Quack at the NTCIR-17 FinArg-1 Task : Boosting and MLM Enhanced Financial Knowledge Sequence Classification

Zih An Lin, Hsiao Min Li, Adam Lin, Yun Ching Kao, Chia Shen Hsu, Yao Chung Fan
Department of Computer Science and Engineering, National Chung Hsing University, Taiwan
hua10155174@gmail.com

Abstract

In the exploration of the task “Identifying Attack and Support Argumentative Relations in Social Media Discussion Threads” (SMDT), we aim to discern differences between proponents and adversaries in financial discussions on the internet. Figure 1 describes three examples from the dataset, illustrating the supportive and opposing relationships between different comments. In this study, we build upon fine-tuning Transformer models like BERT by incorporating the Masked Language Model technique to enrich the model's domain knowledge within the financial field. Furthermore, we optimize the model's performance by adjusting the weights in the loss function. Experimental results confirm that both methods effectively enhance the model’s performance.

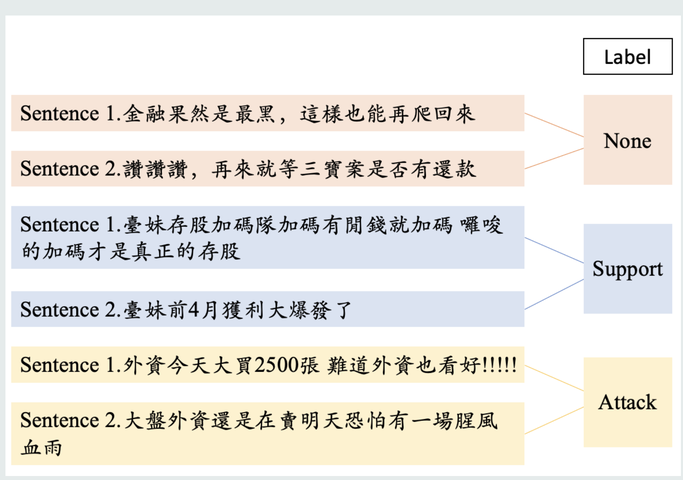

Figure 1: Instances of Support and Opposition in the Dataset

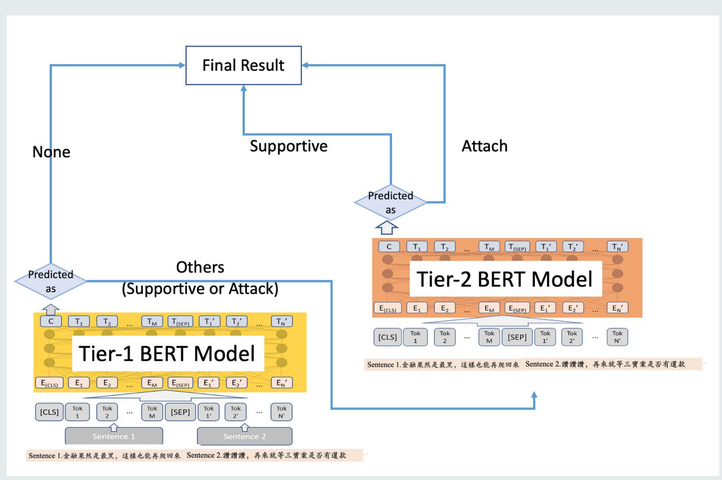

Figure 2: Process of Two-tiered Sequence Classification

Figure 3: Process of MLM prediction

Method

Two-tiered Bert Classification (SMDT_1)

Our approach involves hierarchical prediction using two binary classification models. Refer to Figure 2 or a visual representation. Initially, a BERT model processes the input sentence pair to determine their entailment relationship—whether they are related or unrelated. If the sentences are considered unrelated, we label the output as "none." If they are deemed related, the second-tier model comes into play. This model classifies the sentences as either supporting or attacking.

Mask LM (SMDT_2)

We explore the idea of using Masked Language Modeling (MLM) mechanism to enhance the domain knowledge of the BERT model in the financial domain. We use data from the well-known "STOCK" board in Taiwan's "ptt" forum as the basis, which exhibits similarities with the dataset provided by NTCIR-17. We acquire approximately 32,000 forum articles and conduct MLM fine-tuning training. This process helps the BERT model become more familiar with domain-specific vocabulary in the financial realm (such as stock codes, company aliases, forum terminology). By reinforcing its understanding of financial-related content, we anticipate that BERT will demonstrate even more remarkable capabilities in the financial domain.

Boosting (SMDT_3)

We propose an Adaboost-inspired approach to enhance our classification model and improve overall performance scores. A departure of our method from the original Adaboost lies in the interpretation of weights, where we replace the focus on the proportion of different samples with adjusting proportions of loss. We employ this method to differentiate between "attack" and "support". The working steps are as follows.

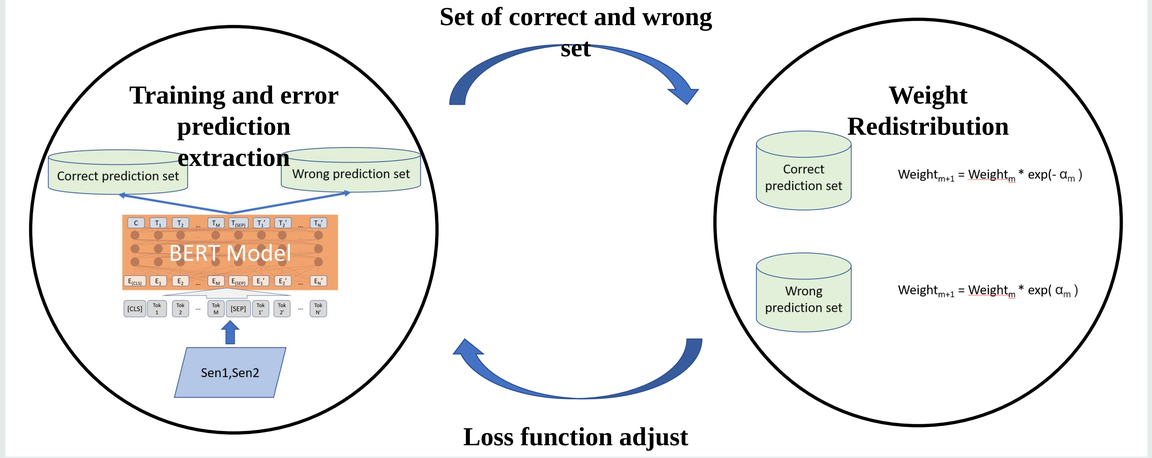

Figure 3 : Adjustment of loss function for Boosting

	pre	rec	f1	acc
Baseline	0.73	0.71	0.72	0.68
SMDT_1	0.71	0.70	0.70	0.68
SMDT_2	0.76	0.72	0.74	0.72
SMDT_3	0.69	0.66	0.67	0.66

Table 1: Result of precision, recall, f1, accuracy on $Test_{official}$ in different way.

Performance Evaluation

In the final stage, we evaluate the SMDT_1, SMDT_2, and SMDT_3 models using the Test official dataset. The results of these evaluations are also displayed in Table 1

A notable enhancement was evident on the testing data provided by NTCIR-17, as reflected in the score statistics shown in Table 1. Specifically, our precision, recall, F1 scores, and accuracy all showed relative improvements compared to the initial SMDT_1 model, with increases of 0.05, 0.02, 0.04, and 0.04, respectively. This illustrates that the efficacy of MLM might vary across different testing data scenarios, and it performs admirably on the Test official dataset.