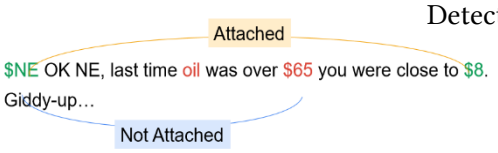
WUST at NTCIR-15 FinNum-2 Task

Xinxin Xia, Wei Wang, Maofu Liu
School of Computer Science and Technology, Wuhan University of Science and Technology

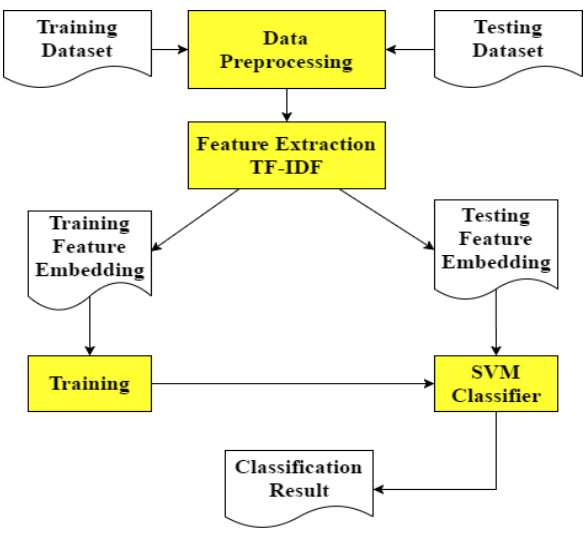
01. Introduction

Numbers are a key part of financial documents.
In order to understand the details of the comments in the financial documents, in-depth analysis of the digital information is also required.



Detect the attached target of the numeral
Target = Cashtag
Source: Financial Tweet

02. System Architecture



1 Data Preprocessing

- removed the emoji codes contained in the text
- cleaned the URLs contained in the text data of the training set.

2 Feature Extraction

- Spliced tweet content with cashtag and number
- used TF-IDF to transform the text into a machine-friendly representation.

3 Classifier

- SVM is an optimal boundary classification method based on VC dimension theory and structural risk minimization criterion

03. Experiments

Table 1. Experimental results

Team	Development	Test
Majority	44.88	44.93
CYUT-1	48.64	48.02
WUST	**82.91**	**54.43**
Caps-m[2]	79.27	63.37
CYUT-2	95.99	71.90
TLR-3	88.87	73.95

In additional experiment, we collected 1360 from the attached data, and then merged with 1360 unattached data. After shuffling, we retrained the model by under-sampling the training set.

Table 2. Additional experiment results

Team	Development	Test
WUST	82.56	64.91

04. Conclusions

We use the SVM model to classify text by concatenating text features. In additional experiment, we retrain the model by down-sampling the training set, and the experimental results show that this is effective.

For classification model, due to the small amount of data, we will think of pre-training model. Such as BERT, XLNET. Both of them can notice location information and make use of contextual information. After we complete the above, we will be able to achieve better results.