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,
Wuhan 430065, China
liumaofu@wust.edu.cn

ABSTRACT

This article introduces how we deal with the FinNum-2 task of NTCIR15. In the FinNum-2 task, the relationship between a number and a given label is the object of classification. In a short text, given a target value and a cashtag, judge whether the given target value is related to the given cashtag according to the content. The classification involved in this topic is essentially a two-classifier, that is, given a short text, determine whether the given value and the label are relevant or not. We use the SVM model to classify the text by splicing text features and analyze the results.

KEYWORDS

Financial Numeral Classification; Numeral Understanding; Financial Social Media; Natural Language Understanding

TEAM NAME

WUST

SUBTASK

Numeral Attachment in Financial Tweets

1 INTRODUCTION

WUST team participated in the NTCIR-15 FinNum-2 task. This report introduces the methods we used in this task and discusses the experimental results.

Digital-related information in financial social media data is the focus. But before understanding digital information, we should first determine the individual corresponding to the number in order to better understand its true meaning.

In FinNum-2, the organizer introduced a new task called Number Attachment to identify the relation between the mentioned stock and the numerals in a financial tweet. Our team regards this task as a text binary classification problem. The system we proposed first divides the training set into two parts for training and verification. Then use the trained model to classify the test set. Due to the small-scale data set and the purpose of binary classification, we chose the SVM model to complete fast classification.

The rest of this report is organized as follows. Section 2 shows the related work of number classification. Section 3 introduces data preprocessing, text feature extraction and classification models. Section 4 shows the official experimental results and some discussion about error cases. Finally, some conclusions are drawn in Section 5.

2 REALATED WORK

In the field of financial research, the study of online media text information has become more and more common. The earliest research on this mainly relied on empirical studies of special circumstances or linear regression models, which simplified the influence of the media on news articles, rather than the influence of their text content [1,2]. With the technological advancement of natural language processing and artificial intelligence, more and more studies are exploring these connections from the perspective of big data. The useful information extracted from Web media is expressed in a machine-friendly form, and various types of analysis models are used to fit the information relationship.

There are three mainstream methods, statistical models in statistics, regression models in econometrics, and machine learning-based models in computer science [3].

Machine learning (ML) models will take high-dimensional data as input, usually connect the features of different information sources into one feature vector, which will be applied to machine learning to explore the relationship between the information. Such as neural networks [4], Bayesian classifiers [5]and support vector machines [6].

Support vector machines has been proven to achieve better classification results in small datasets. Considering the small amount of data in this experiment, we decided to use SVM as the classifier.

3 SYSTEM DESCRIPTION

Based on the analysis of the purpose of the task and the corpus, the task can be regarded as a binary classification problem, which is mainly to judge whether a given number is related to a given cashtag based on Twitter.

Our system includes three main modules, namely data preprocessing, feature extraction and classifier.

Figure 1 details our system architecture.

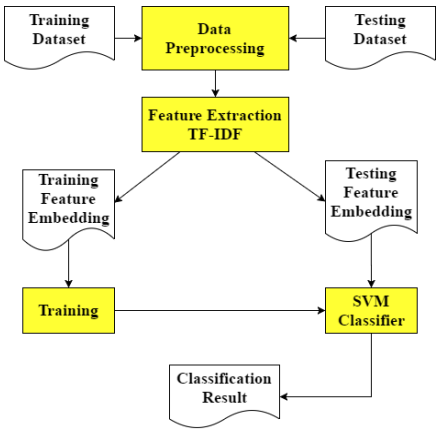

Figure 1:System Overview

3.1 Data Preprocessing

In the data preprocessing, we calculated 7615 training data according to the relation. Figure.2 shows the distribution of

relations in the data.

From the data distribution, the two types of data distribution are not balanced, and the Attachment accounts for 82% of the part. Considering that data distribution will have a biased effect on training, 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.

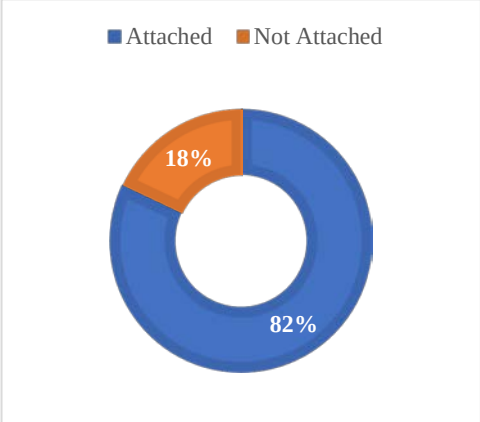

Figure 2:Category Distribution

After observing the data, we removed the emoji codes contained in the text and cleaned the URLs contained in the text data of the training set.

Example 1:

Origin

Target num: 1.56
Target cashtag: SEED
Tweet: I am bullish on $SEED with a target price of $1.56 in 3 mos. on Vetr! http://on.vetr.com/2A3o9MK
relation: 1

After preprocessing

Target num: 1.56
Target cashtag: SEED
Tweet: I am bullish on $SEED with a target price of $1.56 in 3 mos. on Vetr!
relation: 1

Example 2:

Origin

Target num:4
Target cashtag: DPW
Tweet: $DPW newbies I know y\u2019all excited to see 4. But let this one ride out. \ud83d\ude02\ud83d\ude02 lets try for 5
relation: 1

After preprocessing

Target num:4
Target cashtag: DPW
Tweet: $DPW newbies I know yall excited to see 4. But let this one ride out. lets try for 5
relation: 1

3.2 Text Feature

The main text of the task is the tweet content. We spliced it with the objects that need to be classified, cashtag and number, and used TF-IDF to transform the text into a machine-friendly representation.

TF-IDF is an algorithm that can convert text into an expression that can be operated on by a computer. Among them, TF is the abbreviation of term frequency, which refers to the number of times a given word appears in the file.

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (1)$$

Inverse document frequency (IDF) reflects the frequency of a word in all texts. If a word appears in many texts, its IDF value should be low.

$$IDF_i = log \frac{|D|}{1+|\{j : t_i \in d_j\}|} \quad (2)$$

TF-IDF algorithm is to multiply TF and IDF.

$$TF - IDF_i = TF_i * IDF_i \quad (3)$$

3.3 SVM Classification

SVM is an optimal boundary classification method based on VC dimension theory and structural risk minimization criterion [7,8]. The schematic diagram of the optimal separating hyperplane is shown in the Figure 3. Its advantage lies in solving small sample, nonlinear and high-dimensional regression and binary classification problems [9].

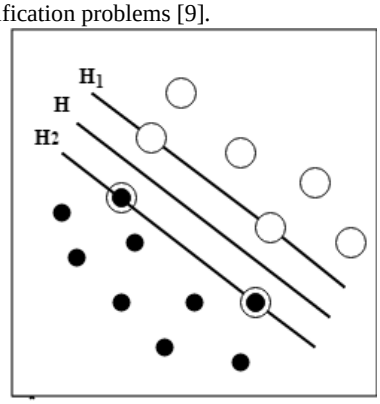

Figure 3: Optimal Separating Hyperplane

For nonlinear problems, SVM non-linearly maps the training text to a high-dimensional space, so that the problem of linear inseparability in low-dimensional space is transformed into linearly separable in high-dimensional space without increasing the number of adjustable parameters. Then, in the high-dimensional feature space obtained by this transformation, the optimal classification surface is constructed using the solution idea in the linearly separable case.

Finally, the constructed optimal classification surface is mapped back to the original space to obtain the decision function of the classifier. In order to avoid the possible dimensionality disaster, the inner product of the transform space can be replaced by introducing an appropriate kernel function. In this experiment, we chose the Gaussian kernel function.

$$K(x_i, x_j) = \exp\left[-\frac{[x_i \cdot x_j]^2}{\sigma^2}\right] \quad (4)$$

σ is the radial basis radius.

4 EXPERIMENTS

4.1 Dataset and Evaluation Metrics

In this experiment, we used the NumAttach 2.0 dataset proposed by NTCIR-15[10], which contains 7187 training sets, 1044 development sets, and 2109 test sets. According to official

evaluation criteria, we use the macro-F1 score to evaluate the experimental results.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 - measure = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (7)$$

Among them, the TP refers to the category where cases are truly classified into the positive class, FP refers to the category where cases are wrongly classified into the positive class, FN refers to the category where cases are wrongly classified into the positive class.

Macro-F1 first calculates the accuracy and recall rate of each category and its f1 score, and then obtains the F1 score on the entire sample by averaging.

4.2 Experimental results

We submitted one system result to NTCIR-15 of FinNum-2 task. The official evaluation results of performance are listed in the Table 1.

Table 1 shows that the SVM model has achieved good performance in the development set.

However, there is still much room for improvement. We surmise that the data imbalance may cause this not good results in test set. Therefore, in additional experiment, we collected 1360 from the attached data, and then merged with 1360 unatached data. After shuffling, we retrained the model by under-sampling the training set.

The experimental results are shown in the Table 2.

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

From table2, we find that by down-sampling the unbalanced training set, the results in test set are greatly improved.

4.3 Error Analysis

In this section, based on comparing the correct classification results with the classification results of our model, our team analyzed several examples of classifications and discussed the reasons for the errors. The examples are shown in table 3.

Most of the Attached data has been judged correct, but still a small part of the Attached data is judged incorrectly.

Table 3. Analysis examples

	Target num	cashtag	Tweet
Example 3	2018	ISR	2018 shareholder meeting, Thurs, December 14, 2017, beginning at 11:00 a.m. local time, 8701 East Pinnacle Peak Road, Scottsdale, AZ $ISR
Example 4	10	GDXJ	$IDXG Part 6: $JNUG $NUGT $GDX $GDXJ and this is just a juicy set up for gold next month with monthly above mid B.B. and SMA10
Example 5	14017 77666 06896	WFT	$WFT Watch lists, technical analysis assistance, workspace setups, Robinhood/general market advice, $15/month www.facebook.com/groups/140177766606896
Example 6	8	ENTL	WATCHLIST $GEF $TOPS $AVGR $GLBR $WAC $SAGE $ENTL $SIGM $DPW $DCIX $MRNS $ALXN $GALT $GLBS $OC $GLBR $TROV $CLDR $INFI $LC

Example 3:

The given label “ISR” and the given number “2018” are far apart in the text “2018 shareholder meeting, Thurs, December 14, 2017, beginning at 11:00 a.m. local time, 8701 East Pinnacle Peak Road, Scottsdale, AZ $ISR”. The features extracted based on the word frequency cannot demonstrate location information.

On the contrary, most of the Unattached data is judged to be Attached. The following are two typical examples.

Example 4:

According to the official report, the number of multiple-cashtag with attached numbers in the dataset is far greater than not attached. And this text contains multiple cashtag, which interferes with classification. The lack of this part of the training data makes the model unable to distinguish such situations well.

We also select a few examples of correctly classified Not Attached data. The correct classification has the following two characteristics.

Example 5:

The text “$WFT Watch lists, technical analysis assistance, workspace setups, Robinhood/general market advice, $15/month www.facebook.com/groups/140177766606896” contains obviously long URLs. During the model training process, this part of the training set was observed to be washed out in the preprocessing stage. This result shows that this operation is effective.

Example 6:

Except for the tags, the text “8 WATCHLIST $GEF $TOPS $AVGR $GLBR $WAC $SAGE $ENTL $SIGM $DPW $DCIX

$MRNS $ALXN $GALT $GLBS $OC $GLBR $TROV $CLDR $INFI $LC" contains short content and many cashtag. The text is so short that only a single word can provide more information, which will cause the classifier to classify it into not attached.

5 CONCLUSIONS

This paper details our participation in the Numeral Attachment in Financial Tweets task, which is a subtask of the NTCIR-15. In the FinNum-2 task, the relationship between a number and a given label is the object of classification.

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. Next 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.

REFERENCES

[1] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. Review: Text mining for market prediction: A systematic review. *Expert Syst. Appl.* 41, 16 (November, 2014), 7653–7670. DOI:https://doi.org/10.1016/j.eswa.2014.06.009

[2] Zheludev, I., Smith, R. & Aste, T. When Can Social Media Lead Financial Markets?. Sci Rep 4, 4213 (2014). https://doi.org/10.1038/srep04213

[3] Li, Q., Chen, Y., Wang, J., Chen, Y., & Chen, H. (2018). Web Media and Stock Markets: A Survey and Future Directions from a Big Data Perspective. *IEEE Transactions on Knowledge and Data Engineering*, *30*(2), 381-399. [8068217]. https://doi.org/10.1109/TKDE.2017.2763144

[4] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, "The Graph Neural Network Model," in *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61-80, Jan. 2009, doi: 10.1109/TNN.2008.2005605.

[5] Binder, J., Koller, D., Russell, S. et al. Adaptive Probabilistic Networks with Hidden Variables. Machine Learning 29, 213–244 (1997). https://doi.org/10.1023/A:1007421730016

[6] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (*COLT '92*. Association for Computing Machinery, New York, NY, USA, 144–152. DOI:https://doi.org/10.1145/130385.130401

[7] J. A. K. Suykens and J. Vandewalle. 1999. Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.* 9, 3 (June 1999), 293–300. DOI:https://doi.org/10.1023/A:1018628609742

[8] Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* 2 (3/1/2002), 45–66. DOI:https://doi.org/10.1162/153244302760185243

[9] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (*COLT '92*). Association for Computing Machinery, New York, NY, USA, 144–152. DOI:https://doi.org/10.1145/130385.130401

[10] Chen, C.C., Huang, H. H., Takamura. H., Chen, H.H. 2020.Overview of the NTCIR15 FinNum-2 Task: Numeral Attachment in Financial Tweets.