WUST at the NTCIR-14 FinNum Task

Wei Wang, Maofu Liu, Zhenlian Zhang
School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
liumaofu@wust.edu.cn

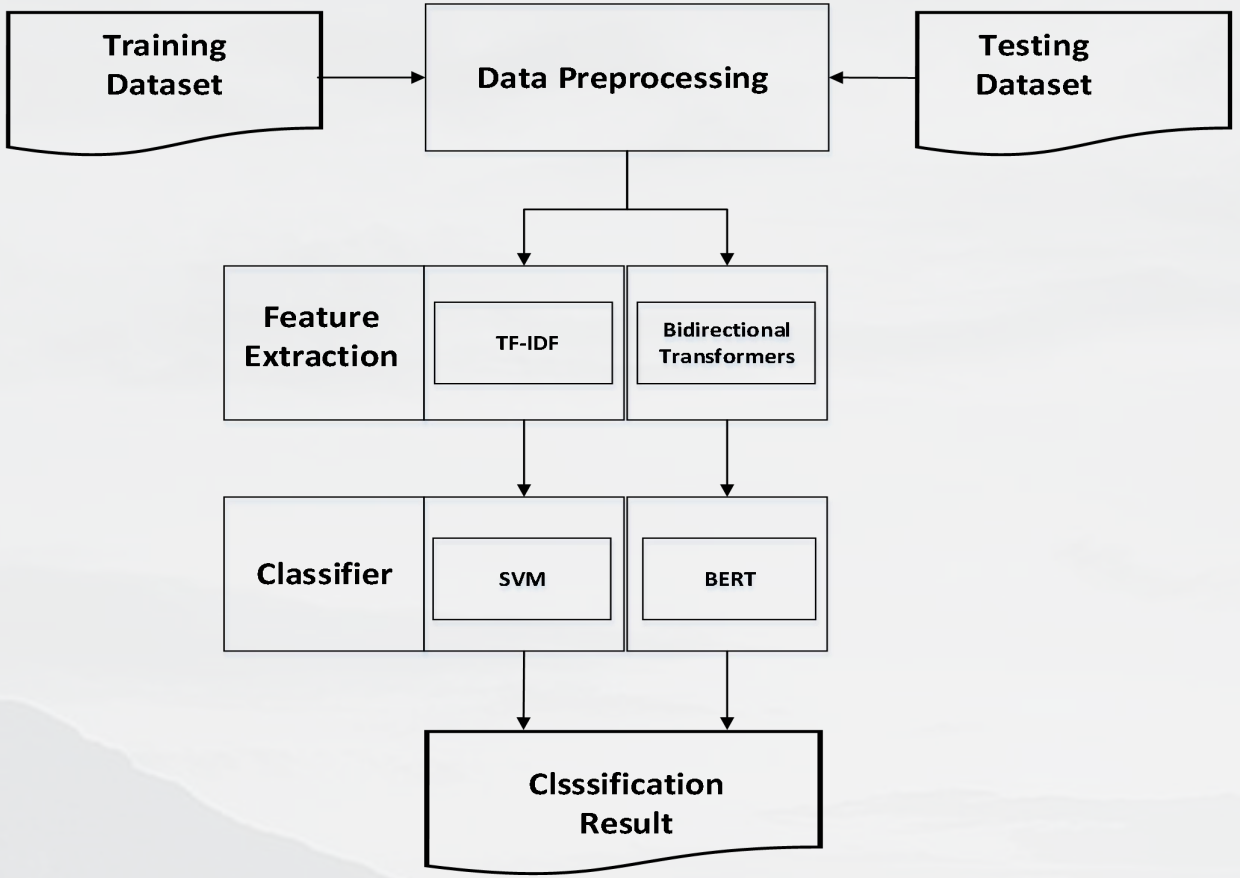
CONTENTS

01/Introduction

- In NTCIR-14, the FinNum task is dedicated to identifying the category of a numeral from the financial social media data, i.e. tweet, for fine-grained numeral understanding.
- Our team carry out the FinNum task as the text classification problem. The Subtask1 is one seven-classification task, and the subtask2 is one seventeen-classification task.
- In our system, we construct the classification model based on Support Vector Machines (SVM) to identify the categories of numerals. In additional experiments, we adopt the Bidirectional Encoder Representations from Transformers (BERT) model to act as a multi-classifier.
- The experimental results show that our proposed both SVM and BERT models are effective.

02/System Architecture

- Our system consists of three main modules, i.e. data preprocessing, feature extraction, and classifier.



Data preprocessing

- In the data preprocessing, due to the uneven data distribution, our team tries to expand the training data, ensuring that one target numeral corresponds to one tweet and one category, divide tweets with the target numeral, and combine the segmented data into the training set. As a result, the divided data are tripled.

Example :

Original data:

Target num: 22.5, 20

Tweet: $BZUN I sold 22.5 and 20 puts for income. Don't think I'll be assigned.

Expanded data:

Target num: 22.5

Tweet: $BZUN I sold 22.5 and 20 puts for income. Don't think I'll be assigned.

Target num: 20

Tweet: $BZUN I sold 22.5 and 20 puts for income. Don't think I'll be assigned.

Feature extraction

- Before classifying with SVM, we extract textual features by Term Frequency-Inverse Documentation Frequency (TF-IDF), and TF-IDF is a statistical method used to evaluate the importance of a word to a document set or one of the documents in the corpus. The TF-IDF algorithm is simple in principle, fast in the calculation, and has strong universality for extracting keywords for FinNum task.

- In additional experiments, we adopt the BERT model, which trains a universal language understanding model on a large-scale corpus, a deep bidirectional Transformer encoder is used to retain contextual features and positional information when extracting textual features.

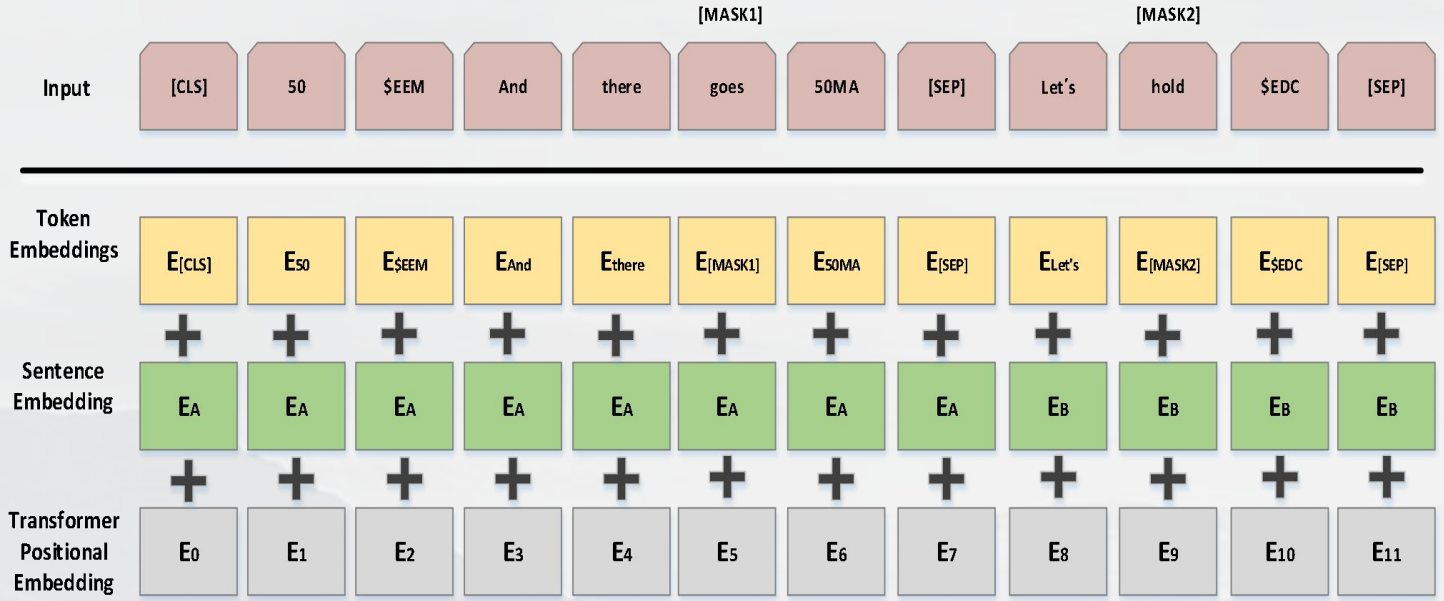
SVM Classifier

- The SVM model has the characteristic of high precision and multidimensional data processing, and strong robustness, so the SVM model is favored in the study of classification problems. The SVM is usually used for binary classification problems, but the FinNum task is a multi-classification problem, so a multi-classification SVM needs to be constructed.
- The specific formula is shown as follows.

$$k(x, x_i = \exp(-\gamma \left\| x_i - x_j \right\|^2)), \lambda > 0$$

BERT Classifier

- The BERT uses a simple approach, in which a special classification embedding ([CLS]) inserted as the first token and a special token ([SEP]) added as the final token. It masks out 15% of the words in the input, runs the entire sequence through a deep bidirectional Transformer encoder, and then predict only the masked words. the Sentence Embedding as shown below.



03/Experiments

- We submitted one system result to NTCIR-14 for the FinNum task. The official evaluation results are listed in Table 1.

Table 1. Performances of the CRF model

Task Name	Micro	Macro
Subtask1	0.7402	0.6371
Subtask2	0.6088	0.5293

- In additional experiments, we selected the BERT model as a classifier. The experimental results are shown in Table 2.

Table 2. Performances of the BERT model

Task Name	Micro	Macro
Subtask1	0.9450	0.8862
Subtask2	0.8725	0.8307

04/Conclusions

- we separately employ SVM and BERT models for classification, both of which have got good performances. Moreover, we have mainly taken statistical features into consideration, and we will extract and select more rules and semantic features for our system to improve the system’s accuracy.

THANK YOU