ICIC Express Letters
Part B: Applications
Volume **7**, Number **12**, December **2016**
 ISSN 2185-2766
pp. **2691–2698**

MINING QUALITY DIFFERENCES OF CHINESE INFORMATION DISCLOSURE

Xinying Qiu[1], Kebin Deng[2], Shengyi Jiang[1] and Likun Qiu[3,*]

[1]CISCO School of Informatics
Guangdong University of Foreign Studies
No. 2, North Baiyun Ave., Baiyun Dist., Guangzhou 510420, P. R. China

[2]School of Economics and Commerce
South China University of Technology
No. 381, Wushan Road, Tianhe Dist., Guangzhou 510641, P. R. China

[3]School of Chinese Language and Literature
Ludong University
No. 186, Hongqi Mid-Road, Zhifu District, Yantai 264025, P. R. China
*Corresponding author: qiulikun@gmail.com

Received May 2016; accepted August 2016

Abstract. *We present a series of experiments designed to mine the factors behind the quality differences in information disclosure in China. Our main research questions are: 1) whether certain quality categories evaluation is easier to be transferred to machine-learning based approaches than the others; 2) if and which sectional or temporal factors may contribute to the predictive differences in disclosure quality. We build multi-class text categorization models combined with different feature selection methods to predict different disclosure quality. Among our experiments results, we found that "state-owned" firms stand out in their performance on predicting "Excellent" and "Fail" quality report, while "non-state" firms perform significantly better in predicting "Good" and "Pass" reports. The results may provide insights for regulators overseeing quality evaluation standards. Our findings may also point to features of particular interest to certain classes of disclosure quality, as well as help discovering prototypical models that better represent different qualities.*
Keywords: Prediction, Information disclosure quality, Text mining application

1. **Introduction.** For any capital market, high quality information disclosure plays an important role in maintaining market efficiency. However, in the realm of applying machine learning and text mining techniques to the study of annual reports, we have seen a lot more research on using English-language annual reports to study various research questions. These include: using soft information within annual reports to analyze financial risks among companies [1]; developing algorithm to automatically identify risk factors from annual reports [2]; analyzing the association of annual reports features at document and textual level with different financial factors [3]; mining the sentiment trends within annual reports [4]; and combining investor sentiment with historical pricing information to predict stock price direction [5,6].

On the other hand, the study about Chinese annual reports and their disclosure quality mainly comes from researchers of finance, economics, and accounting domains. For example, Courtis and Hassan [7] examined the reading ease of the English and Chinese versions of annual reports. Other research [8] found that greater disclosure transparency has positive effect and is one of the key factors influencing lower cost of equity capital in China. Machine learning and text mining technologies have been proven effective in utilizing English financial reports for prediction, but have rarely been explored on studying Chinese reports. Aside from the language differences, the different economical contexts

for these filings between China and western countries may contribute to special and interesting observations which are probably easier to be discovered with text mining and machine learning approaches and methodologies. These observations have motivated our effort in conducting a series of research looking into the factors behind disclosure quality in China.

One important feature about Chinese annual reports is the manual ratings of the information disclosure quality by Chinese analysts. These manual ratings of disclosure quality are available for companies traded at the Shenzhen Stock Exchange. Researchers in the accounting and finance field have explored the disclosure quality ratings to show how disclosure quality is related to cost of equity capital [9], and stock liquidity [10]. The methods employed are generally semi-automatic statistical analysis and regression modeling. In this paper, we utilize analysts' manual ratings of Chinese annual reports to build predictive models for analyzing different disclosure qualities, and to explore the features and factors that may influence quality differences of information disclosure in China. More specifically, we address the following problems: 1) evaluate and compare different models' performance in assessing Chinese information disclosure quality; 2) discover patterns and factors that may make a difference in quality evaluation performance. We hope to explore insights from our experiment results about how well human judgment of quality differences may be transferred into machine-learning based methods, and to examine the implications if any behind the challenges arisen from the machine learning based approaches.

We will show that the study of these research problems is not merely transference of existing technologies to data of the same domain but just in another language. We hope to contribute to the applied research of information retrieval and text mining in the following ways. 1) the challenges in automatic evaluation of certain categories of Chinese disclosure quality may provide enlightenment in re-examination of human judgment of annual reports; 2) the factors or features discovered by our automatic methods influencing the disclosure quality evaluation may extend and broaden our current understanding about disclosure quality.

The rest of the paper is organized as follows. Section 2 provides detailed discussion of our methods and experiment design. Section 3 presents experiments results and analyzes how the results provide insights to our research questions. We conclude in Section 4 with summary of our contribution, findings, and discussions.

2. **Methodology and Design.**

We formulate our design to mine the differences in disclosure quality with a multi-class classification approach. We use the analysts' manual quality ratings for annual reports at Shenzhen Stock Exchange as our gold standard. To validate the system's feasibility and evaluate the model's performance, we conduct a series of stratified cross-validation experiments. We further evaluate how the best model from cross-validation would perform in practice with simulation experiments. Figure 1 presents the component models and flow of experimentation of this paper. The details of our approaches and design are presented as follows.

As shown in Figure 1, our research structure consists of the following steps. 1) We first build representation model for our document collection with feature selection method to filter out uninformative term. We select document classification algorithms and use randomly selected development set to tune the parameters needed for the classifiers. 2) We use cross-validation experiments to decide on the best classification model paired with its most suitable document model. 3) With the best model, we perform analysis on the impact of the corporate features on the model's performance. 4) To evaluate the model's performance in practice, we simulate the model's operation and use temporal analysis to assess the influence of historical data on the model's predictive ability.

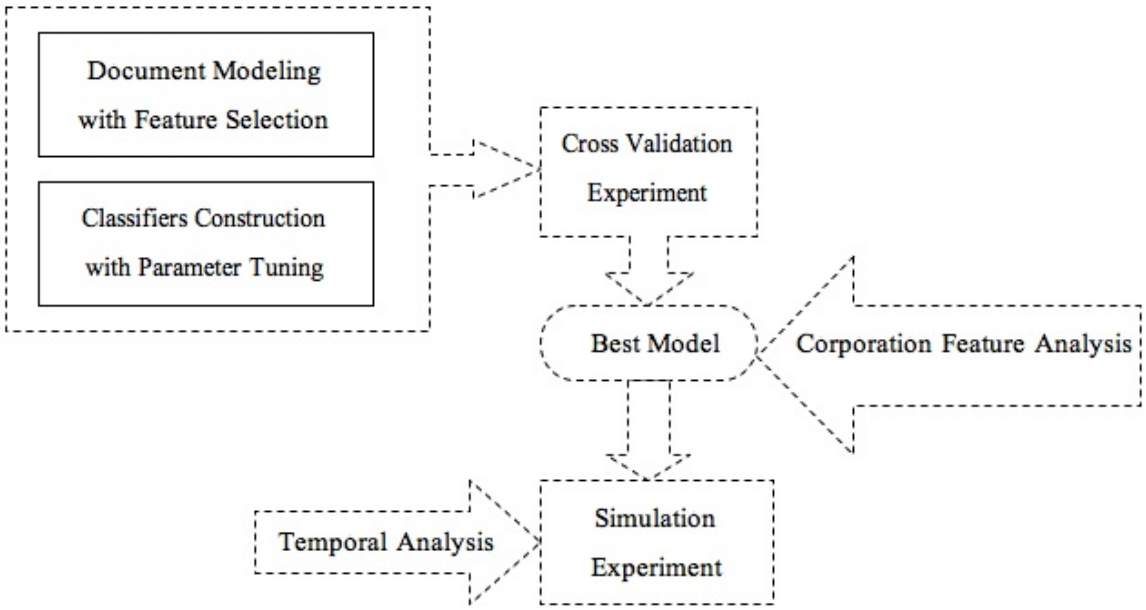

FIGURE 1. Research framework

2.1. **Data collection and class definitions.** We automatically retrieved all the Chinese annual reports with disclosure quality ratings for companies traded at Shenzhen Stock Exchange from 2001 to 2010. The disclosure quality rating scheme is a 4-level ranking of "Excellent", "Good", "Pass", and "Fail". We also retrieved financial data from CSMAR database[1], to distinguish firms as "glamour", "value", "large", "small", "state-owned" and "non-state" firms. After filtering out firms with incomplete sectional or financial data, as well as reports with errors, we obtain a sample set of a total of 5800 company annual reports spanning from 2001 to 2010[2]. The distribution of the reports along with quality ratings is indicated in Table 1.

2.2. **Document modeling.** In information retrieval, documents are typically modeled as vectors of terms with weighting for each term to indicate the importance of term in contributing to the documents' main content. We adopt the typical bag-of-word representation model with TF*IDF weighting scheme to build document vector model. This model is the most successful and widely used where the positions of terms are ignored and the term weighting scheme measures the descriptive information and the importance of terms. Based on other explorative studies with similar data set, we find that a proper TF*IDF weighting scheme for Chinese annual report vectors to be "ltn" weights is calculated as follows:

$$(1 + \log(tf_{t,d})) \log \frac{N}{df_t} \tag{1}$$

where $tf_{t,d}$ is term t's raw term frequency in document d; df_t is the term t's document frequency in the corpus; N is the total number of documents in the corpus.

2.3. **Feature selection.** In Chinese language, terms may be composed of single words as well as multi-word phrases. Based on our pilot study, we use Lucene system and Lucene's ICTCLAS dictionary to segment and index documents. Our indexing experiment

[1]http://www.gtarsc.com/.

[2]Since annual reports of a given year will be available by April the following year, as we prepared for this manuscript in 2014, we could only obtain data up till year 2012. We therefore conduct research on the 10 years of data from 2001 to 2010 as it would be a more well-organized and complete data set for a decade.

TABLE 1. Distribution of annual reports with quality assessment

Year	Total # of Docs	By Quality Labels				By Market -to-book		By Firm Size		By Ownership	
		Excellent	Good	Pass	Fail	Glam	Value	Large	Small	Non-state -owned	State -owned
2001	420	28	169	198	25	233	187	171	249	74	346
2002	434	32	204	166	32	150	284	203	231	101	333
2003	456	39	244	151	22	109	347	236	220	123	333
2004	445	27	275	126	17	48	397	241	204	127	318
2005	332	25	176	106	25	31	301	147	185	118	214
2006	538	53	289	170	26	154	384	254	284	214	324
2007	637	62	336	215	24	529	108	309	328	296	341
2008	715	77	432	191	15	218	497	350	365	352	363
2009	764	93	521	134	16	612	152	414	350	410	354
2010	1059	143	728	173	15	816	243	576	483	738	321
Total	5800	579	3374	1630	217	2900	2900	2901	2899	2553	3247

Note: "Glam" (i.e., glamorous) versus "Value" firms are distinguished with the median of Market-to-book ratio. "Large" and "Small" firms are distinguished with the median firm size value. "State-owned" and "Non-state" indicator data are retrieved from CSMAR database directly. The Quality labels of "Excellent", "Good", "Pass", and "Fail" refer to disclosure quality score for the annual reports provided by analysts of Shenzhen Stock Exchange.

originally extracted 62687 terms (including single word and multi-word terms). After discarding meaningless symbols to preserve a feature set of 40797 Chinese terms, we planned to further reduce the feature set without sacrificing our predictive accuracy.

We employ feature selection methods in hope of preserving a smaller vocabulary that would provide better interpretation for disclosure quality without sacrificing predictive accuracy. Based on previous research for feature selection, we considered two feature selection approaches: 1) choose a document frequency threshold (noted as DF threshold in this paper) to discard terms that appear in a few documents [11]; and 2) select the top-K most meaningful terms for each document [12]. For both approaches, we use grid search to find the best parameter value for DF and K, with a random sample of 10% of our data set. Our experiments yield an optimal DF threshold of 15 (i.e., using terms with document frequency of 15 and above) and K of 1000 (i.e., for each document, only use terms with the top 1000 highest TF*IDF weights).

2.4. **Classifiers construction.** Our quality assessment model is based on support vector machine (SVM) classifiers. We consider two different options for our four-class classification problem. First, we perform a one-against-rest classification for each class. We combine the predictive scores of the four binary classifiers and use the highest score to assign the class label. For each binary classifier, we use the SVM-light implementation of SVM with linear kernel. We noted this model as SVM-score. Second, we use algorithms designed specifically for multi-class classification. In order to decide on the optimal value for c (i.e., the trade-off between training error and margin) in the multi-class-classification model, we randomly sampled 10% (i.e., 579 reports) of data from our dataset to perform classification with cross-validation. We found the optimal value for c to be 15. We noted this model as SVM-multi.

2.5. **Experiment design.** We consider two experiment designs. One is 10-fold cross-validation with stratification to build models, and evaluate with average accuracy and paired two-tailed t-test to find the best model. The other is a simulation model where we use data of years previous to test year to build classification model and predict report quality class labels in the test year. This is an implementable model that can be applied directly to practice. We use the majority vote results for the four classes as the baseline

predictive accuracy. The majority class in the data set is class "Good", resulting in a baseline accuracy of 58.17%.

3. Results and Analysis.

3.1. Overall and by-class accuracy.

Overall, with two multi-class classification mechanism (i.e., SVM-score and SVM-multi) and two feature selection methods (i.e., DF and Top-K), we evaluated six different models against our baseline of majority vote. Table 2 presents results from 10-fold cross validation with paired two-tailed t-test for significance analysis. We observe that all predictive models can achieve accuracies significantly better than majority baseline. However, the predictive models perform similarly with no particular model standing out in predictive accuracy. Considering feature selection methods, since the vocabulary sizes for "no-feature-selection", "DF", and "Top-K" are 40797, 18620, 40780 respectively, we decide to select DF+SVM-score for the following experiment and analysis due to its smaller feature set, easier interpretation, and equivalent accuracy.

As shown in Table 3, with DF+SVM-score model, we achieve class-specific accuracies of: 91.05% for class "Excellent", 66.86% for class "Good", 73.1% for class "Pass", and 96.33% for class "Fail". We look into the details of the predictive accuracy with a contingency table analysis of DF+SVM-score model. As shown in Table 4, the predictions for the "Excellent" and "Fail" categories of annual reports achieve the highest accuracy of up

Table 2. T-test for comparing average accuracy of four-class classification models from cross-validation

P-Value	**SVM-score (63.15%)**	**SVM-multi (63.48%)**	**DF+SVM -score (63.43%)**	**DF+SVM -multi (63.48%)**	**Top-K+SVM -score (63.7%)**	**Top-K+SVM -multi (63.76%)**
Baseline (58.17%)	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
SVM-score (63.15%)		0.5075	0.5587	0.5301	0.4554	0.4569
SVM-multi (63.48%)			0.9998	0.0008	0.732	0.655
DF+SVM -score (63.43%)				0.7465	0.5387	0.4913
DF+SVM -multi (63.48%)					0.6074	0.5189
Top-K+SVM -score (63.7%)						0.8104

Note: Values in parentheses are average accuracies from cross-validation except for "Baseline". Cell entry values are p-values from two-tailed t-test for significance analysis.

Table 3. Accuracy of DF+SVM-score binary classifier for predicting each class

Average accuracy of DF+SVM-score classifier from cross-validation				
Class label	**Excellent**	**Good**	**Pass**	**Fail**
Average accuracy	91.05%	66.86%	73.1%	96.33%

Table 4. Contingency table of DF+SVM-score multi-class models

DF+SVM -score Model	True Class Label of Information Disclosure Quality				Total
	Excellent	Good	Pass	Fail	
Predicted Excellent	2.69%	1.47%	0.21%	0%	4.36%
Predicted Good	7.16%	50.59%	17.67%	1.22%	76.64%
Predicted Pass	0.14%	6.09%	10%	2.36%	18.59%
Predicted Fail	0%	0.03%	0.22%	0.16%	0.41%
Total	9.98%	58.17%	28.1%	3.74%	100%

to 96%. However, the largest prediction percentage by the model comes from predicting "Good" class of reports at 76.64% with the true percentage of "Good" class at 58.17%. This implies that the multi-class model is able to identify "Excellent" or "Fail" quality reports with high precision, but makes the most mistakes in judging of the "Good" quality reports. Another observation is that the largest incorrect classification errors occur for predicting "Pass" reports as "Good" report (17.67%). On the contrary, our model did not make any mistakes in predicting "Excellent" as "Fail" (0%) or predicting "Fail" as "Excellent" (0%). This implies that it is a lot easier to transfer human judgment of "Excellent" and "Fail" class of disclosure quality into a machine-learning based system, but far less so for "Good" and "Pass" class of disclosure quality.

3.2. **Analysis by firm features.** Table 5 presents the sectional analysis about how reports of different firm-level features (i.e., glamour vs. value, large vs. small, state-owned vs. non-state) may differ in their predictive accuracy. We observe that distinguishing firms by market-to-book ratio or by firm size does not contribute to the predictive accuracy of disclosure quality, given the relatively larger p-value from comparing the cross-validation results. An interesting observation, however, is the clear distinction in predictive accuracy between "state-owned" and "non-state-owned" firms, with the former's predictive accuracy significantly worse than overall benchmark, and the latter's significantly better. This is of particular interest because: 1) "state-owned" and "non-state-owned" firm category is unique in China and not considered as a firm-level feature for firms in western countries; and 2) the results imply that it is more challenging to conduct automatic assessment of "state-owned" firms' disclosure quality, which could be inferred as difficulty lying in either the data or the evaluation criteria.

Table 6 contrasts the class-specific performances for state-owned and non-state firms. We observe that for predicting "Excellent" reports, both "state-owned" and "non-state" firms perform significantly better than benchmark model (i.e., DF+SVM-score) with all data. However, "state-owned" firms are even more significantly better than "non-state"

TABLE 5. T-test analysis of overall predictions for firms of different sections

P-Value	**State-owned (61.33%)**	**Non-state (65.36%)**	**Glam (62.17%)**	**Value (62.61%)**	**Large (63.04%)**	**Small (62.19%)**
Overall (63.43%)	**0.0193**	**0.003**	0.049	0.3746	0.5601	0.0971

Note: Values in parentheses are average accuracies from cross-validation. Cell entry values are p-values from two-tailed t-test for significance analysis.

TABLE 6. T-test analysis of class-specific predictions for state-owned and non-state firms

Excellent	**State-owned (95.56%)**	**Non-state (92.53%)**	**Good**	**State-owned (64.6%)**	**Non-state (68.82%)**
Overall (91.05%)	< 0.0001	< 0.0001	**Overall (66.86%)**	0.0033	0.0456
State-owned (95.56%)	–	< 0.0001	**State-owned (64.6%)**	–	0.0033
Pass	**State-owned (72.1%)**	**Non-state (74.13%)**	**Fail**	**State-owned (96.85%)**	**Non-state (95.56%)**
Overall (73.1%)	0.0049	0.0826	**Overall (96.33%)**	< 0.0001	< 0.0001
State-owned (72.1%)	–	0.0001	**State-owned (96.85%)**	–	< 0.0001

Note: Values in parentheses are average accuracies from cross-validation. Cell entry values are p-values from two-tailed t-test for significance analysis.

firms (95.56% vs. 92.53%). For predicting "Good" reports, "state-owned" firms' predictions are significantly worse than benchmark, while "non-state" firms' reports are significantly better to predict than benchmark. For predicting "Pass" reports, "state-owned" firms' predictions are significantly worse than benchmark. However, "non-state" firms perform just the same as benchmark though significantly better than "state-owned". For predicting "Fail" reports, "state-owned" firms are significantly better than benchmark while "non-state" firms' are significantly worse. Overall, "state-owned" and "non-state" firm tie in their competition for better predictive accuracy. However, their winning patterns are very different. It is much easier to predict "Excellent" and "Fail" class of quality for state-owned reports, but easier to predict the "Good" and "Pass" class of quality for "non-state-owned" firms.

3.3. **Simulation experiment analysis.** We conduct an implementable experiment to simulate the real-life application of predictive models. The purpose is two-fold. We want to explore the feasibility of applying models to practice, as well as to examine if and how different amount of historical data may influence the implementable model. We pick year 2010 as our test year, and use 8 sets of historical data covering different length of time-span to build different predictive models. The predictive model applied is DF+SVM-score model. Figure 2 shows the accuracies of these 8 implementable models. We observe that the simulation accuracies are quite consistent around 70%, better than the average accuracy of 63% from 10-fold cross-validation results with the same model. This is also better than the majority vote baseline of 58%. Considering the consistency in predictive performance for models built with 1 up to 9 years of historical data, we may infer that the historical data of different time-spans may not influence the performance of building and applying models to practice.

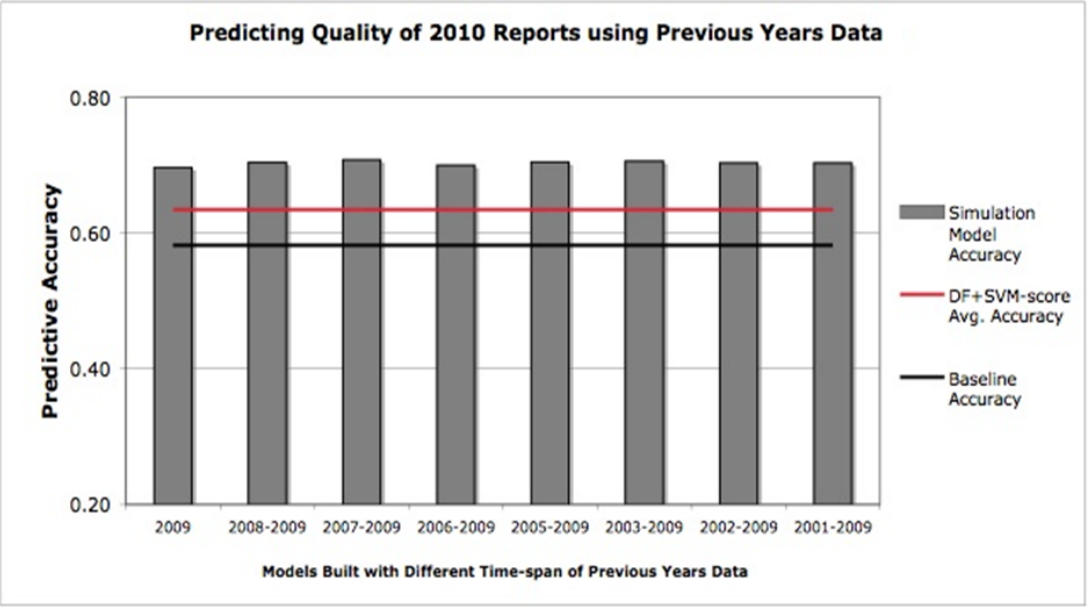

FIGURE 2. Simulation model to predict reports quality of Year 2010 with historical data of different time-spans

4. **Conclusions.** We presented a series of experiments designed to mine the factors behind the quality differences in information disclosure in China. Our main research questions are: 1) whether certain quality categories evaluation is easier to be transferred to machine-learning based approaches than the others; 2) if and which sectional or temporal factors may contribute to the predictive differences in disclosure quality. We address our research goals with the following approaches. We build multi-class text categorization

models combined with different feature selection methods to predict different disclosure quality. We analyze our models' performance by firm features to mine factors contributing to predictive accuracy. We simulate real-life practice of evaluating reports quality and assess the feasibility of applying models in practice with different time-spans of historical data.

Our models perform well in cross-validation as well as real-life simulation. The findings suggest that Chinese annual reports do present some social and economic characteristics underlying their disclosure quality differences that we do not normally find in the study of English annual reports of other countries. We found that the "Excellent" and "Fail" reports are much easier to model and predict than the "Good" and "Pass" reports. Interestingly, "State-owned" firms stand out in their performance on predicting "Excellent" and "Fail" quality reports, while "non-state" firms perform significantly better in predicting "Good" and "Pass" reports. The results imply that it is more challenging to conduct automatic assessment of "state-owned" firms' disclosure quality, which could be inferred as difficulty lying in the quality evaluation criteria. These observations might provide insights for regulators overseeing quality evaluation standards. Our findings may also point us in our future work to look further into features particular to certain classes, as well as discovering prototypical models that better represent different disclosure qualities.

Acknowledgments. This work is partially supported by Grant 12YJAH103 from the Ministry of Education of China Project, Grant GDITSEC-a-2013041 from Guangdong Information Technology Security Evaluation Center Program Project, China Central Fundamental Research Funds (No. 522014Y-3354) and National Natural Science Foundation of China (No. 61572245).

REFERENCES

[1] M. F. Tsai and C. J. Wang, On the risk prediction and analysis of soft information in finance reports, *European Journal of Operational Research*, 2016.

[2] K.-W. Huang and Z. Li, A multilabel text classification algorithm for labeling risk factors in SEC Form 10-K, *ACM Trans. Management Information Systems*, vol.2, no.3, pp.18:1-18:19, 2011.

[3] Y. Bao and A. Datta, Simultaneously discovering and quantifying risk types from textual risk disclosures, *Management Science*, vol.60, no.6, pp.1371-1391, 2014.

[4] J. S. J. Ren, H. Ge, X. Wu et al., Effective sentiment analysis of corporate financial reports, *The 34th International Conference on Information Systems*, Milan, pp.1-9, 2013.

[5] R. Feldman, S. Govindaraj, J. Livnat and B. Segal, Managements tone change, post earnings announcement drift and accruals, *Review of Accounting Studies*, vol.15, no.4, pp.915-953, 2010.

[6] M. Rechenthin, W. N. Street and P. Srinivasan, Stock chatter: Using stock sentiment to predict price direction, *Algorithmic Finance*, vol.2, pp.169-196, 2013.

[7] J. K. Courtis and S. Hassan, Reading ease of bilingual annual reports, *Journal of Business Communication*, vol.39, no.4, pp.394-413, 2002.

[8] W. Wang and G. Jiang, Information disclosure, transparency and the cost of capital, *Economic Research Journal*, vol.7, 2004.

[9] Y. Zeng and Z. Lu, The relationship between disclosure quality and cost of equity capital of listed companies in China, *Economic Research Journal*, vol.2, 2006.

[10] Q. Chen, Disclosure quality and market liquidity, *South China Journal of Economics*, vol.10, 2007.

[11] R. Balakrishnan, X. Y. Qiu and P. Srinivasan, On the predictive ability of narrative disclosures in annual reports, *European Journal of Operational Research*, vol.202, pp.789-801, 2010.

[12] M.-C. Lin, A. J. T. Lee, R.-T. Kao and K.-T. Chen, Stock price movement prediction using representative prototypes of financial reports, *ACM Trans. Management Information Systems*, vol.2, no.3, pp.19:1-19:18, 2011.