TMUNLP at the NTCIR-16 FinNum-3 Task: Multi-task Learning on BERT for Claim Detection and Numeral Category Classification

Tzu-Ying Chen
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
m946110012@tmu.edu.tw

Yu-Wen Chiu
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
linda037@tmu.edu.tw

Hui-Lun Lin
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
m946110005@tmu.edu.tw

Chia-Tzu Lin
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
m946110009@tmu.edu.tw

Yung-Chung Chang*
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
changyc@tmu.edu.tw

Chun-Wei Tung
Institute of Biotechnology and Pharmaceutical Research,
National Health Research Institutes, Taiwan
cwtung@nhri.edu.tw

ABSTRACT

In financial documents, numerals often contain important information in addition to textual data. As a result, understanding the semantics and relations between numerals and words is of great interest and benefit. The main task of the NTCIR-16 FinNum-3 is fine-grained claim detection, uncovering meaning of numerals in financial reports through claim detection and numeral category classification. We proposed a system that can carry out the two tasks concurrently in this paper. Our evaluation shows that the ensemble fine-tuned BERT has the best performance with a micro averaged F_1-score of 94.67% for the numeral category classification, and a micro averaged F_1-score of 92.75% for the claim detection.

KEYWORDS

BERT, Argument Mining, Ensemble Learning, Multi-task Learning

TEAM NAME

TMUNLP

SUBTASKS

Investor's claim detection (Chinese)

1. INTRODUCTION

With the accelerated development and importance of Artificial Intelligence (AI) observed in many industries, numerous AI-driven applications utilizing recent breakthroughs in machine learning, deep learning, and in particular, natural language processing (NLP) technologies, have emerged in the financial industry. Machine learning and deep learning models can identify dynamic patterns and previously unidentified relations in large quantities of samples, easing the difficulties in handling large amount of data and increasing the analytical value of these data. In addition, more companies and investors are sharing their investment opinions, stock reviews, and recommendations with the general public [1]. This information sharing accelerates technological development in the financial domain as more accurate predictions can be made to assist managers in decision-making related to operation, investment, supervision, and risk control. Financial technology, shortened as "FinTech," is broadly defined as any technology that enables or enhances the provision of financial services. Chen *et al.* [2] defined and categorized FinTech into 7 categories, including (1) cybersecurity; (2) mobile transactions; (3) data analytics; (4) blockchain; (5) peer-to-peer (P2P); (6) robo-advising; and (7) internet of things (IoT). Among these categories, data analytics is comparatively more extensive, including the analysis of financial reports, social media, or news articles. Data analytics can be used to extract a substantial amount of rich and essential information from a wide range of articles.

The two tasks of the FinNum-3 competition are to determine if the given numeral is an in-claim numeral and classify the category of the given numeral. To classify and understand these numerals in noisy and unstructured text, it is necessary to employ NLP techniques. In the past, there are some common methods to classify numbers in text, namely Naïve Bayes [3], Convolutional Neural Networks (CNN) [4], Recurrent Neural Network (RNN) [5], and Long Short-Term Memory (LSTM) [5]. To identify the meaning of numbers, argument mining (AM) is one of the most suitable research directions. AM is a crucial but not yet fully developed aspect of AI and involves various research areas from the AI panorama [6]. Therefore, in this competition, building upon the knowledge and theories of our previous work, we also

* Corresponding author

experimented with many different model-building methods in the AM domain. In our literature review, Maximum Entropy and Support Vector Machines were employed by Mochales *et al.* [7] to carry out AM tasks by building statistical classifiers for the detection and classification of argument propositions, as well as completely defining the AM problem. In the following years, statistical and machine learning models have been widely adopted, whereas the application of deep neural networks to AM problems gradually emerged around 2017, such as the work by Eger *et al.* [8]. Their work found that BiLSTM-based labeling models are able to capture inherent long-term dependencies, therefore, providing satisfactory solutions to AM problems. Finally, Lawrence *et al.* [9] discussed AM-related issues and reviewed existing AM datasets in detail, which serves as a very useful reference for future AM research.

A significant breakthrough in the field of NLP recently is the Bidirectional Encoder Representations from Transformers (BERT), which is a language model pre-trained on a large amount of unsupervised data. It contains a multi-layer encoder with bidirectional Transformer blocks, and is trained with masked word prediction and next sentence prediction objectives. BERT demonstrates the ability to learn common language representations, and is shown to be effective for many natural language understanding (NLU) tasks [10]. Similarly, BERT was used for the classification of financial texts. Chen *et al.* [11] proposed a joint intent classification and slot filling model based on BERT to improve the generalization capability of traditional NLU models. The performance of their method showed substantial improvement and evidenced that the BERT model is feasible for the classification task. Liang *et al.* [12] during the NTCIR-15 FinNum-2, proposed using the BERT embedding outputs as the input to a BiLSTM layer followed by an Attention layer, which can perform efficient classification when correlations exist between numbers and cashtags in financial social media data.

In this research, we utilized several pre-trained BERT models and apply fine-tuning for downstream tasks in the NTCIR FinNum3. We propose a multi-task learning approach on top of BERT for concurrent claim detection and numeral category classification.

Table 1. Statistics of analyst's report annotations. (IC: In-Claim; OC: Out-of-Claim)

Category		Train		Development		Test		Total
		IC	OC	IC	OC	IC	OC	
Monetary	Money	428	311	78	57	413	-	1,287
	Change	3	3	-	12	362	-	380
	Price	34	32	8	1	30	28	133
Percentage	Relative	326	335	82	67	351	452	1,613
	Absolute	171	394	37	106	169	572	1,449
Temporal	Date	-	1,775	-	359	-	1,847	3,981
	Time	-	3	-	-	-	1	4
Quantity	Absolute	36	183	19	36	40	165	479
	Relative	-	4	-	-	3	16	23
Product Number		1	100	-	35	1	145	282
Ranking		-	-	-	3	-	6	9
Other		-	80	-	25	-	90	195
Total		999	3,220	224	701	1,369	3,322	9,835

2. MATERIAL AND METHOD

2.1. Dataset

The dataset used in this competition consists of professional finance analysts' reports written in Chinese. Table 1 shows the statistics of the given data. There is a total of 9,835 entries in this dataset, in which 4,219 of the entries are for training, 925 are for development, and 4,691 are for testing. The dataset contains five columns: (1) text, (2) offset, (3) target numeral, (4) claim, and (5) category. The detailed definition of the claim and categories can be found in Chen *et al.* [13].

2.2. Method

During our participation in the FinNum-3, we developed a system that can carry out the two tasks simultaneously. The architecture of the proposed system is shown in Figure 1. The first task is to determine whether the target numeral is an in-claim numeral, and the second task is to classify which category the target numeral belongs to. When text is inputted into the system, the system will extract financial context-dependent features. Then, the knowledge-based component, which is designed to deal with data sparsity, produced other related features. The 4 categories with the least samples are filtered out, and the remaining samples enter the pre-trained BERT model for finetuning. Through combining different designated features and target numerals, several different inputs are passed through the model, and the output results, categories, and claims can be jointly predicted by our system.

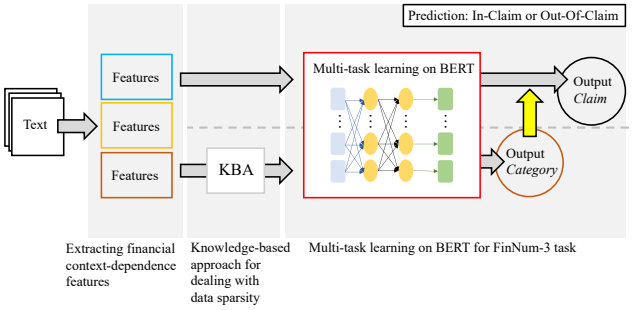

Figure 1. Overall architecture of the proposed multi-task machine learning framework.

Extracting financial context-dependent features

The key issue is how to accurately uncover the meaning of the numerals in the financial context. We discover the following financial context-dependent (FCD) features through literature research and theoretical analysis. Since the tasks are performed on a Chinese data set, we can use a single character to form a unit (which may not necessarily be meaningful) instead of using a word. For word units, the features we create may have the following forms:

- The first two characters or the last two characters of the target numeral

- The first three characters or the last three characters of the target numeral
- From the target numeral to the start/end of the sentence it exists

After rigorous testing of the above features, we found that "the last two characters of the target numeral (a)", "the target numeral and two characters before and after it (b)", and "from the target numeral to the start of the sentence it exists (c)", are three features that can be key to the success of the model. It appears that their effects are superior to using other features. Finally, we enter two feature combinations FCD 1 and FCD 2 into the model, where FCD 1 is (b)+(c), and FCD 2 is (a)+(b)+(c). We also conducted the Log-Likelihood Ratio test with keywords of finance and economics. In addition, we also used the units behind the numeral, such as "%" or "$", to conduct keyword-oriented models, but the results were not satisfactory. In the end, a more sophisticated feature engineering was conducted and we found that although humans may sometimes find words to be meaningless, but machines are capable of uncovering the hidden relations. An example of our FCD method is shown in Table 2.

Table 2. Example of financial context-dependent (FCD) features.

Examples	**……，EPS1.05 元創下歷史新高。我們預估 2018 年營收為 289.6 億元，毛利率 35%，淨利 *59.63* 億元，EPS3.64 元。(……, EPS1.05 dollar achieved an all-time high. We estimate that revenue in 2018 will be 289.6 billion, with a gross profit margin of 35%, net income of *59.63* billion, EPS3.64 dollar.)**
Target numeral	59.63
feature (a)	億元 (billion)
feature (b)	淨利 59.63 億元 (net income of 59.63 billion)
feature (c)	我們預估 2018 年營收為 289.6 億元，毛利率 35%，淨利 (We estimate that revenue in 2018 will be 289.6 billion, with a gross profit margin of 35%, net income of)

Knowledge-based approach for dealing with data sparsity

As mentioned in the Dataset section above, the uneven distribution of data is a serious issue. Since the amount of data in some categories is too small, data sparsity and unsatisfactory classification accuracy can happen. As such, data cleaning is carried out with filtering of outlier-like categories, which enables the model to learn with the cleaner data and obtain better outcomes. However, although the samples belonging to the smaller categories are few, we may lose prominent textual patterns. In this regard, we created the knowledge-based approach (KBA) to solve this problem. For the dataset provided by the organizer, KBA filtered out the four categories with the fewest samples at the beginning of our proposed system, namely, *Change*, *Time*, *Ranking*, and *Quantity*. Next, with the extra features, KBA assists the model to perform better on numeral category classification and claim detection. For the category *Ranking*, since it only accounted for 0.058% of the total data and it represents performance, we used professional knowledge to match the observation of linguistic pattern, and used regular notation to extract patterns such as "前___大" (the first __), "前___名" (the first __ position), etc. In KBA, we designed different mechanisms for different categories, and finally achieved the best approach.

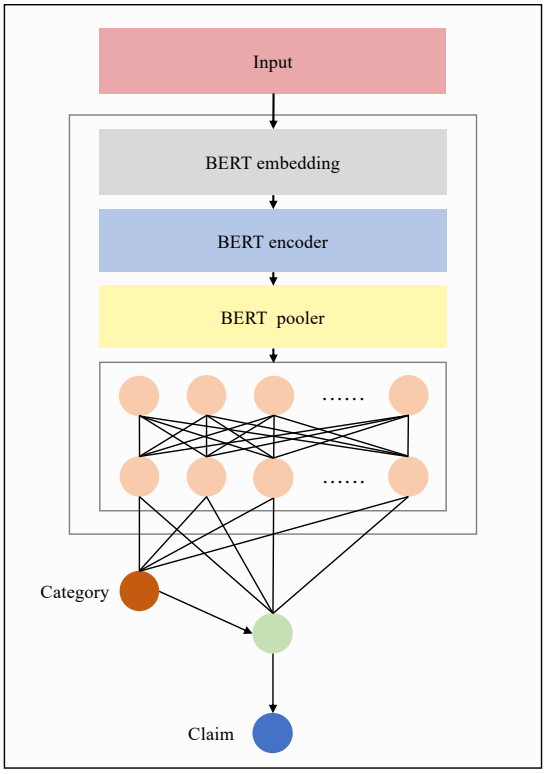

Figure 2. Model architecture.

Multi-task learning on BERT for FinNum-3 task

We propose a method to construct a fine-tuned BERT model to carry out the two tasks concurrently [10]. The pre-trained BERT models were obtained from https://huggingface.co/models, and their hyperparameters were finetuned as the final models architectures were decided. The training of our final architecture was performed on a single machine with one NVIDIA TITAN RTX GPU. It involves a BERT pooler output with the following layers: Dropout (with a dropout probability of 0.3), Linear (fully connected) with hidden sizes in-768 or 1024 for BERT-base or BERT-large, respectively, and 8 for the output that predicts the 8 categories. As for the main task, it concatenates a BERT pooler output to the 8 predicted categories with the following layers: ReLU, Linear (fully connected): size (in-768+8 or 1024+8 for BERT-base or BERT-large, out-2). For the loss, we used cross-entropy. The optimizer was ADAM [14], the learning rate 1e-05,

and a batch size of 16. The max length of the input is 50. This study is implemented with Python (3.6.13) using the Pytorch library (1.7.1) and PyTorch-Transformers (1.2.0). Finally, we selected the pre-trained BERT from Harbin Institute of Technology Xunfei Laboratory and Koichi Yasuoka as our fine-tuning BERT models to enhance the performance on textual multi-task classification [15]. The architecture of our proposed multi-task BERT model is shown in Figure 2.

3. EXPERIMENTAL RESULT AND DISCUSSION

Micro F_1 and macro F_1 scores are used to evaluate the experimental results for claim detection and numeral category classification. In our experiments, we evaluated different pre-trained models and features to predict whether the target numeral is an in-claim numeral and classification of the 8 numeral categories. Table 3 shows our results. We first tried Koichi Yasuoka's chinese-roberta-base-upos pre-trained model with FCD 1, and achieved a micro F_1-score of 90.98% for claim detection and a macro F_1-score of 88.44 % for numeral category classification. To compare the performance of the different models, we next tried FCD 1 with Koichi Yasuoka's chinese-roberta-large pre-trained model and hfl's chinese-roberta-wwm-ext pre-trained model respectively. The Koichi Yasuoka's chinese-roberta-large-upos pre-trained model had a micro F_1-score of 91.42% for claim detection and a macro F_1-score of 89.37% for numeral category classification. As for the hfl's chinese-roberta-wwm-ext pre-trained model, it achieved a micro F_1-score of 91.64% for claim detection and a macro F_1-score of 85.26% for the numeral category classification. From the above experiments, FCD 1 with the hfl's chinese-roberta-wwm-ext pre-trained model has the best performance in the claim detection, and FCD 1 with the Koichi Yasuoka's chinese-roberta-base-upos pre-trained model has the best performance in the numeral category classification, with the micro F_1-score of 91.64% and the macro F_1-score of 89.37% respectively.

Next, we tried these three pre-trained models with FCD 2. The micro F_1-scores for claim detection of the Koichi Yasuoka's chinese-roberta-base-upos, the Koichi Yasuoka's chinese-roberta-large-upos and the hfl's chinese-roberta-wwm-ext pre-trained model are 92.1%, 90.6% and 91.3% respectively. The macro F_1-score for numeral category classification of the Koichi Yasuoka's chinese-roberta-base-upos, the Koichi Yasuoka's chinese-roberta-large-upos and the hfl's chinese-roberta-wwm-ext pre-trained model are 86.77%, 87.79%, and 87.56% respectively. In contrast, FCD 2 with the Koichi Yasuoka's chinese-roberta-base-upos has the best performance in the claim detection and FCD 2 with the Koichi Yasuoka's chinese-roberta-large-upos has the best performance in the numeral category classification.

Overall, for the numeral category classification, the Koichi Yasuoka's chinese-roberta-large-upos pre-trained model with FCD 1 had the best macro F_1-score of 89.37%. After passing FCD 1, which has word unit comprising of the target numeral to the start of the sentence it exists, through the BERT model, key information about the category can be obtained through self-attention mechanism. Therefore, the prediction of the numeral category classification can be improved greatly. As for the claim detection, the Koichi Yasuoka's chinese-roberta-base-upos pre-trained model with FCD 2 had the best micro F_1-score of 92.19%. FCD 2 contains word unit comprising more than two characters next to the target numeral as compared to FCD 1, and the next two words usually contain some crucial information pertaining to claim, such as the status or unit of the target numerals, therefore, FCD2 can improve the performance of claim detection by the Koichi Yasuoka's base pre-trained model.

Table 3. Performance of compared methods on 8 categories data in the development set.

		Claim Detection		Numeral Category	
Pre-trained Model	Feature	micro F_1 (%)	macro F_1 (%)	micro F_1 (%)	macro F_1 (%)
Koichi Yasuoka base	FCD 1	90.98	87.55	92.52	88.44
Koichi Yasuoka large	FCD 1	91.42	88.62	91.97	89.37
hfl chinese-roberta	FCD 1	91.64	88.78	90.43	85.26
Koichi Yasuoka base	FCD 2	92.19	89.53	91.86	86.77
Koichi Yasuoka large	FCD 2	90.65	87.54	91.75	87.90
hfl chinese-roberta	FCD 2	91.31	88.28	91.86	87.56

We selected the best combination of feature and pre-trained model for the claim detection and the numeral category classification as two of the three submissions. For the third submission, we employed ensemble learning with the majority voting mechanism to integrate these three fine-tuned BERT models. FCD 1 is for numeral category classification. FCD 2 is for claim detection. The result of three rounds of submission is shown in Table 4. The difference between Table 3 and Table 4 is that Table 3 show only the results of the classifications of 8 numeral categories by multi-task BERT model. Table 4 show the result of the full dataset, in which 4 numeral categories are classified by our KBA approach, and the remaining 8 numeral categories are classified by multi-task BERT model prediction.

Round 1: Koichi Yasuoka's chinese-roberta-large-upon pre-trained model with FCD

In Round 1, we used the Koichi Yasuoka's chinese-roberta-large-upos as the pre-trained model and FCD 1 as the input text, which is the best combination for the numeral category classification in our experiment. The performance in Round 1 is shown in Table 4. For numeral category classification of the development set, the micro F_1-score is 91.56% and the macro F_1-score is 89.14%. For claim detection of the development set, the micro F_1-score is 91.56% and the macro F_1-score is 88.68%.

Round 2: Koichi Yasuoka's chinese-roberta-base-upon pre-trained model with FCD 2

In Round 2, we used the Koichi Yasuoka's chinese-roberta-base-upos as the pre-trained model and FCD 2 as the input text, which is the best combination for the claim detection in our experiment. The performance in Round 2 is shown in Table 4. For claim detection of the development set, the micro F_1-score is 92.32% and the macro F_1-score is 89.59%. For numeral category classification of the development set, the micro F_1-score is 91.45% and the macro F_1-score is 87.07%.

Round 3: Ensemble fine-tuned BERT model

In Round 3, we utilized ensemble learning with the majority voting mechanism to integrate fine-tuned BERT models. The performance in Round 3 is shown in Table 4. For claim detection of the development set, the micro F_1-score is 92.43% and the macro F_1-score is 89.84%. For numeral category classification of the development set, the micro F_1-score is 92.43% and the macro F_1-score is 89.64%.

Table 4. Performance of our methods on the full development set.

	Development set			
	Claim Detection		Numeral Category	
Submission ID	micro F_1	macro F_1	micro F_1	macro F_1
TMUNLP 1	91.56%	88.68%	91.56%	89.14%
TMUNLP 2	92.32%	89.59%	91.45%	87.07%
TMUNLP 3	92.43%	89.84%	92.43%	89.64%

Table 5. Performance of our methods on the test set.

	Test set			
	Claim Detection		Numeral Category	
Submission ID	micro F_1	macro F_1	micro F_1	macro F_1
Baseline	80.32%	69.19%	62.59%	20.99%
TMUNLP 1	92.82%	89.56%	94.31%	73.68%
TMUNLP 2	91.11%	87.76%	94.03%	72.99%
TMUNLP 3	92.75%	89.68%	94.67%	73.89%

Table 5 shows the official result of the baseline and our submissions. For claim detection, the micro F_1-score is about 91-92% and the macro F_1-score is about 87-89%. For numeral category classification, the micro F_1-score is about 94% and the macro F_1-score is about 72-73%. Compared with the baseline, which was results of Capsule Neural Network models returned by organizers, the final performance can be improved greatly by our method, which demonstrated that (1) KBA is effective in dealing with the sparse data, (2) extracting keywords and key sentences from contexts as features can be beneficial, and (3) utilizing multi-task learning and ensemble learning on BERT is effective.

4. Conclusion

In this paper, we proposed a system that can concurrently carry out both claim detection and numeral category classification of numerals in professional finance analysts' reports [16]. Among the three proposed models, the model with the ensemble fine-tuned BERT as the pre-trained model and FCD 2 as the input feature set had the best performance, with a micro F_1-score of 94.67% for the numeral category classification task, and a micro F_1-score is 92.75% for the claim detection task.

REFERENCES

1. Wang, W., Liu, M., Zhang, Y., Xiang, J., & Mao, R. (2019, June). Financial numeral classification model based on BERT. In *NII Conference on Testbeds and Community for Information Access Research* (pp. 193-204). Springer, Cham.
2. Chen, M. A., Wu, Q., & Yang, B. (2019). How valuable is FinTech innovation?..T*he Review of Financial Studies, 32*(5), 2062-2106.
3. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques Third Edition [M]. *The Morgan Kaufmann Series in Data Management Systems, 5*(4), 83-124.
4. Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. *In European conference on computer vision* (pp. 818-833). Springer, Cham.
5. Lai, S., Xu, L., Liu, K., & Zhao, J. (2015, February). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence.*
6. Cabrio, E., & Villata, S. (2018, July). Five Years of Argument Mining: a Data-driven Analysis. In *IJCAI* (Vol. 18, pp. 5427-5433).
7. Mochales, R., & Moens, M. F. (2011). Argumentation mining. *Artificial Intelligence and Law, 19*(1), 1-22.
8. Eger, S., Daxenberger, J., & Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104.*
9. Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics, 45*(4), 765-818.
10. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*
11. Chen, Q., Zhuo, Z., & Wang, W. (2019). Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909.*
12. Liang, Y. C., Huang, Y. H., Cheng, Y. Y., & Chang, Y. C. (2020). TMUNLP at the NTCIR-15 FinNum-2. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.
13. Chen, C. C., Huang, H. H., & Chen, H. H. (2020, October). NumClaim: Investor's Fine-grained Claim Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1973-1976).
14. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*
15. Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, & Kazunori Fujita. (2022) Designing Universal Dependencies for Classical Chinese and Its Application. *Journal of Information Processing Society of Japan, Vol.63, No.2,* pp.355-363.
16. Chen, C. C., Huang, H. H., & Chen, H. H. (2020, October). NumClaim: Investor's Fine-grained Claim Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 1973-1976).