TMUNLP at the NTCIR-15 FinNum-2

Yu-Chi Liang, Yi-Hsuan Huang, Yu-Ya Cheng, Yung-Chun Chang*
Graduate Institute of Data Science, Taipei Medical University,Taiwan
{m946108001, m946108002, i906108009, changyc} @tmu.edu.tw

Introduction

To understand the details of financial data, we must rely not only on analysis of the content, but also the numbers that may contain important information. This task mainly explores whether there is a correlation between numbers and cashtags on financial social media data. Taking Figure 1 as an example, a financial document may contain more than one number. The number '3' represents the cashtags $BAC was at 3 a share. Therefore, the number '3' is describing the cash label $BAC. In contrast, the number '2009' is not the modifier of $BAC due to the fact that it denotes the year was in 2009 and has nothing to do with the cash label $BAC.

Not related (0) directly related (1)

Remember 2009? $BAC was at 3 a share. The people of Greece all want $NBR to open to get back to normal. Tired of banking notes on Notepads

Figure 1: The relationship between the number and the cashtags in a tweet.

Methodology

We have two preprocessing steps before putting the data into the models. First, we replace the cashtag, target number and URLs in the tweet with "TICKER", "NUM" and "URL", respectively. Next, we remove the emoji in the tweet.After preprocessing, we input the normalized data to the model. There are 3 major rounds in our model. Figure 2 shows detailed architecture of our model.

Figure 2: Model architecture

Results

There are total of 10,340 sentences in the data set. Within it, 70% is set as the training set, 10% as the development set, and 20% as the testing set. We use the macro F1 score to evaluate the experimental results. Table 1 shows the results of two baseline methods (Majority and Caps-m) and the results of our models.

Table 1: Experimental results

Run	Macro-F1 score (%)	
	Development	Test
Majority	44.88	44.93
TMUNLP-3	**87.34**	58.40
TMUNLP -2	85.17	59.77
Caps-m	79.27	63.37
TMUNLP -1	87.02	**64.74**

Conclusion

In this paper, we present a system that can distinguish whether there is a correlation between the numbers and cashtags on financial social media data. We show that the BERT embedding matrix as input into BiLSTM with Attention has the best performance, achieving about 64% in F1 score.

Acknowledgments

This research was supported by the Ministry of Science and Technology of Taiwan under grant MOST 107-2410-H-038-017-MY3, MOST 109-2410-H-038 -012 -MY2, and MOST 107-2634-F-001-005.