

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

¹Institute of Information Management, National Taipei University, New Taipei City, Taiwan

²Zeals Co., Ltd. Tokyo, Japan



Yung-Wei Teng¹



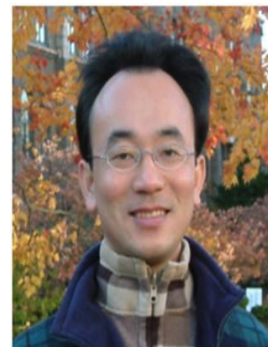
Pei-Tz Chiu¹



Ting-Yun Hsiao¹



Mike Tian-Jian Jiang²



Min-Yuh Day^{1, *}

* myday@gm.ntpu.edu.tw

Outline

- IMNTPU Research Architecture
- IMNTPU Proposed Method
- Performance
- Conclusions

Highlights

- **IMNTPU**

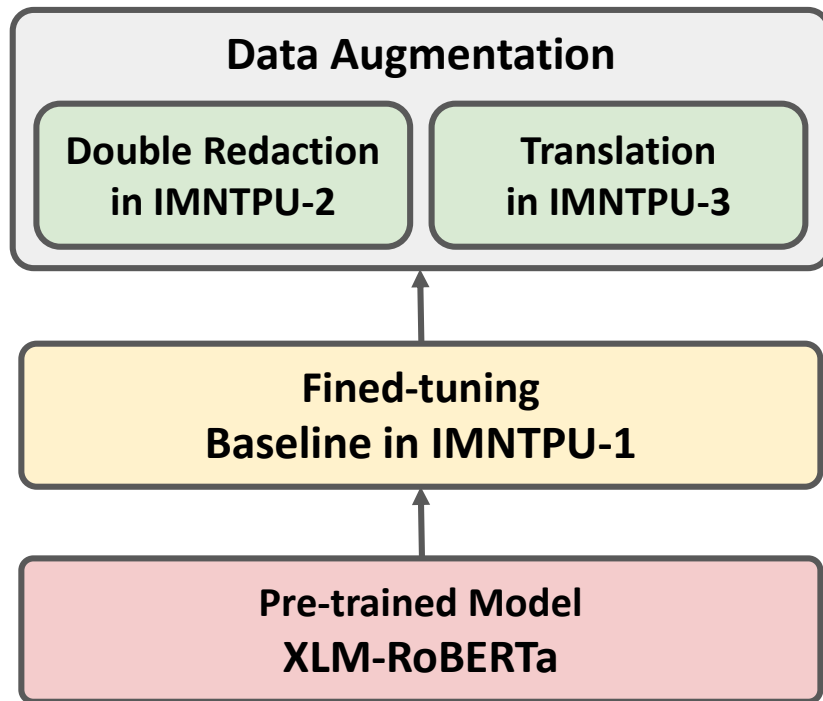
(Information Management at National Taipei University)

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

- IMNTPU Submitted **Three runs** for NTCIR-16 FinNUM3

- IMNTPU1- XLMRoBERTa Baseline Model
- IMNTPU2- Double Redaction
- IMNTPU3- Translation

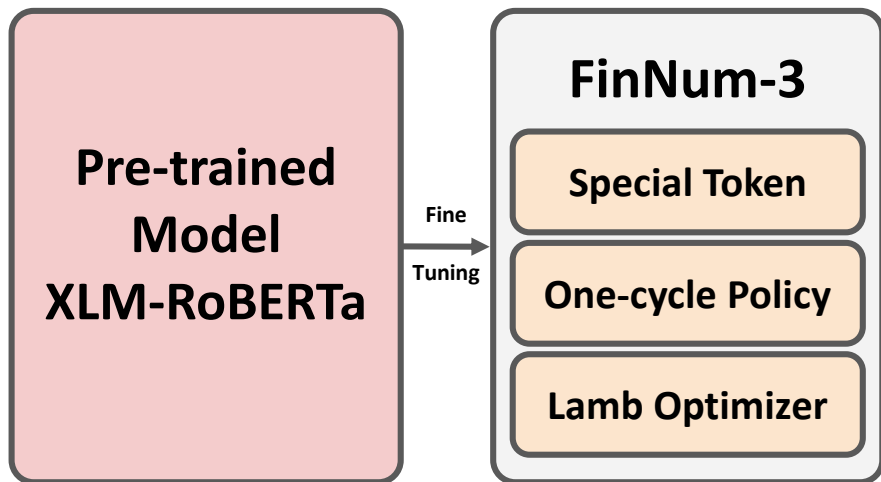
IMNTPU Research Architecture for **NTCIR-16 FinNum-3**



Proposed Method

- IMNTPU1 - We adopted **XLM-RoBERTa Model** without data augmentation as our baseline model.
- IMNTPU2 - We adopt **Double Redaction approach** for data augmentation in XLM-RoBERTa Model.
- IMNTPU3 - We adopt **Translation approach** for data augmentation in XLM-RoBERTa Model.

Fine-tuning of XLM-RoBERTa for IMNTPU at FinNum-3



- **Combine :**
cross-lingual language model (XLM)
- **Tokenizer :** add Special Token
- **Optimizer :** Lamb Optimizer
- **Learning Rate :** One-Cycle Policy

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>`

IMNTPU2- Double Redaction

Algorithm 1 An algorithm of double redaction

```
1: Shuffle the tokens in sentence
2: Delete the duplicated tokens in sentence
3: Copy the remaining tokens as  $\beta$ 
4: SET the  $\delta$  and  $\gamma$ 
5: for specific token in  $\beta$  do
6:   if  $\gamma$  less than  $\delta$  then
7:     Replace original token with <usk> 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 <usk> and <mask>
14: end while
```

Double Redaction- English

Input:

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



Double Redaction for Data Augmentation

Output:

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

Double Redaction- Chinese

Input:

巨大為全球最大自行車製造商，擁有捷安特、Liv、Momentum三個自有品牌，營收比重 70%；代工業務佔 30%，最大客戶為 TREK。主要競爭優勢在生產規模龐大，創造了成本優勢，也使其生產工藝不斷精進。品牌經營則有多面向且細膩的操作經驗，2000年和品牌顧問公司 Interbrand 合作，希望用新品牌精神：啟動探索的熱情 (InspiringAdventure) 連結消費者，開始各項運動行銷操作。



Double Redaction for Data Augmentation

Output:

<mask> 巨大為全球最大自行車製造商，擁有捷安特、Liv、Momentum 三個自有品牌，營收比重 70%；代工業務佔 xxnum 30%，最大客戶為 TREK。主要競爭優勢在生產規模龐大，創造了成本優勢，也使其生產工藝不斷精進。品牌經營則有多面向且細膩的操作經驗，2000 年和品牌顧問公司 Interbrand 合作，希望用新品牌精神：啟動探索的熱情 (InspiringAdventure) 連結消費者，開始各項運動行銷操作。 </s>

IMNTPU3- Translation

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

Run	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

Performance- English

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

Conclusions

- IMNTPU Submitted Three runs for NTCIR-16 FinNUM3
 - IMNTPU1- XLM-RoBERTa Baseline Model
 - IMNTPU2- Double Redaction
 - IMNTPU3- Transaltion
- 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.

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

¹Institute of Information Management, National Taipei University, New Taipei City, Taiwan

²Zeals Co., Ltd. Tokyo, Japan



Yung-Wei Teng¹



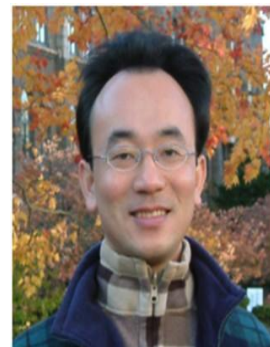
Pei-Tz Chiu¹



Ting-Yun Hsiao¹



Mike Tian-Jian Jiang²



Min-Yuh Day^{1, *}

* myday@gm.ntpu.edu.tw