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

Yung-Wei Teng [1], Pei-Tz Chiu [1], Ting-Yun Hsiao [1], Mike Tian-Jian Jiang [2] and Min-Yuh Day [1,*]

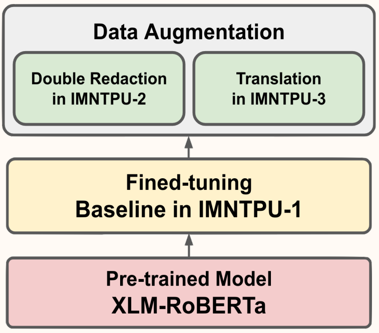
[1] Information Management, National Taipei University, New Taipei City, Taiwan

[2] Zeals Co., Ltd. Tokyo, Japan

[*] myday@gm.ntpu.edu.tw

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.

Research Architecture and Proposed Method



Data Augmentation

Double Redaction in IMNTPU-2

Translation in IMNTPU-3

Fined-tuning Baseline in IMNTPU-1

Pre-trained Model XLM-RoBERTa

IMNTPU1: We adopted XLM-RoBERTa Model without data augmentation as our baseline model.

IMNTPU2: We adopt Double Redaction approach for data augmentation and XLM-RoBERTa Model.

IMNTPU3: We adopt the Translation approach for data augmentation and XLM-RoBERTa Model.

Tokenization Tricks

Input: `Good day and welcome to the Apple Inc. Third Quarter Fiscal Year 2018 Earnings Conference Call. Today's call is being recorded.`

XLM-RoBERTa Tokenizer

Output: `<s> Good day and welcome to the Apple Inc. Third Quarter Fiscal Year xxnum 2018 Earnings Conference Call. Today's call is being recorded. </s>`

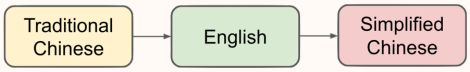
Double Redaction

Output: `<s> <mask> Good day and <mask> to the Apple <mask> Third Quarter Fiscal Year xxnum 2018 Earnings Conference Call. Today's call is <mask> recorded. </s>`

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
```

Translation Approach



Traditional Chinese	English	Simplified Chinese
“ 稅後純益 9.81 億元，YoY+36.36-%，稅後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. ”	“ 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。”

Performance

	Chinese Dataset		English Dataset	
Run	Dev Set F1-Score (%)	Test Set F1-Score (%)	Dev Set F1-Score (%)	Test Set F1-Score (%)
IMNTPU1	90.51	**93.18**	87.13	88.39
IMNTPU2	88.65	91.64	88.82	**89.86**
IMNTPU3	**92.16**	91.64	-	-

Conclusions and Contribution

Conclusions:

The performance with data augmentation method (Double Redaction) in English dataset is superior than without data augmentation.

Contribution:

- The major contribution of the research is that data augmentation approach may help reduce imbalanced situation.
- We have developed a novel method for data augmentation technique, which is double redaction and translation approach, and can decrease the issue of imbalanced dataset.

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.

Information Management, National Taipei University

NTCIR-16 Conference, June 14-17, 2022, Tokyo, Japan

國立臺北大學 National Taipei University