IMNTPU at the NTCIR-16 FinNum-3 Task: Data Augmentation for Financial Numclaim Classification

Yung-Wei Teng
Information Management
National Taipei University
New Taipei City, Taiwan
s711036115@gm.ntpu.edu.tw

Pei-Tz Chiu
Information Management
National Taipei University
New Taipei City, Taiwan
s711036103@gm.ntpu.edu.tw

Ting-Yun Hsiao
Information Management
National Taipei University
New Taipei City, Taiwan
s711036112@gm.ntpu.edu.tw

Mike Tian-Jian Jiang
Zeal Co.,Ltd
Tokyo, Japan
tmjiang@gmail.com

Min-Yuh Day*
Information Management
National Taipei University
New Taipei City, Taiwan
myday@gm.ntpu.edu.tw

ABSTRACT

This paper provides a detailed description of IMNTPU team at the NTCIR-16 FinNum-3 shared task in formal financial documents. We proposed the use of the XLM-RoBERTa-based model with two different approaches on data augmentation to perform the binary classification task in FinNum-3. The first run (i.e., IMNTPU-1) is our baseline through the fine-tuning of the XLM-RoBERTa without data augmentation. However, we assume that presenting different data augmentations may improve the task performance because of the imbalance in the dataset. Accordingly, we presented double redaction and translation method on data augmentation in the second (IMNTPU-2) and third (IMNTPU- 3) runs, respectively. The best macro-F1 scores obtained by our team in the Chinese and English datasets are 93.18% and 89.86%, respectively. The major contribution in this study provide a new understanding toward data augmentation approach for the imbalanced dataset, which may help reduce the imbalanced situation in the Chinese and English datasets.

KEYWORDS

Data Augmentation, Double Redaction, Binary Classification, XLM-RoBERTa, Financial Claim

TEAM NAME

IMNTPU

SUBTASKS

Investor’s Claim Detection (Chinese)
Manager’s Claim Detection (English)

1 INTRODUCTION

Data mining has usually been a crucial task in various domains, especially in finance. Researchers have proposed dynamic methods for understanding (extracting) a deeper information from texts. In previous studies, numerals provide an essential contribution to infer from the textual information in the finance domain. FinNum-1 [3] and FinNum-2 [4] focused on analyzing the targeted numerals in the financial social media. However, in FinNum-3, we consulted professional analyst’ s reports and earning conference calls from multilingual datasets.[1]

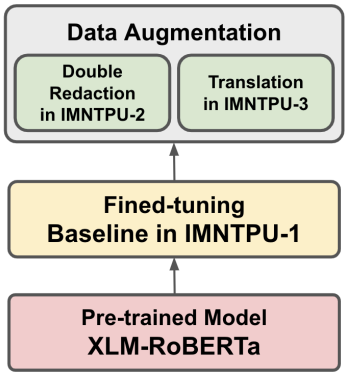
Data Augmentation
Double Redaction in IMNTPU-2
Translation in IMNTPU-3
Fined-tuning Baseline in IMNTPU-1
Pre-trained Model XLM-RoBERTa

Figure 1: The proposed research architecture of IMNTPU at the NTCIR-16 FinNUM-3.

The problem of the shared task as a binary classification is defined to identify the estimation interpreted by the managers or the investors and to classify whether the target numeral in the given claims made by them belongs to “in-claims” or “out-claims.” However, the label distribution of the datasets we collected are heavily imbalanced. Thus, to address the problem of imbalanced label distribution, we adopt two data augmentation methods.[7]

The remainder of this paper is structured as follows. Section 2 presents the detailed approaches in multilingual datasets of the shared task. It includes which model we applied and how we did feature engineering. Section 3 provides the experimental result, the configuration of the models, and the error analysis of the result. Finally, Section 4 presents the conclusion and the future work.

2 PROPOSED METHOD

In this section, we will present our models and the applied techniques for feature engineering in detail. The remainder of this section is organized as follows. Section 2.1 provides an introduction of XLM-RoBERTa adopted in this work. Section 2.2 discusses the details of the double redaction for data augmentation. Section 2.3 presents the translation technique for another data augmentation method. The proposed research architecture of IMNTPU at the NTCIR-16 FinNum-3 is shown in Figure 1.

2.1 XLM-RoBERTa Based Model

We adopt XLM-RoBERTa [5], which is multilingual version of RoBERTa, so that we can utilize this model in our multilingual datasets. The training method of XLM-RoBERTa is based on BERT model [6]. The difference between XLM-RoBERTa and BERT is that the latter trains Masked Language Models (MLM) and Next Sentence Prediction (NSP) tasks whereas the former omits the NSP task, which aims to learn whether the two sentences are connected. The XLM-RoBERTa model focuses on the MLM task, which masks the word in sentences so that the mask pattern will change dynamically when training the model using the training dataset.

2.2 Double Redaction

We adopt data augmentation techniques for feature engineering because of a few given textual data in the training dataset. Double redaction aims to change some token as <mask> or <unk> token randomly in sentences to enable the model to learn more patterns about textual analysis in the training phase. For <mask> and <unk> token, the former means covering the original token whereas the latter denotes the conversion of the original token into an unknown token. Therefore, these tokens are regarded as noise.

The major concepts and steps of double redaction as follows. First, we set a list of special tokens, such as <xxnum>, <s> and </s>, which are unchangeable token α in each training data. Second, we shuffle and remove the duplicated tokens in a sentence. Then, the remaining tokens are copied as a set of double redaction tokens β. Moreover, we set a float as an unknown token probability δ and given a random number γ. If a random number is less than the unknown token probability, then noise will be set as <unk>. If not, then the noise will be <mask>. Lastly, the specific token in double redaction tokens will be replaced with noise, and the model will predict the original token, which is replaced with <mask> or <unk>. The double redaction algorithm is shown in Algorithm 1. Briefly, the double redaction approach will let a model learn more for the analysis of text so that great results can be expected on the financial Numclaim classification task.

Algorithm 1 An algorithm of double redaction

```
Shuffle the tokens in sentence
Delete the duplicated tokens in sentence
Copy the remaining tokens as $\beta$
SET the $\delta$ and $\gamma$
for specific token in $\beta$ do
    if $\gamma$ less than $\delta$ then
        Replace original token with <usk> token
    else
        Cover original token as <mask> token
    end if
end for
while True do
    Model predict the original token of <usk> and <mask>
end while
```

Table 1: Configuration of the hardware and software

Items	Version
System Type	X64
Processor	AMD® Ryzen R9 3900X CPU
RAM	64GB
Display Card	NVIDIA GeForce GTX 3090 24GB
OS	Ubuntu 18.04
Python Version	3.6.9

2.3 Translation

During the data generation process, we first translate the textual training data into English and into Simplified Chinese afterward. Based on the synonymous word translation techniques, we acquired additional training data without losing the original meaning. For example, in the original training set, ”税後純益 9.81 億元，YoY+36.36%，税後 EPS2.62 元，優於預期。” (The tax proceeds were $981 million, YoY+36.36 percent and EPS 2.62 percent, higher than expected.) which is written in Traditional Chinese will become “税后净利润为 9.81 亿美元，YoY+36.36%，扣除 ESP 2.62 税后利润比预期的要高。” which is written in Simplified Chinese through the translation process. The pipeline of the translation approach for data augmentation is shown in Figure 2. We also reset the offset position of the target number.

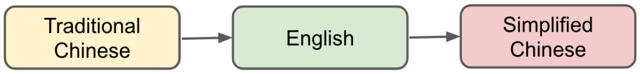

Figure 2: The pipeline of translation approach for data augmentation.

3 EXPERIMENT RESULTS & DISCUSSION

In this section, we listed our model configuration and our experimental result in detail. The experiments below run under the configuration of software and hardware (Table 1).

We present the experimental result for the Chinese and English datasets in each run. In IMNTPU-1 for Chinese dataset, we adopted XLM-RoBERTa model in fastai version. With the One-Cycle schema in fastai version, we run four cycles by using a batch-size of 8 and 2 epochs, 3 epochs, 1 epoch, and 1 epoch in sequence. As for IMNTPU-2, to solve the issue of few datasets, we utilized double redaction for data augmentation approaches. Lastly, for IMNTPU-3, we applied the translation pipeline presented by Hugging Face as our special data augmentation methods. The results of Chinese development test and test set is shown in Table 2.

In addition, in IMNTPU-1 for English dataset, we used XLM-RoBERTa model as our baseline model by adjusting the parameters and optimizing the hyper-parameter of the fastai version. Furthermore, we proposed the double redaction approach for data augmentation. The result is shown in Table 3.

The remainder of the Section 3 is structured as follows. Section 3.1 provides a comprehensive result in IMNTPU-1 for the Chinese and English datasets. Section 3.2 provides the comparison result in

Table 2: IMNTPU macro F1-score result in both development and test set in Chinese dataset

	Dev Set F1-score (%)	Test Set F1-score (%)
IMNTPU-1 (Baseline)	90.51	**93.18**
IMNTPU-2 (Double Redaction)	88.65	91.64
IMNTPU-3 (Translation)	**92.16**	91.64

Table 3: IMNTPU macro F1-score result in both development and test set in English dataset

	Dev Set F1-score (%)	Test Set F1-score (%)
IMNTPU-1 (Baseline)	87.13	88.39
IMNTPU-2 (Double Redaction)	88.82	**89.86**

IMNTPU-2 for the Chinese and English datasets. Section 3.3 only presents the result for the Chinese dataset in IMNTPU-3. Finally, Section 3.4 provides the error analysis for explaining why the performance will increase or decrease via different data augmentation approaches.

3.1 Baseline in IMNTPU-1

For IMNTPU-1 in both Chinese and English dataset, we adopt XLM-RoBERTa model in fastai version. In addition, we set up the class weight which is the inverse of the class distribution to solve the issue of unbalanced label distribution in training dataset. As for the optimizer of the model, we simply use Lamb optimizer which has improved the performance and greatly decrease the training time for training a transformer-based model[8].

The performance of IMNTPU-1 in Investor’ s Claim (Chinese) and Manager’ s Claim (English) are shown in Tables 2 and 3, respectively. In the Chinese dataset, the macro F1-score for the development set is 90.51% and 93.18% for the test set. As for the English dataset, the macro F1-score for the development test is 87.13% and 88.39% for the test set. The performance of the test set outperforms that of the development set, and we realized that the label distribution and the textual data of the test set is more similar to the training set.

3.2 IMNTPU-2 Double Redaction

For IMNTPU-2 in the Chinese and English datasets, we use double redaction for data augmentation approaches. We inputted the augmented dataset into the XLM-RoBERTa model, and the training cycle, epoch numbers, batch size, and the optimizer configuration remain the same. In the Chinese dataset (Table 2), the macro F1-score for the development set is 88.65% and 91.64% for the test set. As for the English dataset (Table 3), the macro F1-score for the development set is 88.82% and 89.86% for the test set.

3.3 IMNTPU-3 Translation

For IMNTPU-3, we only focused on the Chinese dataset. To determine whether the performance will raise under various data augmentation approaches, we applied the translation pipeline presented by Hugging face. The macro F1-score for the development set is 92.16% and 91.64% for the train set (Table 2).

3.4 Error Analysis

In the English and Chinese datasets, we discover some instances with wrong prediction in all runs. First, we deciphered that the instance with a certain pattern often shows a lower correction. Second, we realized that the category may affect the prediction, wherein several certain categories show lower accuracy, and the distribution is shown in the following table.

In the first instance, we discussed about the target number shown in the pattern of “xx 到 xx,” “xx 至 xx,” “xx to xx,” and “xx - xx” in the text accounts for the vast majority of our incorrect prediction, in which the former target number and the latter one are in the same category, and once the model predict the former incorrectly, the latter usually went wrong too, thereby reducing accuracy.

As for another instance, we count all the incorrect prediction categories in each run. We find out that among all the categories in the Chinese and English datasets, the most categories are absolute, money, relative, and quantity_absolute. However, compared with the train set of the Chinese and English datasets, the aforementioned categories do not take the most portion of the train set. In addition, our methods did not consider the “category” as our input parameter. Therefore, it suggests one of the future research directions, which is the use of the rule-based name entity recognition method to identify the categories as the model input.

4 CONCLUSION & FUTURE WORK

4.1 Conclusion

We proposed our approach for financial Numclaim classification based on XLM-RoBERTa model for Chinese and English datasets in three runs as follows: baseline, double redaction, and translation. In the Chinese dataset, our results have reached the highest macro F1-score among all the participants. As for the English dataset, the macro F1-score of the one with data augmentation earn the second highest F1-score among all the teams.[2]

In conclusion, the results of our approach show that we can achieve a great performance based on data augmentation for financial Numclaim classification in multilingual dataset. Due to the lack of sampling data and the language structure restriction, utilizing the same data augmentation approach in different language datasets is difficult.

The major contribution of the research is that the data augmentation approach may help reduce imbalanced situation. The managerial implication of the paper is that the method we presented may help the public pay more attention on the estimation from the investors and the managers.

4.2 Future Work

Our approach was limited by the different unbalanced label distributions between the train set and the test set. For this reason, more information on the sampling dataset would help us to establish a greater state-of-the-art model for financial Numclaim classification on this matter.

With regard to IMNTPU-1 and IMNTPU-2, the data augmentation approach seems to show different results for the financial Numclaim classification in Chinese and English. The double redaction approach has greatly improved the performance of IMNTPU-1 and IMNTPU-2 in the English dataset. In contrast, based on the different language structures described in Section 2.2, the performance of the Chinese dataset for double redaction does not improve the performance. Therefore, more experiments must be conducted if the token delimiter for double redaction and translation approaches for data augmentation changes based on its language feature in the Chinese and English datasets to determine performance improvement.

ACKNOWLEDGMENTS

This research was supported in part by the Ministry of Science and Technology (MOST), Taiwan under grant number 110-2410-H-305-013-MY2, and National Taipei University (NTPU) under grant number 110-NTPU-ORDA-F-001, 111-NTPU-ORDA-F-001, and 111-NTPU-ORDA-F-003.

REFERENCES

[1] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. NumClaim: Investor's Fine-grained Claim Detection. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* 1973–1976.

[2] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. Overview of the NTCIR-16 FinNum-3 Task: Investor’s and Manager’s Fine-grained Claim Detection. *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan.*

[3] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies.* 19–27.

[4] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 Task: Numeral attachment in financial tweets. *Development* 850, 194 (2020), 1–044.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Mike Tian-Jian Jiang, Yi-Kun Chen, and Shih-Hung Wu. 2020. CYUT at the NTCIR-15 FinNum-2 task: tokenization and fine-tuning techniques for numeral attachment in financial tweets. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies.* 92–96.

[8] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962* (2019).