CYUT at the NTCIR-16 FinNum-3 Task: Data Resampling and Data Augmentation by Generation

Xie-Sheng Hong, Jia-Jun Lee, Shih-Hung Wu
Chaoyang University of Technology
Taichung, Taiwan

Mike Tian-Jian Jiang
Zeals Co, Ltd
Tokyo, Japan

ABSTRACT

We submitted 3 runs in both two subtask in shared task.
Attempting to solve the problem through data augmentation by ***data resampling and data generation.***
Also, we did additional runs to test the validity of our original proposed methods by conducting more oriented attempts.

Official Runs and Additional Runs

CYUT-1: MacBERT / RoBERTa with BiLSTM
⇒ Baseline for our all systems
⇒ Data resampling
CYUT-2: MacBERT / RoBERTa with AWD-LSTM
⇒ Replace BiLSTM with AWD-LSTM
⇒ Data resampling
CYUT-3: MacBERT / RoBERTa with Additional data
⇒ GPT-2 generates additional data

- ➢ Morel Data: To generate more additional data
- ➢ More Seeds: Using more text types to generate
- ➢ 1000 Data: Using only the first 1000 additional data
- ➢ No Change: Only use MacBERT / RoBERTa with BiLSTM

Result

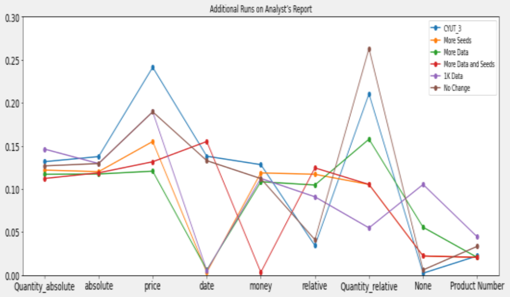
Analyst’s Report			
Run	Macro-F1	Micro-F1	Recall
CYUT-2	86.76%	91.73%	90.32%
CYUT-3	88.20%	92.16%	88.76%
CYUT-1	88.80%	92.11%	87.34%
No Change	88.75%	92.52%	89.30%
More Data and Seeds	89.23%	92.86%	89.92%
More Seeds	89.30%	93.14%	**91.66%**
1000 Data	89.97%	93.16%	89.52%
More Data	**90.24%**	**93.43%**	90.31%

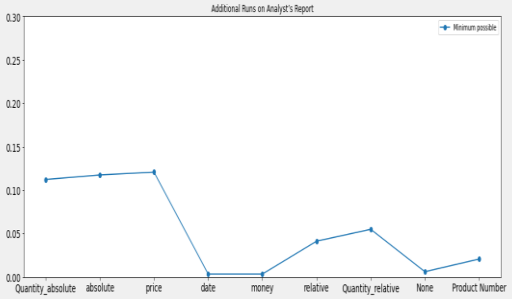
Earnings Conference Call			
Run	Macro-F1	Micro-F1	Recall
CYUT-1	85.53%	94.67%	79.82%
More Seeds	85.93%	95.00%	80.74%
More Data	86.73%	95.93%	84.78%
More Data and Seeds	87.17%	95.76%	83.25%
1000 Data	87.28%	95.73%	83.15%
CYUT-2	87.49%	95.64%	82.39%
CYUT-3	87.88%	**96.43%**	**87.25%**
No Change	**88.15%**	96.22%	85.03%

Discussion and Conclusions

1. The quality of additional texts may be more important than the amount
2. Pay attention to possible overfitting
3. There are advantages and disadvantages among systems in each category

⇒ *Build a large multi-model system that leverages the strengths of each system*
⇒ *A system with a low theoretical error rate*





GPT-2 Data Generation

A fixed text + A random number input GPT-2 => Generate subsequent text

Analyst‘s Report (Chinese data):
Input: ***我們推測會上升(We predict an increase of X)***
Output(Max length: 100): *我們推測會上升*X%，明天早晨大跌...
(We predict an increase of X% and a big fall tomorrow morning...)
X: meaning a random number (range: 0 - 1000)

Earnings Conference call (English data):
Input: ***We anticipate a X increase***
Output(Max length: 50): *We anticipate a X increase* in the number of cases with...