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 provides 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**

**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**

**Double Redaction**

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

### Algorithm of Double Redaction

1. Shuffle the tokens in sentence
2. Delete the duplicated tokens in sentence
3. Copy the remaining tokens as β
4. SET the δ and γ
5. for specific token in β do
6. if γ less than δ then
7. Replace original token with <unk> token
8. else
9. Cover original token as <mask> token
10. end if
11. end for
12. while True do
13. Model predict the original token of <unk> and <mask>
14. end while

### Translation Approach

**Traditional Chinese**

“税後純益 9.81 億元，YoY+36.36%，稅後 EPS2.62 元，優於預期。”

**English**

“The tax proceeds were $981 million, YoY+36.36% and EPS 2.62, higher than expected.”

**Simplified Chinese**

“稅後净利润为 9.81 亿美元，YoY+36.36%；扣除 ESP 2.62 税后利润比预期的要高。“

### Performance

<table>
<thead>
<tr>
<th>Run</th>
<th>Dev Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMNTPU1</td>
<td>90.51</td>
<td>93.18</td>
</tr>
<tr>
<td>IMNTPU2</td>
<td>88.65</td>
<td>91.64</td>
</tr>
<tr>
<td>IMNTPU3</td>
<td>92.16</td>
<td>91.64</td>
</tr>
</tbody>
</table>

### 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 the 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.