional kernels in the convolutional layer, ReLU activation in the hidden layer, and Sigmoid activation in the output layer. As the pseudo label is a new distribution instead of binarization vectors, we transform task from multi-label classification to regression and minimizing the mean squared error (MSE) loss from the network outputs to the pseudo labels.

After training a series of local classifiers, we need to build a global classifier G_k by integrating all local classifiers from the root node to level k from top to bottom. The multiplication between the output of the parent classifier and child classifier can be explained by conditional probability formula:

$$\begin{aligned} p\left(D_i \in C_c\right) &= p\left(D_i \in C_c \cap D_i \in C_p\right) \\ &= p\left(D_i \in C_c | D_i \in C_p\right) p\left(D_i \in C_p\right) \end{aligned} \quad (6)$$

where, class C_c is the child of class C_p. When the formula is called recursively, the final prediction can be obtained by the product of all local classifier outputs on the path from the root node to the target node.

Self-training with unlabeled real documents

To take advantage of semantic information in the real documents, we utilize the prediction of real documents as supervision in the self-training procedure iteratively. However, if the predictions are used as the supervision for the next iter self-training directly, the self-training can hardly go on because the model has been convergent in pre-training. To obtain more high-confidence training data, we adopt a confidence enhancement mechanism. Specifically, we calculate the confidence of predictions by Eq 7 and only reserve data with high-confidence as training data.

$$conf(q) = -\frac{\log(\sum_{i=1}^m q_i + 1)}{\sum_{i=1}^m q_i \log q_i}. \quad (7)$$

where $m \geq 2$ is the number of children of C_j.

In addition, we notice the true label of a real document is either zero or one, thus, we conduct a normalization on G_k's predictions by the following formula:

$$label^{**}(D_i)_j = \frac{label^*(D_i)_j}{\max_{j'}(label^*(D_i)_{j'})} \quad (8)$$

When the change rate of G_k's outputs of real documents is lower than δ, the self-training will stop earlier.

5 Experiments

Three things will be demonstrated in this section. First, the performance of our method is superior to various baselines for the weakly-supervised hierarchical multi-label financial news classification task (Section 5.2). Second, we carefully evaluate and analyze the components in our method proposed in Section 4(Section 5.3). Third, we reveal the business significance in the task of hierarchical multi-label classification for financial news by an application(Section 5.4).

5.1 Experiments setup

Dataset

We collect a dataset from a well-known Chinese financial news website, Cailianshe, to evaluate the performance of our method.

The dataset statistics are provided in Table 1: the news corpus consists of 7510 pieces of financial news with 2 super-categories and 11 sub-categories, covering the major institutions and product categories in China mainland financial markets. The label hierarchy refers to Figure 2 for details, in which the colored italics are user-provided keywords for leaf labels.

Figure 2: The label hierarchy for Chinese financial market.

dataset	classes (level1+level2)	docs
FIN-NEWS-CN	2+11	7510

Table 1: Dataset Statistic

It should be noted that we maintained an unbalanced dataset to truly reflect the market size and shares of the Chinese financial market. For example, financial futures account for only 10% but stocks account for 53% in the dataset. This is because there is a mature stock market in China, while the beginning of financial futures in China is late and the initial stage comes into being until China Financial Futures Exchange (CFFEX) launches CSI 300 futures in 2010 to some extent.

Baselines

- **WeSHClass** [Meng *et al.*, 2019] provides a top-down global classifier for the hierarchical classification, which supports multiple weakly supervised sources.
- **HiGitClass** [Zhang *et al.*, 2019] utilizes HIN encoding to solve a hierarchical classification task of GitHub repositories, with user-provided keywords as weak seed information.

Note that WeSHClass and HiGitClass can only output at most a single-label at each level. To compare with our method, we adjust the activation and loss function of baselines to fit a multi-label classification task, but they are still unable to generate multi-label pseudo documents.

Evaluation Metrics

According to the common standards for evaluating classification, we use Micro-F1 and Macro-F1 scores as metrics for classification performances at level 1, level 2, and overall classes respectively.

5.2 Performance comparison with Baselines

Table 2 demonstrate the superiority of our proposed method over baselines on the financial news dataset. It can be observed from Table 2 that our method has a significant improvement over baselines, whether at level 1, level 2, or overall classes. This is because we borrow the self-training mechanism of WeSHClass and HIN encoding of HIGitClass at the same time, and propose a suitable multi-label pseudo documents generation module in addition. However, for fine-grained labels, our method is still far from excellent although the average F1 scores improvement approaches 20% at level 2 comparing to baselines, which reflects the difficulty of this task.

5.3 Components Performance Evaluation

To evaluate each components, we carefully analyze performance of models with or without different components in Figure 3 and Figure 4.

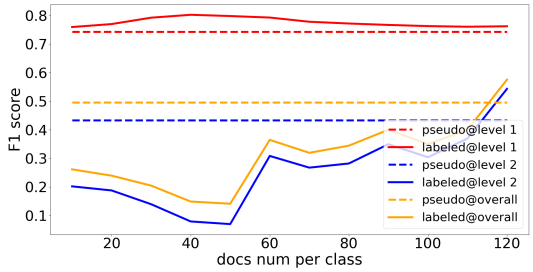

Figure 3: Performances Comparision of classificaton with pseudo documents and manual annotation documents

Qualitatively, the effectiveness of the multi-label pseudo document generation module has been demonstrated in previous training, and its quantitative value will be carefully analyzed by replacing the pseudo documents with manually labeled data. As we can observe in Figure 3, F1 score of

Method	Macro (level1)	Micro (level1)	Macro (level2)	Micro (level2)	Macro (overall)	Micro (overall)
WeSHClass	0.71373	0.80225	0.34627	0.48468	0.4085	0.60728
HiGitClass	0.68329	0.86769	0.24623	0.40716	0.31338	0.47073
Our method	**0.743**	**0.89723**	**0.44765**	**0.60173**	**0.50185**	**0.73977**

Table 2: Performance comparison for all method, using Micro-F1 and Macro-F1 scores as metrics at all levels.

pseudo documents training model is slightly lower than **labeled documents training model** at level 1, but for level2 and overall classes, the former stays lower than the latter until the number of labeled documents reaches 120 per class. To some extent, this component can save 1560 ($120\ per\ class \times 13\ classes$) pieces of documents labeling cost.

To analyze the effect of heterogeneous information and self-training, we conduct model ablation experiments to compare performances of two variants (**No heterogeneous information** and **No self-training**) and our **Full method**. Here, the method of No heterogeneous information means heterogeneous information is not included in the word embedding process, and the method of No self-training means the self-training process is removed from the complete model. Overall F1 score in Figure 4 illustrates that both No heterogeneous information and No self-training perform are worse than the Full method. Therefore, embedding words with heterogeneous information and self-training with unlabeled real data play essential roles in financial news classification.

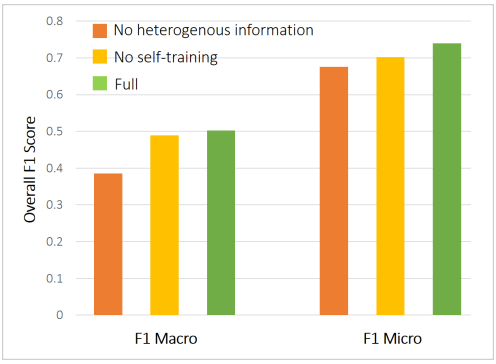

Figure 4: Comparison among No heterogeneous information, No self-training and Full method.

5.4 Application

A good classification can not only label each document appropriately but also can mine the hidden information behind the corpus. This section gives an example of a practical application, that is, discovering a correlation of business significance behind labels. In brief, we calculate the Pearson coefficients across all labels to draw a label correlation matrix in Figure 5, whose colors from shallow to deep represent the labels correlation is from weak to strong.

We only analyze the lower triangular matrix due to its symmetry, observing following two phenomena: (1) Correlations between different exchanges and products are different (e.g., CFFEX has a strong correlation with financial futures) and correlations between different exchanges are different as well (e.g., there is a strong correlation between Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE)). This phenomenon implies the main products of exchanges and their relationships. (2) Commodity futures are highly uncorrelated with stock exchanges or securities products such as stocks, while financial futures are not. This is because commodity futures (e.g., petroleum futures) take spot commodities as subject matter but financial futures (e.g., stock indexes futures) take securities products as subject matter. These phenomena are aligned with the reality of China's financial market, which demonstrates that targeting the financial news annotation task as hierarchical multi-label classification does have its practical application value, such as quickly understanding the relationship between different institutions, products, and concepts in complex financial markets.

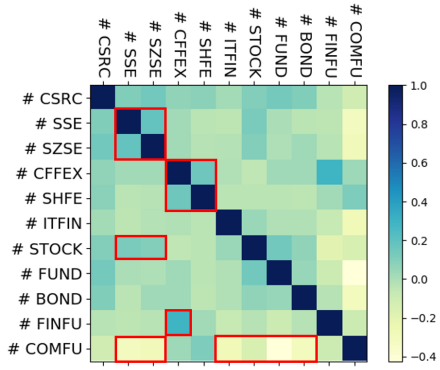

Figure 5: The labels correlation matrix, reflecting information about relationship between different financial concepts.

6 Conclusion

In this paper, we proposed a weakly-supervised hierarchical multi-label classification method with three modules for financial news, which enables us to effectively overcome challenges of supervision scarcity and the multifaceted nature of financial news. Experiments on a Chinese financial news dataset demonstrate the performance of our near-zero cost solution for hierarchical multi-label classification. Besides, we reveal the practical value and business significance of hierarchical multi-label classification in a real-world application. In the future, we would like to improve the quality of pseudo documents by label promotion methods such as the label propagation mechanism. With more accurate labels for pseudo documents, the performance of the model trained with pseudo documents will be further improved.

References

[Bai *et al.*, 2019] Haodong Bai, Frank Z Xing, Erik Cambria, and Win-Bin Huang. Business taxonomy construction using concept-level hierarchical clustering. *arXiv preprint arXiv:1906.09694*, 2019.

[Banerjee *et al.*, 2005] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005.

[Cai and Hofmann, 2004] Lijuan Cai and Thomas Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 78–87, 2004.

[Chen *et al.*, 2015] Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. Dataless text classification with descriptive lda. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Dumais and Chen, 2000] Susan Dumais and Hao Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263, 2000.

[Kogan *et al.*, 2009] Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 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, 2009.

[Li *et al.*, 2016] Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. Effective document labeling with very few seed words: A topic model approach. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 85–94, 2016.

[Liu *et al.*, 2005] Tie-Yan Liu, Yiming Yang, Hao Wan, Hua-Jun Zeng, Zheng Chen, and Wei-Ying Ma. Support vector machines classification with a very large-scale taxonomy. *Acm Sigkdd Explorations Newsletter*, 7(1):36–43, 2005.

[Meng *et al.*, 2018] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992, 2018.

[Meng *et al.*, 2019] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6826–6833, 2019.

[Seong and Nam, 2019] NohYoon Seong and Kihwan Nam. Predicting stock movements based on financial news with systematic group identification. *Journal of Intelligence and Information Systems*, 25(3):1–17, 2019.

[Shang *et al.*, 2016] Jingbo Shang, Meng Qu, Jialu Liu, Lance M Kaplan, Jiawei Han, and Jian Peng. Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. *arXiv preprint arXiv:1610.09769*, 2016.

[Song and Roth, 2014] Yangqiu Song and Dan Roth. On dataless hierarchical text classification. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[Sun and Han, 2012] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.

[Sun *et al.*, 2016] Andrew Sun, Michael Lachanski, and Frank J Fabozzi. Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, 48:272–281, 2016.

[Tang *et al.*, 2015] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015.

[Wang *et al.*, 2019] Chenyu Wang, Zhongchen Miao, Yuefeng Lin, and Jian Gao. User and topic hybrid context embedding for finance-related text data mining. *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 751–760, 2019.

[Wei and Zou, 2019] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Yao *et al.*, 2016] Rongpeng Yao, Guoyan Xu, and Jian Song. Micro-blog new word discovery method based on improved mutual information and branch entropy. *Journal of Computer Applications*, pages 2772–2776, 2016.

[Zhang and He, 2013] Pu Zhang and Zhongshi He. A weakly supervised approach to chinese sentiment classification using partitioned self-training. *Journal of Information Science*, 39(6):815–831, 2013.

[Zhang *et al.*, 2019] Yanyong Zhang, Frank F. Xu, Sha Li, Yu Meng, Xuan Wang, Qi Li, and Jiawei Han. Higitclass: Keyword-driven hierarchical classification of github repositories. *2019 IEEE International Conference on Data Mining (ICDM)*, pages 876–885, 2019.