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 Day1,*

* myday@gm.ntpu.edu.tw

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

NTCIR-16 Conference, June 14-17, 2022, Tokyo, Japan
IMNTPU Research Architecture for NTCIR-16 FinNum-3

Data Augmentation
- Double Redaction in IMNTPU-2
- Translation in IMNTPU-3

Fined-tuning Baseline in IMNTPU-1

Pre-trained Model
- XLM-RoBERTa

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

Pre-trained Model XLM-RoBERTa

FinNum-3

- Special Token
- One-cycle Policy
- Lamb Optimizer

Fine-tuning

NTCIR-16 Conference, June 14-17, 2022, Tokyo, Japan
Good day and welcome to the Apple Inc. Third Quarter Fiscal Year 2018 Earnings Conference Call. Today's call is being recorded.
**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。

Output:
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>
“税后净利润为9.81亿美元，YoY+36.36%，税后EPS2.62元，优于预期。”

“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%，税后EPS2.62元，优于预期。”
# Performance - Chinese

<table>
<thead>
<tr>
<th>Run</th>
<th>Dev Set F1-Score(%)</th>
<th>Test Set F1-Score(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMNTPU-1 (Baseline)</td>
<td>90.51</td>
<td>93.18</td>
</tr>
<tr>
<td>IMNTPU-2 (Double Redaction)</td>
<td>88.65</td>
<td>91.64</td>
</tr>
<tr>
<td>IMNTPU-3 (Translation)</td>
<td>92.16</td>
<td>91.64</td>
</tr>
</tbody>
</table>

NTCIR-16 Conference, June 14-17, 2022, Tokyo, Japan
# Performance - English

<table>
<thead>
<tr>
<th>Run</th>
<th>Dev Set F1-Score(%)</th>
<th>Test Set F1-Score(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMNTPU-1(Baseline)</td>
<td>87.13</td>
<td>88.39</td>
</tr>
<tr>
<td>IMNTPU-2(Double Redaction)</td>
<td>88.82</td>
<td>89.86</td>
</tr>
</tbody>
</table>

NTCIR-16 Conference, June 14-17, 2022, Tokyo, Japan
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.
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 Day1,*

* myday@gm.ntpu.edu.tw

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