Automatic Term and Sentence Classification Via Augmented Term and Pre-trained Language Model in ESG Taxonomy texts

Ke, Tian[1] , **ZePeng, Zhang**[2] , **Hua, Chen**[2]
[1]Rakuten Group, Inc, Japan
[2]School of Computer and Information Engineering, Jiangxi Normal University, China
tianke0711@gmail.com, gottenzzp@jxnu.edu.cn, hua.chen@jxnu.edu.cn

Abstract

In this paper, we present our solutions to the FinSim4 shared task which is co-located with the FinNLP workshop at IJCAI-2022. This new edition of FinSim4-ESG is extended to the "Environment, Social and Governance (ESG)" related issues in the financial domain. There are two sub-tasks in the FinSim4 shared task. The goal of sub-task1 is to develop a model to classify correctly a list of given terms from ESG taxonomy domain into the most relevant concepts. The aim of sub-task2 is to design a system that can automatically classify the ESG Taxonomy text sentences into sustainable or unsustainable class. We have developed several classifiers to automatically predict the concepts of terms with augmented terms and word vectors and classify sentences into sustainable or unsustainable label with pre-trained language models. The final result leaderboard shows that our proposed methods yield a significant performance improvement compared to the baseline which ranked 1st in the sub-task2 and 2rd (Mean Rank) in the sub-task1.

1 Introduction

Natural Language Processing (NLP) is a kind of computational techniques which processes and analyzes large volume of natural language data, such as document text. In the last decade, term frequency-inverse document frequency (tf-idf) [wikipedia,] word vector and word embedding such as word2vec [Mikolov *et al.*, 2013] and Glove [Jeffrey Pennington and Manning, 2014] are widely used in the NLP tasks which became the default standard features in many NLP tasks, such as text classification. Recently, transformer model which utilizes the mechanism of self-attention is considered as a breakthrough for NLP [Vaswani *et al.*, 2017] and computer vision field [Dosovitskiy *et al.*, 2020]. The transformers model has caused the paradigm shift in NLP domain such that pre-trained language models which are applied widely in NLP tasks. Pre-trained language model has been gained wide attention after BERT achieved state-of-the-art results on a variety of NLP tasks [Devlin *et al.*, 2018]. OpenAI GPT [Radford and Narasimhan, 2018], BERT [Devlin *et al.*, 2018], DistilBERT [Sanh *et al.*, 2019], RoBERTa [Liu *et al.*, 2019], XLNet [Yang *et al.*, 2019] and XLM [Lample and Conneau, 2019] are examples of pre-trained language model (PLM) that could be applied to a wide range of NLP tasks. In the finance domain, there is a large volume of document texts to be processed for analysing financial markets, investment support, trading and so on. One of tasks is to classify these document text sentences into proper classification. The word vectors and PLMs are implemented widely for text classification in the finance text field [Araci, 2019] [Tian and Peng, 2019b] [Tian and Peng, 2019a] [Tian and Chen, 2021]. Recently "Environment, Social and Governance (ESG)" related issues in the financial domain are gained more and more attention with the goal of building sustainable environment. The aim of the FinSim4 shared task [Organizer,]is to elaborate an ESG taxonomy (ESG related concepts representations) based on the document data like companies' annual reports, sustainability reports, environment reports, etc. and utilizes them to analyse how the economic activity complies with the taxonomy. There are two sub-tasks in the FinSim4 task. Regarding the sub-task1, there is a number of terms which are selected from ESG taxonomy texts. For example, the given terms: "low-carbon", "carbon footprint" et al., These terms are related to the "Carbon factor" concept. The goal of sub-task1 is to develop a model to classify correctly a list of given terms from ESG taxonomy domain into the most relevant concepts. Regarding the sub-task2, there are selected sentences texts from the sustainability reports and other documents about sustainable or unsustainable activities. The aim of sub-task2 is to design a system that can automatically classify the ESG Taxonomy text sentence into sustainable or unsustainable class.

As sub-task1, we make use of on-line data such as Wikipedia data to augment the financial terms with terms' definition to be term sentence. Moreover, we combine the given dataset composed of financial and non-financial reporting documents files with augmented terms' sentences to train word2vec with the context-free Word2vec model. The Logistic Regression and Deep Attention Model [Tian and Chen, 2021] by inputting word2vec and tf-idf vectors are implemented to predict the concepts of the test terms. As sub-task2, the Bert, Albert, Distil BERT, Roberta, XLNet are applied to this task. Based on the results of the experiments, the proposed models have achieved good performance for each task.

Section 2 describes the task data and the term augmenta-

Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP), pages 224 - 227

Label	Num	Label	Num
Sustainable Transport	46	Biodiversity	29
Board Independence	27	Waste management	16
Energy efficiency and renewable energy	59	Community	27
Sustainable Food & Agriculture	54	Human Rights	10
circular economy	47	Carbon factor	19
Injury frequency rate for subcontracted labor	59	Share Capital	2
Injury frequency rate	35	Audit Oversight	7
Employee engagement	37	Board Make-Up	23
Employee development	22	Emissions	39
Product Responsibility	51	Future of work	18
Recruiting and retaining employees	11	Human Rights	10
Executive compensation	32	Shareholder rights	38

Table 1: The numbers of each label in training data

tion method. Section 3 describes our proposed methods for two tasks. Section 4 shows experimental configurations and discusses the results. Then, we conclude this paper in Section 5.

2 Data Description and Augmented Terms

As the sub-task1, the number of training and test set data are 647 and 145 respectively. There are 24 categories of concepts in the training data. The number of each concept in training data is listed as Table 1.

Based on the above table, the concept of "Energy efficiency and renewable energy" has the largest number of label data. Some concepts like the "Injury frequency rate"," SHARE CAPITAL" and "Board Independence" have just 2 label data. We found that some concepts are very similar, such as the "Injury frequency rate for subcontracted labor" and "Injury frequency rate" concepts, "Waste management and Water" and "waste-water management" concepts which have common words. Moreover, some terms are composed of a single word like "Strikes", "Contraceptives" terms which are not easy to understand the meaning of terms. We augment the terms with terms' definition in the training and test set data. The Wikipedia terms' definitions are utilized to describe the meaning of the terms. We take the "Recycle" as an example to describe how we augment the term with term's definition. The definition of "Recycling" in the Wikipedia is "Recycling is the process of converting waste materials into new materials and objects". We merge the term word "Recycle" and definition of "Recycling" with a space as a new sentence: "Recycle Recycling is the process of converting waste materials into new materials and objects". Some terms like "Tobacco 5% Revenues" which have intuitive meaning, we keep the term words as sentence without adding additional definition. Since the number of provided training and test data is still limited for training a good word embedding. We combine the given dataset composed of financial and non-financial reporting documents files with augmented term sentences to train the word embedding. There are a total of about 196615 sentences for training word2vec. As the task 2, there are 2265 rows in the training data and 205 rows in the test data. The number of sustainable and unstainable sentences are 1223 and 1042 respectively.

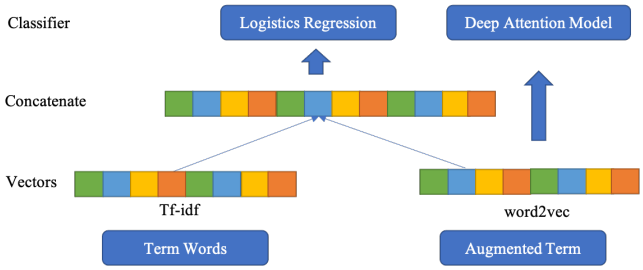

Figure 1: Structure of proposed method for sub-task1.

3 Methods

3.1 TF-IDF Vector, Word Embedding for Sub-task1

As the task1, we mainly use the tf-idf vector and work2vec for creating features. The Logistics Regression and Deep Attention Model are applied for classifier. The overall structure of proposed method is shown in Figure 1.

We observed that some key words in term words are strongly related to concept category, for example, the "low-carbon", "carbon footprint" terms have "carbon" word which indicates the concept is "Carbon factor". The tf-idf is a kind of numerical statistics that could reflect how important a word is to a document in a collection or corpus [wikipedia,]. We extracted key word features for term words using the tf-idf vector. We trained the tf-idf vector with the scikit-learn library's TfidfVectorizer class, we set the 300 dimensions features for the tf-idf vector. We trained the 100 dimensions word2vec using the genism library with augmented term sentence and given financial and non-financial reporting pdf documents. All sentence texts are preprocessed with the following steps: removing stop words, deleting punctuation, and using word stemming to replace word in text. We have implemented two classifiers: Logistics Regression and Deep Attention Model. As the Deep Attention Model, the input vector is word2vec, as the logistics regression, the tf-idf vector and word2vec are concatenated as 400 dimensions for input vector features.

3.2 Pre-trained Language Models for Sub-task2

As sub-task2, we have implemented different PLM models: BERT, Roberta, Albert, DistillBert, and XLNet with related PLM's tokenizer. The sentence label classification has been fine-tuned by adding dropout, linear layer and Relu function after PLM's output as shown in the Figure 2.

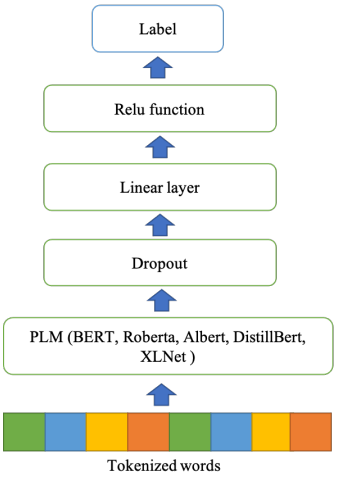

Figure 2: Structure of proposed method for sub-task2.

Input layer	Model	Accuracy
Word2vec trained via train & test data	Logistics Regression	73.84%
Word2vec trained via train & test data + tf-idf vector	Logistics Regression	81.53%
Word2vec trained via train & test data pdf documents + tf-idf vector	Logistics Regression	87.69%
Word2vec trained via train & test data pdf documents	Deep attention model	81.5315%

Table 2: Validation result of each model for sub-task1

4 EXPERIMENT AND RESULT

4.1 Experiment Design

In order to select the best classifier model in the training stage, the label data are split into train and valid data with ratio 9:1 for both two sub-tasks. In the training stage, we mainly test two models: Logistics Regression, Deep Attention Model for sub-task1. We have tested different vector combination for Logistics Regression and Deep Attention Model. As sub-task2, we have implemented the pre-trained models with hugging face library and fine tuning the model with adding linear layer. After the validation stage, the best performance models are selected to predict the test data for final submission.

Input layer	Model	Accuracy
Train & Valid data	BERT	92.1%
Train & Valid data	Roberta	94.0%
Train & Valid data	ALBert	92.29%
Train & Valid data	Distil Bert	91.7%
Train & Valid data	XLNet	93.6%

Table 3: Validation result of each model for sub-task2

Input layer	Model	Accuracy
Word2vec trained via train & test data + tf-idf vector	Logistics Regression	66.2%
Word2vec trained via train & test data pdf documents + tf-idf vector	Logistics Regression	74.48%
Word2vec trained via train & test data pdf documents	Deep Attention Model	75.17%

Table 4: Test result of each model for sub-task1

Input layer	Model	Accuracy
Train & Valid data	BERT	89.75%
Train & Valid data	Roberta	94.63%
Train & Valid data	Distil Bert	89.267%
Train & Valid data	XLNet	92.68%

Table 5: Test result of each model for sub-task2

4.2 Result and Discussion

The result for each model in the experiment is shown in the Table 2 and Table 3. In the validation stage, as sub-task1 we could find that the Logistics Regression based on word2vec and tf-idf vectors achieved better than other classifiers. As the test stage, we submitted 3 model prediction results for test terms. The final score is shown as Table 4 . It can conclude that the Deep Attention Model outperforms than other models in test stage although the accuracy of deep attention model is worse than Logistics Regression's result in the validation stage. As sub-task2, we have implemented different PLM, we found that the Roberta model is the best in the validation stage. In test leader board, Roberta model outperforms obviously better than other three models as shown in Table 5.

5 CONCLUSION

This paper mainly presents kaka team how to tackle the FinSim4 shared tasks. We approach the two tasks using different modes. As sub-task1, we implemented Logistics Regression and Deep Attention Model with different word vectors. As sub-task2, several PLM are implemented with fine tuning to classify the sentences into sustainable or unsustainable class. The experimented result show that our methods could effectively solve the goal of the two tasks. However, our method still needs to be improved to achieve better performance in the following direction. Firstly, it is better to do more parameter tuning in each model to improve the accuracy. Secondly, as sub-task1, there is significant gap between our score and the best result in the final test leaderboard, we could make more efforts in feature engineer like text similarity for models.

Acknowledgments

We would like to thank organizers for holding this shared task and also building the training data. This work is financially supported by the Jiangxi Double Thousand Plan-Long term young innovation project (grant nos.0299/09030022).

References

[Araci, 2019] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[Jeffrey Pennington and Manning, 2014] Richard Socher Jeffrey Pennington and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Lample and Conneau, 2019] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[Organizer,] FinSim4-ESG Organizer. Shared task finsim4-esg. https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022/shared-task-finsim4-esg?authuser=0.

[Radford and Narasimhan, 2018] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. In *Technical report, OpenAI*, 2018.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[Tian and Chen, 2021] Ke Tian and Hua Chen. aiai at the finsim-2 task: Finance domain terms automatic classification via word ontology and embedding. In *The 1st Workshop on Financial Technology on the Web (FinWeb) The Web Conference*, page 320–322, Ljubljana Slovenia, April 2021. Association for Computing Machinery.

[Tian and Peng, 2019a] Ke Tian and Zi Jun Peng. aiai at finnum task: Financial numeral tweets fine-grained classification using deep word and character embedding-based attention model. In *The 14th NTCIR Conference*, pages 198–202, Tokyo, Japan, June 2019.

[Tian and Peng, 2019b] Ke Tian and Zi Jun Peng. aiai at FinSBD task: Sentence boundary detection in noisy texts from financial documents using deep attention model. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 88–92, Macao, China, August 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[wikipedia,] wikipedia. tf-idf. https://en.wikipedia.org/wiki/Tf\OT1\textendashidf.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.