WUST at NTCIR-15 FinNum-2 Task

Content

Introduction

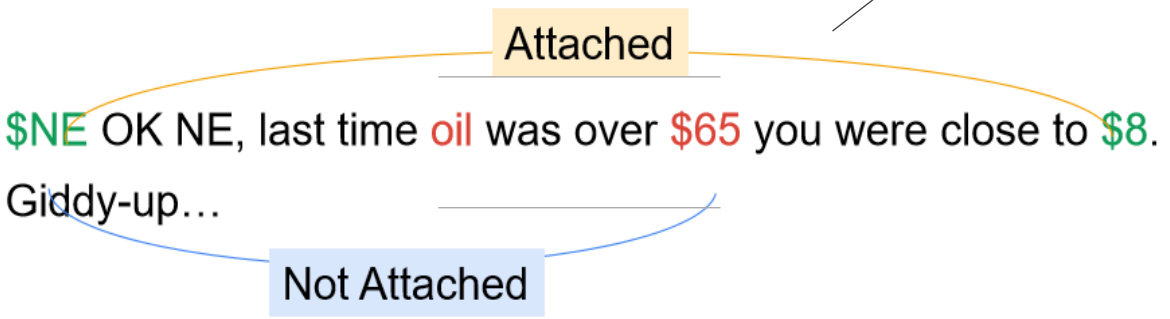
1.1 Introduction

In the era of big data, the continuous improvement of computer technology makes the data type more abundant, the text big data has become the data that the computer can interpret and analyze, and can study the economic phenomena in non-traditional fields.

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.

1.2 Task Definition

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



1.3 Data Description

{

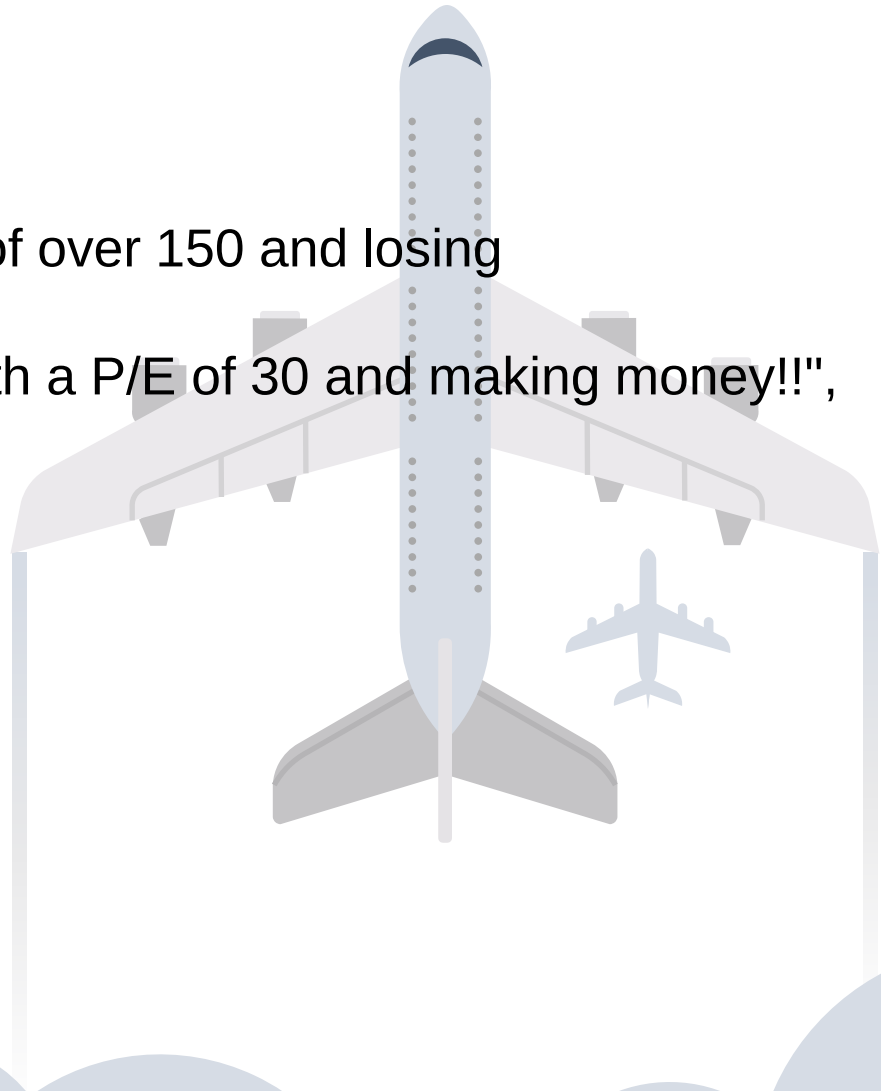
"tweet": "$SQ is $39 per share with a P/E of over 150 and losing money...should be at least as good as them with a P/E of 30 and making money!!",

"target_num": "39",

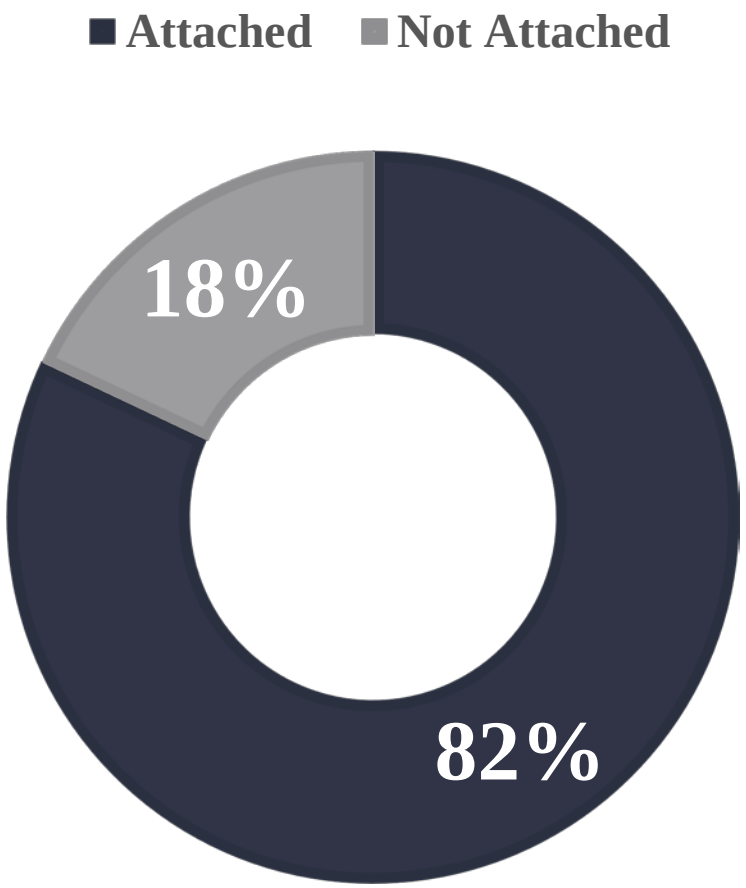
"target_cashtag": "SQ",

"relation": 1

}

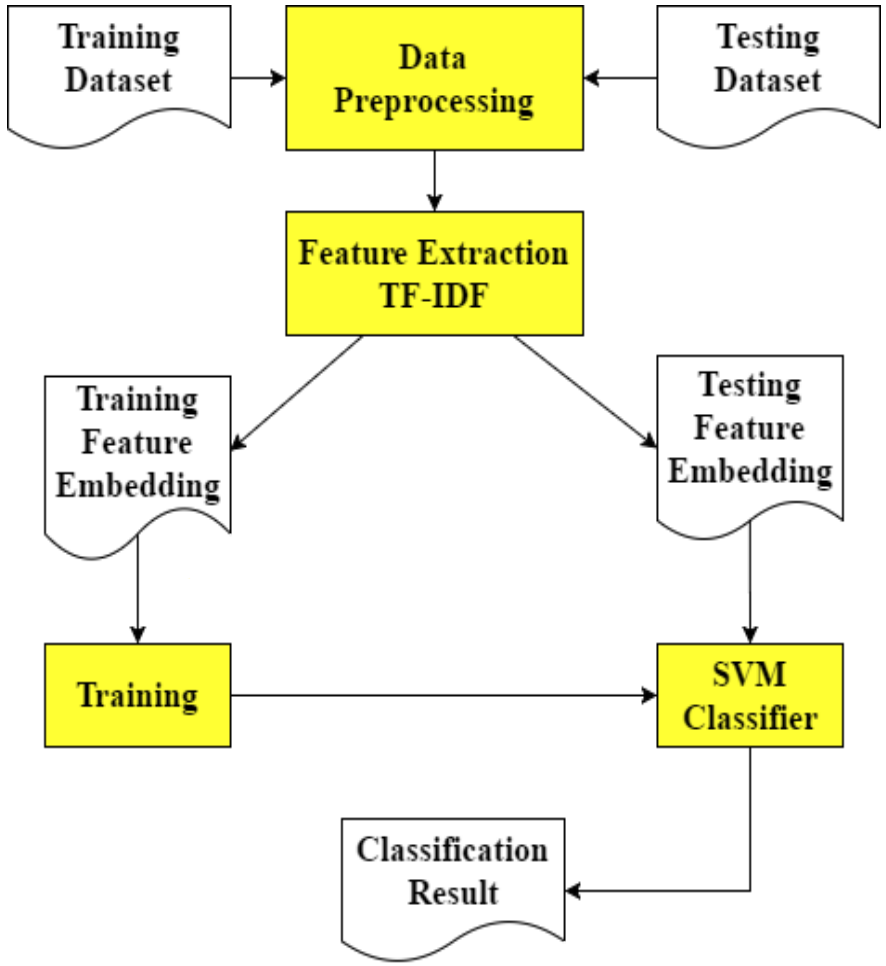


1.4 Data Distribution



System Architecture

2 System Architecture



01

Data Preprocessing

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

02

Feature Extraction

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

03

Classifier

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

Experiments

3.1 Additional Experiment

Our team tried to expand the data that is not relevant. Due to the uncertainty of the label criteria, only copy unattached data multiple times and disrupt it.

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.

3.2 Experiment results

The official evaluation results are listed below

Table 1. Experimental results

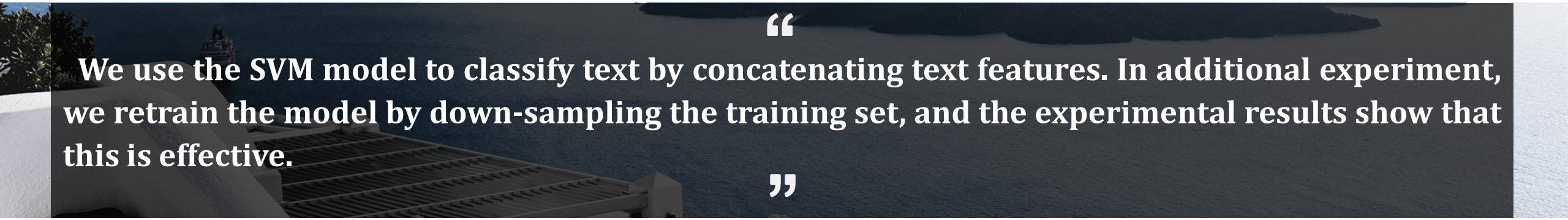
Team	Development	Test
Majority	44.88	44.93
CYUT-1	48.64	48.02
WUST	**82.91**	**54.43**
Caps-m[2]	79.27	63.37
CYUT-2	95.99	71.90
TLR-3	88.87	73.95

Table 2. Additional experiment results

Team	Development	Test
WUST	82.56	64.91

Conclusions

4 Conclusions



We will consider dealing with the problem of data imbalance and adopt reasonable data enhancement methods to improve the generalization ability of the model.

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.

Thanks