TEXT OPTIMIZATION ANALYSIS FOR THE FINANCIAL CORPUS

Besufikad Enideg Getnet *
School of Computer and Information Engineering, Beijing Technology and Business University, Beijing, 100048, China

Abstract
The corpus is a useful tool for the linguistics statistical analysis to check occurrences or validate linguistic rules within a specific language territory. The general corpus is very extensive, such as Google N-Grams Corpus and American National Corpus, etc, while they cannot satisfy the specific need of the financial field, in which some especial financial words always didn't be included and the text analysis results can't be good enough for the applications. In this paper, we take the downloaded financial news as the original corpus, and use them as the input of the text classification system. This whole process forms a closed loop to get the optimized corpus. By the simulation for the financial news analysis, we compared the prediction results for the stock tendency between the optimized corpus and the original corpus, the results show the predictions are greatly developed by the optimized corpus.

Keywords: Text Classification; Optimization of the Corpus; Feature Selection Methods; Prediction of Time Series Data.

1. INTRODUCTION

Recently, the financial news has played the more and more important role in the valuation process for the investors and institutional traders evaluating stock prices. In fact, the financial news carries information about the firm's fundamentals and qualitative information influencing expectations of market participants [1].

The financial new analysis can be carried the text classification which has been used to apply in many fields, such as information retrieval, news catalog, digital library, E-mail assortment, searching engines, and text data base, etc [2]. Nowadays, text classification for the financial news analysis becomes more and more important because the analysis result can provide effective information for stock prices forecasting [3]. For example, in stock prediction, today's financial news are input into the text classification system, the result would be given to show the tendency of the tomorrow's stock prediction.

Corpus is the key part in the text classification system. The results have shown that the good corpus is very necessary for the good prediction results. But the general corpus can't meet the demand of the financial news analysis.

Researches has given some achievements in text classification such as the feature selection, the size of the sample space, etc [4,5]. But it is less mentioned to optimize the corpus while use it. After the corpus optimized, the financial professional words become the main words in the corpus, which will improve the performance of text classification and reduce the cost of time when the corpus is use to the stock analysis. While in the general corpus, the effect words are less. For instance, in Google N-Grams Corpus, there are about 155 billion words in all, but only thousands of words are financial related. Clearly, corpus optimization can make the corpus more specialized and convenient to use, and we assure it can be used more and more widely in text classification [6,7].

This paper gives a method to get the optimized corpus in order to achieve an advanced financial news analysis. The rest of paper was organized as follows. Corpus optimization method and the steps of text classification are described in Section 2. The corpus optimization method, are described in Section 3. The experimental results are analyzed in Section 4. Conclusions are given in Section 5.

2. CORPUS OPTIMIZATION METHOD

Usually, the process of text classification can be divide into two modules. The first is the learning module which can also divide into two parts, such as training process and testing process. The training process can create the classifier, which is used by the testing process. Specially, in this module we develop the feedback structure to optimize the corpus. The second module is the classify module in which the classifier is used to classify new text documents and output the results of new documents are output. We make the following improvements by using optimized corpus to train new model. First, we train all of the texts, and test them. So these texts are both training sets and test sets. Then, we rule out all the texts which are predicted wrong. After that, we use the rest of the texts to train a new model, which is more targeted by using the optimized corpus [8,9].

1) When we use this new model to test one new text, if the result is right, this text can be added into the corpus, so

we can get a bigger and more comprehensive corpus.

2) We can do the operation of the above recycled using this new corpus, so we can get a feedback structure, which is more targeted.

2.1. Training and Test Process

In training process, we should convert the text into a table that contains columns which are composed by its basic semantic unit firstly. Then the feature of key words should be selected, by which the classification, such as SVM, Neural net, etc.

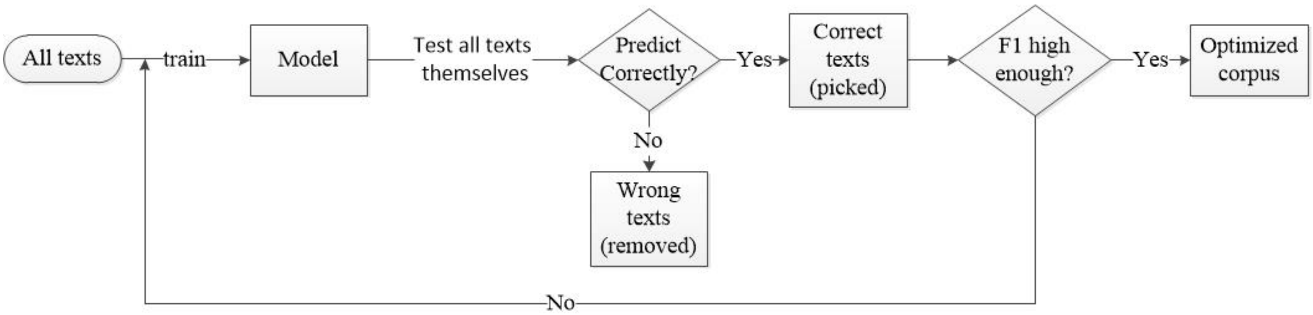

Fig. 1. The Process of Optimizing Corpus

To choose the appropriate basic semantic unit is called as word segmentation. There are two main ways for word segmentation: dictionary-based method and rule-based method. The former is to classify the same meaning words as a class, and to establish specialized thesaurus, and the latter is to strip each word suffix according to certain rules, in order to obtain the root revealed the basic meaning [10].

Stop words are those words that can't reflect the theme. In Chinese, "however", "therefore", "but" this kind of words can only reflect the sentence grammatical structure, "to", "must", "the" this kind of words are just part of auxiliary, which we call as stop words. Stop words contain only small amounts of classified information, so its ability to distinguish the text is very weak. It is necessary to delete these stop words from the text. In the actual application process, the text classification system will usually set up a stop words list to directly filter out the words appeared on the list. Not only does this method is relatively simple, but also can filter out the useless words for classification, don't let them to appear in the final text vector space. The text of text collections now are represented as feature vector form, which still has a large number of words. But the fact is that it is necessary to select as less as words to describe the text.

$\mathrm{TF}-\mathrm{IDF}$ (Term Frequency-Inverse Document Frequency weighting) is known as the most widely used weight calculation methods in text processing. The method is based on the following reasons: first the more times the feature *i* appears in the document *j*, the more important it is; second the more the number of the texts which contains the feature *i*, the least important they are. Now the formula of TF-IDF is commonly given by

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}*\mathrm{IDF} \quad (1)$$

where TF is Term Frequency, and IDF is Inverse Document Frequency.

Before text classification, there is an urgent need to reduce the space dimension of text feature, at the same time to delete the noise in the text feature space. Nowadays, feature selection is a good method to reduce the space dimension of the text feature. Feature selection refers to a process, which select a small portion of characteristics from the original feature set and then form a new feature set. The new feature set is a subset of the original feature set.

At present, there are many methods of feature selection for text classification [11], commonly used methods are: Information Gain (IG), CHI-Squared (CH2), Mutual Information (MI) and Cross Entropy (CE) and so on. In different cases, the effect of these methods is different [12]. We have used these methods in the experiment, and compared their effect in the same situation. The details of the feature selection method are in the appendix.

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [13,14]. As to the text classification, compare to the traditional classification method, using SVM can get better classification results which is generally accepted, so this paper will take SVM as the text classification method for research. After the training part of the system, we can get the training model. Next, we use feedback structure to optimize the corpus. We use this model to test all the texts, and only leave the tests which can get classified correctly. After that, we retrain these left texts, and repeat the above process, until the results meet the requirements. The whole left texts at last form the optimized corpus. Finally, we can use this optimized corpus to predict new text [15].

3. THE CORPUS OPTIMIZATION METHOD

This paper optimized the corpus by the innovative concept of corpus circulatory adding, so that we can get a feedback structure. Fig. 1 shows the details of the feedback structure, and Fig. 2 shows the way to get expanded optimized corpus.

As shown in Fig. 1, we train all the texts and use the model to test themselves. Some of the texts are tested correctly,

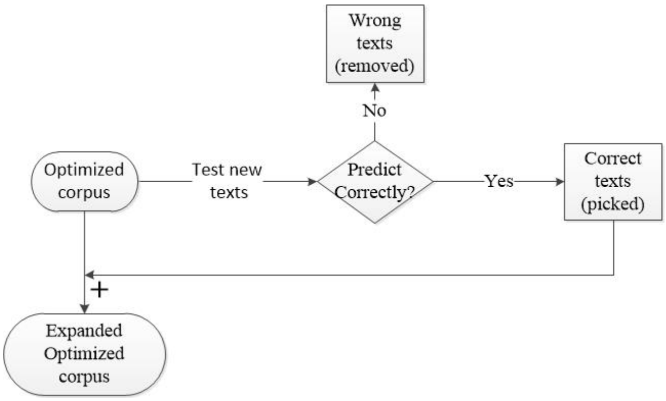

Fig. 2. The Way to Get Expanded Optimized Corpus

which are picked to be the optimized corpus, and some are tested wrongly, which will get removed. The results of the test is measured by F_1, which will be explained in 4.1. The higher the F_1 is, the better the results get. So, if F_1 of this test isn't high enough, we can use the current optimized corpus as "all texts" and run first step again until we get the corpus whose test results reach the required F_1. In conclusion, this feedback structure can make sure us get the optimized and targeted corpus. After that, as shown in Fig. 2, we can use the optimized corpus to test new texts, and add the predicted correctly texts into optimized corpus to get expanded optimized corpus.

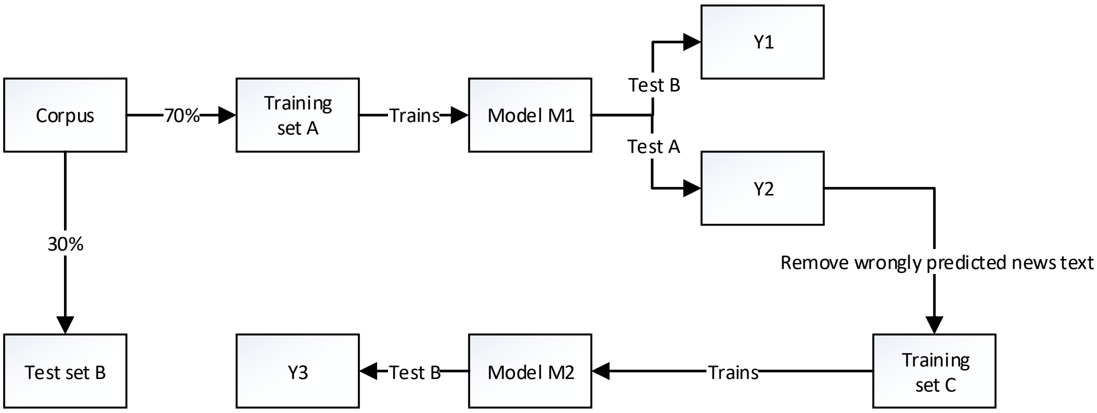

Fig. 3. Corpus Optimizing Flow Chart

As shown in Fig. 3, the specific steps of the feedback structure are as follows: Take the first 70% of collected press as a training set, treating it as original corpus and marking it as A. And use the remaining 30% of news as a test set, marking as B. First, conduct training A, and use the resulting model M1 to test B, recording the results Y1.After that, we still use the model M1 for testing the training set A, retaining exactly predicted news texts as new training set C, according to the test result Y2.At this time the set C is the corpus optimized. Remove wrongly predicted news text. Next conduct training C, use the resulting new model M2 to test B and get the test results Y3.To see if the optimization is effective, we compare Y3 with Y1.After and so on, continue to optimize and update the train set, to generate a more targeted corpus, so as to obtain more accurate test results.

4. EXPERIMENT AND ANALYSIS

We use the finance and economics news of several stock and forex, which are Poly Real Estate, Petro China, Ping An insurance and the exchange rate between euro and dollar. Among them, the Poly Real Estate is a large estate state-owned real estate company. Petro China, a state-owned company, in the leading position of China oil-gas industry, being the largest oil-gas manufacturer and retailer, is also one of the companies that have the largest sales revenue in China and one of the largest petroleum companies in the world. Ping An insurance (Group) Company of China Ltd., is finance company listed in Hong Kong Stock Exchange and Shanghai Stock Exchange, main businesses of which are offering diversified finance services and products and its core is the insurance service.

When collecting news of every stock and the exchange rate between euro and dollar, we first gathered enough news, predicting the share price and the exchange rate in the following day, from websites like Snowball Network and Hexun.com. Then we saved every piece of news in the form of TXT and divide them into two part: rose and fell, allowing the test of Text Classification System. At last, through the test result we could judge whether the stock and the exchange rate was influenced by the news [16].

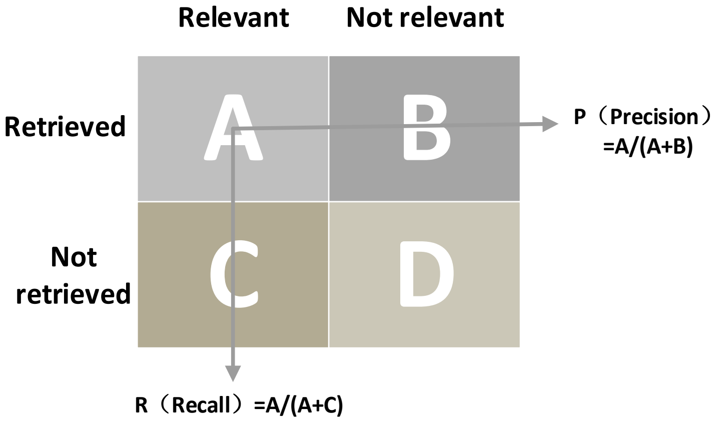

Fig. 4. The Principle Diagram of Precision and Recall Rate

4.1. The Evaluation Index

There are two most basic indicators in areas such as information retrieval, classification, recognition, translation, which are the precision and recall rate [17]. The formulas are as follows:

$$\text{Recall} = \frac{\text{number of the retrieved documents}}{\text{the total number of all relevant documents}}$$

$$\text{Precision} = \frac{\text{number of the retrieved documents}}{\text{the total number of all the retrieved documents}} \quad (2)$$

As is shown in Fig. 4, A+B+C+D refers to the number of all the documents, and A refers to the number of both relevant and retrieved documents, and B/C/D are all in the same way with A. We can easily get to the conclusion:

$$\text{Recall} = \frac{\text{A}}{\text{A+C}}, \text{ Precision} = \frac{\text{A}}{\text{A+B}} \quad (3)$$

Fig. 4. The principle diagram of precision and recall rate

Precision and recall rate always influence each other. We want both precision and recall rate are high, but usually precision and recall rate can’t be a high level at the same time. If both them are low, of course, that's something go wrong. We hope the result can be great, and meanwhile, we also hope to cover more news [18]. So, we usually use the measure that combines precision and recall rate, which is the harmonic mean of precision and recall, called balanced F-score (F_1).

$$F_1 = \frac{2*\text{P}*\text{R}}{P+R} \quad (4)$$

where the value of F_1 is between 0 and 1, and the higher value of F_1 means the performance of the method is the better.

4.2. Results Analysis

4.2.1. Comparison of Different Feature Selection Method

We use the information gain (IG), mutual information (MI), cross entropy (CE), statistical Evidence (CHI) and the Weight of Evidence for Text (WET), these five kinds of feature selection methods to analysis news text of 3 stocks. The result is showed in Table 1.

Table 1. Results of Different Feature Selection Methods (Poly Real Estate).

	Recall	**Precision**	F_1
IG	85.49%	90.78%	0.880516
MI	57.97%	64.47%	0.610506
CE	85.76%	88.17%	0.869457
CHI	84.71%	91.54%	0.879939
WET	86.01%	89.12%	0.875343

Table 2. Results of Different Feature Selection Methods (Petro China).

	Recall	**Precision**	F_1
IG	85.03%	85.95%	0.85483
MI	79.70%	79.41%	0.795552
CE	81.64%	82.20%	0.819165
CHI	79.72%	79.94%	0.798263
WET	81.64%	82.20%	0.819165

According to the former three tables, IG shows a better performance comparing to other feature selection methods in the prediction of stock.

Due to the difference in stock news and foreign currency news, we chose the Euro and US dollar exchange rate to test whether the former conclusion also apply to the currency news.

Table 3. Results of Different Feature Selection Methods (Ping An Insurance).

	Recall	Precision	F_1
IG	81.58%	85.31%	0.834012
MI	74.09%	78.44%	0.762027
CE	81.83%	83.75%	0.827754
CHI	80.05%	84.06%	0.820095
WET	79.03%	83.72%	0.813064

Table 4. Results of Different Feature Selection Methods (Euro-Usd Exchange Rate).

	Recall	Precision	F_1
IG	75.00%	87.18%	0.806322
MI	85.00%	91.43%	0.880974
CE	90.00%	93.94%	0.919273
CHI	57.50%	81.52%	0.674356
WET	77.50%	88.16%	0.824861

According to the Table 4, CE shows a better effect, and gets a better results.

4.2.2. Compared before and after corpus optimization

Based on original corpus we did an optimization for the corpus, and we compared the prediction results before and after optimization. The results shows in Table 5-8.

Table 5. Results of Before and After Optimization (Poly Real Estate).

	Recall	Precision	F_1
Before	85.49%	90.78%	0.880516
After	88.29%	92.40%	0.902957

Table 6. Results of Before and After Optimization (Ping an Insurance).

	Recall	Precision	F_1
Before	81.58%	85.31%	0.834012
After	90.81%	91.60%	0.912038

Table 7. Results of Before and After Optimization (Petro China).

	Recall	Precision	F_1
Before	85.03%	85.95%	0.85483
After	86.95%	88.95%	0.879366

Table 8. Results of Before and After Optimization (Euro-Usd Exchange Rate).

	Recall	Precision	F_1
Before	90.00%	93.94%	0.919273
After	92.11%	95.31%	0.936815

According to the tables, after corpus optimization, there is an increase in F_1. Specifically, the optimization of corpus can rise F_1 at least 1.76%, with maximum 7.8%. The precision can approach or exceed 90%, which is much better.

4.2.3. Results overview

Through comparing multiple stocks and foreign currency text classification system, we found that IG is the best method for stock news and CE is the best method for foreign currency news under the premise that feature space, text classification, and other parameter remain unchanged. If we optimize the corpus circularly, we can get more targeted corpus, so that we can get a better prediction. These verified financial news have great extent influence on the trend of stock and currency price.

5. CONCLUSION

Taking the financial market as the object of study, this paper studies the influence of financial news on stock and foreign exchange market trend. We used the new method that enrich the corpus circularly to optimize the corpus. This method of optimizing the corpus can effectively make the corpus more targeted, more financial related and smaller, so that the whole process of prediction can be more simple and accurate. In the result, we can get the better prediction performance.

6. ACKNOWLEDGEMENTS

This work is partially supported by NSFC under Grant No. 61273002, the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions No. CIT&TCD201304025 and the Key Science and Technology Project of Beijing Municipal Education Commission of China No. KZ201510011012.

7. REFERENCES

[1]. Hagenau, Michael, M. Liebmann, and D. Neumann. "Automated news reading: Stock price prediction based on financial news using context-capturing features." *Decision Support Systems* 55.3(2013):685–697.

[2]. Luss, R., and A. D'Aspremont. "Predicting abnormal returns from news using text classification. Wroking Paper from ORFE." *Quantitative Finance* (2009).

[3]. Antweiler, Werner, and M. Z. Frank. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance* 59.3(2004):1259-1294.

[4]. Bozhao L, Na C. AND Jing W., "Text categorization system for Stock prediction." *International Journal of u- and e- Service, Science and Technology*,2015,8(1), pp.4-5.

[5]. Liang, Xun, et al. "Associating stock prices with web financial information time series based on support vector regression."*Neurocomputing* 115(2013):142-149.

[6]. Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." *Wseas Transactions on Computers*4.2(2005):966-974.

[7]. Kaya, M. İ Yasef, and Karslıgil, M. Elif. "Stock price prediction using financial news articles." *Information and Financial Engineering (ICIFE)*, 2010 2nd IEEE International Conference on IEEE, 2010:478-482.

[8]. Laursen, A. L., B.,Mousten, AND V., Jensen. "Using an AD-HOC Corpus to Write About Emerging Technologies for Technical Writing and Translation: The Case of Search Engine Optimization", *Professional Communication, IEEE Transactions*, 2014,57(1),pp.2-10.

[9]. S. Biber, S. Conrad, and R. Reppen, "Corpus Linguistics: Investigating Language Structure and Use. Cambridge", *UK: Cambridge Univ. Press*, 1998.

[10]. Li, Xiangdong, and C. Zhang. "Research on enhancing the effectiveness of the Chinese text automatic categorization based on ICTCLAS segmentation method." *Software Engineering and Service Science (ICSESS)*, 2013 4th IEEE International Conference on IEEE, 2013:267-270.

[11]. Uğuz, Harun. "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm." *Knowledge-Based Systems* 24.7(2011):1024–1032.

[12]. Xu, Y.Q.A New Feature Selection Method Based on Support Vector Machines for Text Categorization, *ProQuest Dissertations and Theses* ,2006, pp.6-15.

[13]. Cao, Jianfang, and H. Wang. "An improved incremental learning algorithm for text categorization using support vector machine." *Journal of Chemical & Pharmaceutical Research* (2014).

[14]. Hsu, C. W. AND Lin, C. J. A Simple Decomposition Method for Support Vector Machines ,*Machine Learning*, 2002,46, pp.291-314.

[15]. Manne, Suneetha, et al. "Features Selection Method for Automatic Text Categorization: A Comparative Study with WEKA and RapidMiner Tools." *ICT and Critical Infrastructure*: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II. Springer International Publishing, 2014.

[16]. Hussein, Ashraf S., I. M. Hamed, and M. F. Tolba. "An Efficient System for Stock Market Prediction". *Intelligent Systems*'2014. Springer International Publishing, 2015:871-882.

[17]. Rose, Stuart J., W. E. Cowley, and V. L. Crow. "Systems and Processes for Identifying Features and Determining Feature Associations in Groups of Documents." US, US20130173257 A1. 2013.

[18]. Wolf, Christian, et al. "Evaluation of video activity localizations integrating quality and quantity measurements." *Computer Vision & Image Understanding* 127.10(2014):14–30.