The Pennsylvania State University

The Graduate School

College of Engineering

DATA MINING ON CORPORATE FILLING BASED ON BAYESIAN LEARNING APPROACH

A Thesis in

Industrial Engineering and Operations Research

by

Xiaocheng Hu

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

August 2015

The thesis of Xiaocheng Hu was reviewed and approved* by the following.

Tao Yao

Associate Professor of Industrial and Manufacturing Engineering

Thesis Advisor

Terry P. Harrison

Professor of Supply Chain and Information Systems

Smeal College of Business

Harriet B. Nembhard

Professor of Industrial Engineering

Interim Department Head of Industrial Engineering and Manufacturing Engineering

*Signatures are on file in the Graduate School.

ABSTRACT

Most of the researches on corporate filling mainly focus on qualitative analysis. This thesis used quantitative method-Bayesian Learning Machine in analyzing the information content of future prediction statements (FPS) in the Management Discussion and Analysis section of 10-Q fillings.

The thesis proposed a new approach that involved a combination of mathematical methods and text analysis. Naïve Bayesian machine learning approach was used to examine the prediction of future performance of the company.

In conclusion, the average profit tone of FPS is negatively associated with profit predict and negatively associated with other predict which includes the prediction related to employees, regulations, accounting and other. The liquidity tone is negatively associated with other predict. The overall tone is negatively associated with other predict.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Tao for being my thesis adviser. His guidance, support and patience throughout the research and writing of the thesis were invaluable. I also want to thank Dr. Harrison for being my thesis reader. His support and guidance were helpful in completing my thesis within the given time limits.

Finally, I would like to thank my family for their continuing support in everything.

Chapter 1

INTRODUCTION

Data mining is the technology that can be used to manage large volume of information and make inferences to assist in decision-making.

1.1 Data and information

There are various definitions of data in current literatures. According to Juris Kelley (2002), data was comprised of the basic, unrefined, and generally unfiltered information. Debra M. Amidon (1997) defined data as elements of analysis. Thomas H. Davenport (2000) defined data as sets of discrete, objective facts about events and structured records of transaction. In computer science, data is defined as a set of quantitative or qualitative values. Whereas, in the real world, data could be either facts, numbers or text. There are many ways to process data, such as classification, aggregation, and summarization. Currently, organizations are accumulating large volume of data in various formats and database. This includes operational or transactional data, nonoperation data, and meta data.

Information is an alternative form of data that has been processed. Amrit Tiwana (2001) defined information as processed data that is formalized, explicated and can be easily packaged into reusable forms. Georg Von Krogh, Ichijo and Nonake (2000) defined information as data putting in context, which is related to other pieces of data and formed the basis for knowledge. According to Ikujiro Nonaka and Hirotaka Takeuchi (1995) information is a flow of messages. In all, information is derived from data and

conveyed as messages. Information is the organization of raw data in a meaningful manner. . There are various methods to process data, and data mining is one of them.

1.2 Data Mining

There are many definitions of "Data Mining". Some define data mining as the process of analyzing data from different perspectives and summarizing it into useful information [1]. The others define data mining as the practice of automatically searching large volumes of data to discover patterns and trends that go beyond simple analysis [2]. In general, data mining could be regarded as an analytic process to explore consistent patterns among data, and then apply the detected patterns on new subsets of data in order to draw useful information.

There are few steps involved in data mining.

The first step is data exploration which involve three aspects: (1) data cleaning, (2) data transformations, (3) selection of records and features.

The second step involves model building and validation. In this stage, various models were generated and the model with the best performance was selected. There are various model building techniques such as, Bagging, Boosting, Stacking and Meta-Learning.

The final step is model deployment. The best model that was selected in the previous stage was tested with independent set of new data to generate prediction or estimates of the expected outcome. Also, there are other forms of data mining to identify natural grouping in the data. This is known as "grouping". For example, a model might identify

the segment of the population that has an age within specified range. With data mining, we can make prediction, answer the right question, and understand the data.

Today, data mining is mostly used by companies in retail, marketing and so on. However, data mining is rarely used in text analysis.

1.3 Text mining

Text mining is one approach to transform qualitative or unstructured raw data into computer-readable data. The Oxford English Dictionary defines text mining as the process or practice of examining large collections of written resources in order to generate new information, typically using specialized computer software. The purpose of text mining is to extract meaningful numeric indices from the text and make the information contained in the text accessible to the data mining algorithms. In this way, clusters of words in documents can be analyzed. In general, text mining can be defined the process of turning text into numerical data, and then cooperating the numerical data in analysis project such as predictive data mining projects. The general approach of text mining is to 'numericize' text, and then index all input documents [3]. Words are counted in order to compute a table of documents and words. We refine the process to exclude words such as, 'the', 'a', 'this', and to combine different grammatical forms of the same words such as, 'teach' and 'taught'. Once a table of words by documents has been derived, all data mining techniques can be applied to derive dimensions or to identify some kinds of words or terms that could best predict other outcome variables of interest.

Industry companies use text mining to analyze customers' behaviors in order to be more competitive. In academia, text mining is used to deliver efficiencies and new

knowledge in area such as particle physics and media. The results of text mining can be subsequently used into analysis, which is called as text analysis.

1.4 Thesis outline

The text analysis in this thesis is based on Naïve Bayesian Learning Machine. Bayes theorem provides us a way to calculate the posterior probability. Naïve Bayesian is based on Bayes' theorem with independence assumptions between predictors [4]. In this thesis, it was assumed that probability of each word in the document was mutually unaffected. This assumption simplified the computation and avoided the problem of 'curse of dimensionality'. However, this assumption has some limitations. For example, the words 'adverse effect' are mostly appeared with 'material'. But, empirical results from other fields imply that the assumption may have little effect on result.

The thesis will be presented in the following format. Chapter 2 will review the literatures in the areas of researches pertaining to the topic. The formulation of the Naïve Bayesian and text classification will be discussed in Chapter 3. Chapter 4 will contain the implementation of the model for processed data set using the Perl language followed by SPSS Analysis. The relationship between tones of FPS and company's future performance will be studied. Finally, Chapter 5 will include concluding remarks and some recommendations for further research.

Chapter 2

Literature Review

This chapter reviews the relevant works from the literatures before proceeding to the problem description and model formulation.

2.1 Data Mining

A brief review of some of the articles discussing data mining and processing methods is presented below.

D. Hand (1998) translated data mining as a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning, and other areas [6]. It was stated that data mining is aimed to find previously unsuspected relationships which are of interested or value to the database owner. Statisticians can make significant contributions in this discipline.

U. Fayyad, G. Piatesky-Shapiro and P. Smyth (1996) provided an overview on data mining research filed and clarified hoe data mining and knowledge discovery in database are related to each other and to related fields, such as machine learning, statistics and databases [12]. The particular real-world applications were mentioned in the article and future research directions in the field were also pointed out.

X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. MClachlan, A. Ng, B. Liu (2008) discussed top ten data mining algorithms identified by the IEEE and ICDM [16].. They also provided a description of each algorithm and discussed the impact

of the algorithm. Moreover, they reviewed the current and future research on the algorithm.

M. E. J. Newman (2003) described empirical studies on the structure of network and the mathematical modeling of network, including random graph models and their generalizations, exponential random graphs and Markov graphs [18]. In the last part, they discussed the process, such as epidemic process, network failure, model displaying phase transitions, and dynamical systems.

2.2 Text Mining

Hearest (1999) has wrote the nascent field of "text data mining". At that time, the meaning of neither "text mining" nor "text data mining" was clear. Hearst discussed the relationship between text data mining and data mining, information access and corpus-based computational linguistics, but he didn't define "text data mining".

Sebastiani (2002) provided the definition of "text mining" as a phrase to denote any systems that analyzed large quantities of natural language text and detect lexical usage patterns in an attempt to extract probably useful information.

Raymond J Mooney and Un Yong Nahm (2003) described text mining as a method used to look for patterns in unstructured text. In the paper, a framework for text mining call DISCOTEX was presented, which used a learned information extraction system to transform text into more structured data and then to be mined for particular relationships.

Milos Radovanovic and Mirjana Ivanovic discussed approaches to the identification of patterns, including dimensionality reduction, automated classification and clustering. They also illustrated pattern exploration by two applications.

Przemyslaw Maciolek and Grzegorz Dobrowolski presented a prototype implementation of an Open Source Intelligence system which could extract and analyze significant quantities of accessible information.

The rapid development in web technologies led to the generation of massive amounts of data. Nikou Gunnemann (2001) presented an application case using D-VITA which was an interactive text analysis system that exploited dynamic topic mining to detect the latent topic structures and topic dynamics in a collection of documents.

Chapter 3

MODEL DESCRIPTION AND COMPUTATION

3.1 Model Description

To process text analysis, there are two general approaches. One is "rule-based approach", such as dictionary approach, and the other is "statistical approach". The first approach sets up a list of categories in which computer program will automatically classify words into different categories according to particular rules, such as the Linguistic Inquiry and Word Count software.

The second approach that is also called "qualitative approach" relies on statistical technique to classify documents. For example, we can calculate the correlation between the frequency of some keywords and the document type.

In this thesis, statistical approach was employed. There are several reasons to choose statistical approach instead of rule-based approach. First, there is no dictionary which is available for corporate fillings. Thus, rule-based approach may not work well for corporate fillings.

Second, the environment of a sentence is not considered in rule-based approach. For example, the word "decrease" would be a negative word if it is about revenue. But, it should be a positive word when it is related to cost. Thus, when rules are created in one substantive area and applied to another problems, it may cause serious errors.

Third, the rule-based approach seldom uses prior knowledge that researchers may have about the text. For example, a sentence in Management Discussion and Analysis section report is neutral. We should just leave it as neutral, unless we have convincing evidence that the sentence is of positive tone

Finally, we can use training data to classify content with higher efficiency by the statistical approach. Naïve Bayesian algorithm is applied in this thesis. Bayesian algorithm (Bayesian classifier) is an effective way to make text classification.

3.1.1 Naïve Bayesian Algorithm

In this thesis, Bayesian learning algorithm was used to make calculation. Naïve Bayesian algorithm is a way to calculate the probability. Bayesian method uses random variables, or more general unknown quantities to model all sources of uncertainty in statistical models [2]. In this study, sentences was departed to words and each word was weighted by time they appeared in the sentence. I did this in order to classify the sentence into particular category. In the thesis, there are three possible tone categories that are positive, negative and neutral (uncertain is combined in neutral category). The Naïve Bayesian algorithm automatically chooses the category which best suit the sentence by solving the following problem:

$$category = \underset{category \in categoriess}{\arg\max} \frac{P(words \mid category)P(category)}{P(words)}$$

$$TONE_i = \frac{1}{k}\sum TONE_{i,k}$$

Since the frequency of a word in the sentence is fixed in different categories, we can eliminate the $P(words)$. Then the problem becomes:

$$category = \underset{category \in categories}{\arg\max} P(words \mid category)P(category)$$

We assume that $w_1, w_2, w_3 ..... w_n$ are the words in the document. Then we can transform the problem as:

$$category = \underset{category \in categories}{\arg\max} P(w_1 \mid category) * P(w_2 \mid category) * .... * P(w_n \mid category) * P(category)$$

This is the formula used in the thesis to category each sentence.

In Naïve Bayesian algorithm, words are assumed to be independent from each other. This means the presence of one word will not affect the probability of presence of another word.

3.1.2 Text Classification

Naïve Bayes classifiers are a family of simply probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. In order to apply the Naïve Bayesian algorithm, text classification was processed to construct training data. All the 10-Q fillings of Microsoft between 2009 and 2012 were downloaded from the SEC Edgar Web site. Then, the Management Discussion and Analysis section were extracted from them. Next, the contents of the sections were splited into sentences using model in Perl. The detail process can be found in appendix A. The last step is to extract future prediction statements from these sections. For the rule of

selecting future prediction statements, sentences with word “would”, “will”, “can”, “could”, “should”, “expect”, “may”, “might”, “believe”, “anticipate”, “plan”, “seek”, “project” “hope”, “intend”, “target”, “aim”, “forecast”, “objective”, “goal” are all defined as future prediction statements.

Then, I implement “ stemming” and “stopword” process with Perl. Stemming, which is a process for reducing inflected words to their root form. For example, “students” and “student” are the same after stemming. I use the Lingua::Stem::Snowball module in Perl to complement this process. “Stopwords” are words that are frequently occurred in a language, such as: “or”, “and”, “it”, and “are”. I used Lingua::StopWords module in Perl to implement this process.

In order to get training data, I classified 54 future prediction statements which are automatically selected by excel program according to the tones. There are three tones for FPS: positive tone, negative tone, and neutral tone. Each sentence belongs to one of them. Also, content is another factor to classify selected sentences. I divided all the sentences into 12 categories according to their context, such as profit, operations, investing, and employee. Since the same word will have different tone in different situations, the details of my training data preparation are shown in appendix B.

3.2 Computation

Algorithm::NaiveBayes module is used in my thesis to process the computation. First of all, the word vectors should be converted to hash variable and then the hash variables learn from 54 manually coded sentences in the Bayesian classifier and run the training process. The two dimensions (category and tone) of all the future prediction statements

are output after computation. There are more than 5000 sentences.

Last step of calculation is to add the tone of each sentence to get the tone of the whole content. For each future prediction sentence i, I define its tone as the following:

$TONE_i = 1$ sentences with positive tone;

$TONE_i = -1$ sentences with negative tone;

$TONE_k = 0$ sentences with neutral tone.

Then, for Microsoft fillings in quarter j, I defined its overall filling's tone as the average tone of all the predicted sentences in that fillings as follow:

$$TONE_j = \frac{1}{i}\sum TONE_{j,i}$$

where i is the total number of future prediction statements. $TONE_{ij}$ is a variable that is 1, -1 or 0, in which 1 being completely positive, -1 being completely negative and 0 being neutral.

Chapter 4

ANALYSIS OF THE RESULT

4.1 Result

Table 4-1 is the source data which was generated from Perl. The tones are listed according to categories. There are profit tone, liquidity tone and other tone. Since some categories have a small proportion, I combined some of them. Three categories are left which included profit, liquidity and other. The profit_pre, liquidity_pre and other_pre are the percentages they devote to the content.

Table 4-1 Data Source

	PROFIT_TONE	LIQUIDITY_TONE	OTHER_TONE	PROFIT_PRE	LIQUIDITY_PRE	OTHER_PRE
20090122	-0.364	0	-0.2273	0.58333	0.08333	0.32346
20090423	-0.3243	0	-0.16216	0.54286	0.1	0.0007
20091023	-0.3571	0	-0.21053	0.50909	0.14545	0.34545
20100128	-0.4194	0.09091	-0.1	0.508197	0.180328	0.32787
20100422	-0.3871	0	-0.22727	0.5	0.14516	0.35484
20101028	-0.30303	0	-0.08333	0.6111	0.16667	0.2222
10110127	-0.25	0	-0.09091	0.64286	0.16071	0.19643
20110428	-0.27273	0.09091	-0.15789	0.59459	0.14865	0.25676
20111020	-0.26667	-0.076923	-0.05	0.5769	0.16667	0.25641
20120119	-0.26531	-0.06667	-0.130434	0.56977	0.17442	0.26744
20120419	-0.2353	-0.07692	-0.136364	0.6	0.152941	0.25882
20121018	-0.24324	-0.1111	0.045455	0.544118	0.13235	0.32353

4.2 Analysis

To analyze the result, SPSS software was used. Table 4-2 shows the statistics information for FPS tones for the final sample. We can see from the table, the future prediction statement of Microsoft is negative, as indicated by a mean of TONE of -0,4474.

Table 4-2 Descriptive Statistics of each category's tone

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
PROFIT_TONE	12	.184100	-.419400	-.235300	-.30734833	.017862395	.061877151	.004
LIQUIDITY_TONE	12	.202010	-.111100	.090910	-.01248275	.018107708	.062726941	.004
OTHER_TONE	12	.272755	-.227300	.045455	-.12756108	.022860881	.079192415	.006
TONE	12	.305485	-.614370	-.308885	-.44739217	.029374318	.101755622	.010
PROFIT_PRE	12	.142860	.500000	.642860	.56523458	.012995161	.045016558	.002
LIQUIDITY_PRE	12	.096995	.083333	.180328	.14639017	.008412133	.029140485	.001
OTHER_PRE	12	.189271	0196429	.385700	.29684325	.017705072	.061332169	.004
Valid_N(listwise)	12							

In table 4-2, the last three rows are namely the overall proportion of revenue, cost, profit, and operations in the content as predicted by the algorithm, the overall proportion of liquidity, investing, and financing in the content as predicted by the algorithm and all the rest categories proportion in the content as predicted by the algorithm. The mean of PROFIT_PCT, LIQUIDITY_PCT and OTHER_PCT are 56.52%, 14.64% and 29.68%.

Table 4-3 T-test of three category's tones

	Test Value=0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
PROFIT_TONE	-17.206	11	.000	-.30734833	-.34666320	-.26803347
LIQUIDITY_TONE	-.689	11	.505	-.01248275	-.05233755	.02737205
OTHER_TONE	-5.580	11	.000	-.12756108	-.17787754	-.07724462
TONE	-15.231	11	.000	-.44739217	-.51204460	-.38273973

In the t-test, the mean of overall TONE is significantly different from 0. The TONE included three different components: PROFIT_TONE (the average tone of sentence in categories: sales, income, cost, profit and so on), LIQUIDITY_TONE (the average tone of sentence in categories: capital, liquidity, interest coverage, financing and so on), OTHER_TONE (the average tone of sentence in categories: employees, regulations, accounting and other).

The tone of revenue, cost, profit, operations in total are significant. The tone of employees, regulations, accounting and other are also significant. But, the tone of liquidity, investing and financing are not significant.

Table 4-4 Correlation of tone and content

Correlations								
		PROFIT_TONE	LIQUIDITY_TONE	OTHER_TONE	PROFIT_PRE	LIQUIDITY_PRE	OTHER_PRE	TONE
PROFIT_TONE	Pearson Correlation	1	-.600*	.556	.693*	.165	-.655*	.671
	Sig. (2-tailed)		.039	.061	.013	.609	.021	.017
	N	12	12	12	12	12	12	12
LIQUIDITY_TONE	Pearson Correlation	-.600*	1	-.450	-.130	.046	.099	-.099
	Sig. (2-tailed)	.039		.142	.687	.888	.759	.760
	N	12	12	12	12	12	12	12
OTHER_TONE	Pearson Correlation	.556	-.450	1	.242	.370	-.428	.839*
	Sig. (2-tailed)	.061	.142		.449	.237	.165	.001
	N	12	12	12	12	12	12	12
PROFIT_PRE	Pearson Correlation	.693*	-.130	.242	1	.076	-.787**	.529
	Sig. (2-tailed)	.013	.687	.449		.815	.002	.077
	N	12	12	12	12	12	12	12
LIQUIDITY_PRE	Pearson Correlation	.165	.046	.370	.076	1	-.656*	.416
	Sig. (2-tailed)	.609	.888	.237	.815		.021	.179
	N	12	12	12	12	12	12	12
OTHER_PRE	Pearson Correlation	-.655*	.099	-.428	-.787**	-.656*	1	-.670*
	Sig. (2-tailed)	.021	.759	.165	.002	.021		.017
	N	12	12	12	12	12	12	12
TONE	Pearson Correlation	.671	-.099	.839*	.529	.416	-.670*	1
	Sig. (2-tailed)	.017	.760	.001	.077	.179	.017	
	N	12	12	12	12	12	12	12

In Table 4-4, the correlations for tone and content are listed. There is a significant negative correlation between tone and the proportion that sentences about employees, regulations, accounting and other devote to the content. It indicates when there are more discussions of employees, regulations, accounting and other the average tone is more negative.

Also, the average tone of liquidity, investing, financing are negatively related to the proportion that sentences about revenue, cost, profit, operations sentences devote to the content. It indicates when there are more discussions of revenue, cost, profit and operations the average tone of liquidity, investing, financing is more negative.

Moreover, the correlations between the average tone of revenue, cost, profit, operations and the proportions that sentences about employees, regulations, accounting, and other devote to the content are also negative. This means when there are more discussion of employees, regulations, accounting and other the average tone of revenue, cost, profit and operation is more negative.

Chapter 5

CONCLUSIONS AND FUTURE RSEARCH

5.1 Conclusions

In this thesis, I examined the implication of the FPS in the Management Discussion and Analysis section of 10-Q fillings for future performance. I used Naïve Bayesian machine learning algorithm to classify all the sentences. The result showed when there are more discussions of employees, regulations, accounting and other the average tone is more negative. The result also told us when there are more discussion of employees, regulations, accounting and other the average tone of revenue, cost, profit, operations is more negative. Moreover, The tone of revenue, cost, profit, operations in total are significant which is significant different from 0. The tone of employees, regulations, accounting and other are also significant which is significant different from 0. But, the tone of liquidity, investing and financing are not significant which is not significant different from 0.

5.2 Future Research

We can apply dictionary approach on the future prediction statements in Management Discussion and Analysis section data mining and do some comparisons between the results of these two approaches. Then, we can find a better method and make it clear whether the statistic approach is better or not. Also, we may find in some particular situations rule-based approach may be better than statistical approach.

Moreover, there is a lot of scope for future research on this topic. We can use other statistic approach to process the text analysis. We can do linear regression to have statistic prediction and combine it in data mining or we can use nonlinear regression to achieve the same goal.

We can also apply Naïve Bayesian machine learning algorithm to data mining with different data source, such as financial risk or web user demand explore.

BIBLIOGRAPHY

[1] Text Mining: Finding Nuggets in Mountains of Textual Data, J. Dorre, P. Gerstl, and R. Seiffert, *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, California, August 15-18, 1999, 398-401.

[2] Bayesian Networks for Data Mining, David Heckerman, *Data Mining and Knowledge Discovery*, Volume 1, Issue 1, 1997. Spring 2015 Presenter: Tucker Tirey.

[3] Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification (2001). Eui-Hong (Sam) Han, George Karypis, and Vipin Kumar, *PAKDD*, 2001.

[4] The Information Contentof Forward-Looking Statements in Corporate Filings—A Na¨ıve Bayesian Machine Learning Approach, Feng Li, *Journal of Accounting Research* Vol. 48 No. 5 December 2010

[5] Document Categorization and Query Generation on the World Wide Web Using WebACE (1999). Daniel Boley, Maria Gini, Robert Gross, Eui-Hong (Sam) Han, Kyle Hastings, George Karypis, Vipin Kumar, Bamshad Mobasher, and Jerome Moore, *AI Review*, Vol. 13, No. 5-6, 1999.

[6] Data Mining: Statistics and More, D. Hand, *American Statistician*, 52(2):112-118.

[7] A Framework for Analyzing Categorical Data (2009). Varun Chandola, Shyam Boriah, and Vipin Kumar, *Proceedings of SIAM Data Mining Conference*, April 2009, Sparks, NV

[8] Summarization - Compressing Data into an Informative Representation (2005). Varun Chandola and Vipin Kumar. *Proceedings of 5th International Conference on Data Mining (ICDM)*, 2005

[9] Data Mining, G. Weiss and B. Davison, *Handbook of Technology Management*, John Wiley and Sons, expected 2010.

[10] HICAP:Hierarchial Clustering with Pattern Preservation (2004). Hui Xiong, Michael Steinbach, Pang-Ning Tan, and Vipin Kumar, *In Proc. of the Fourth SIAM International Conf. on Data Mining (SDM'04)*, Florida, USA, 2004.

[11] Finding Topics in Collections of Documents: A Shared Nearest Neighbor Approach (2003). Levent Ertoz, Michael Steinbach, and Vipin Kumar, *Clustering and Information Retrieval, forthcoming 2003*, Kluwer Academic Publishers.

[12] From Data Mining to Knowledge Discovery in Databases, U. Fayyad, G. Piatesky-Shapiro & P. Smyth, *AI Magazine*, 17(3):37-54, Fall 1996.

[13] A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, G. Batista, R. Prati, and M. Monard, *SIGKDD Explorations*, 6(1):20-29, 2004.

[14] Challenging Problems in Data Mining Research, Q. Yiang and X. Wu, I*nternational Journal of Information Technology & Decision Making*, Vol. 5, No. 4, 2006, 597-604

[15] Learning from Little: Comparison of Classifiers Given Little of Classifiers given Little Training, G. Forman and I. Cohen, *in 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 161-172, 2004.

[16] Top 10 Algorithms in Data Mining, X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. motoda, G.J. MClachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, *Knowledge Information System* (2008) 141-37

[17] Induction of Decision Trees, R. Quinlan, *Machine Learning*, 1(1):81-106, 1986

[18] Generative Oversampling for Mining Imbalanced Datasets, A. Liu, J. Ghosh, and C. Martin, *Third International Conference on Data Mining (DMIN-07)*, 66-72.

[19] The Pagerank Citation Ranking: Bringing Order to the Web, L. Page, S. Brin, R. Motwani, T. Winograd, *Technical Report*, Stanford University, 1999

[20] Expert Agreement and Content Based Reranking in a Meta Search Engine Environment using Mearf (2002). B. Uygar Oztekin, George Karypis, and Vipin Kumar, *WWW,* 2002.

[21] Mining Association Patterns in Web Usage Data (2002). Pang-Ning Tan, and Vipin Kumar, *International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet*. 2002

[22] A High Performance Two Dimensional Scalable Parallel Algorithm for Solving Sparse Triangular Systems (1997). Mahesh V. Joshi, Anshul Gupta, George Karypis, and Vipin Kumar, *4th International Conference on High Performance Computing*, (HiPC'97).

[23] Mining Frequent Patterns without Candidate Generation, J. Han, J. Pei, and Y. Yin, Proc. 2000 ACM-SIGMOD Int. *Conf. on Management of Data (SIGMOD'00)*, Dallas, TX, May 2000.

[24] Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules, S.D. Lee, David Cheung and Ben Kao, *Data Mining and Knowledge Discovery*, Volume 2, Issue 3, 1998, 233-262.

[25] Improving Generalization with Active Learning, D Cohn, L. Atlas, and R. Ladner, *Machine Learning* 15(2), 201-221, May 1994.

[26] Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction, G. Weiss and F. Provost, *Journal of Artificial Intelligence Research*, 19:315-354, 2003.

[27] Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection, P. Chan and S. Stolfo, *KDD* 1998.

APPENDIX A

TEXT CLASSIFICATION

The rule of how to select future prediction statements is that sentences with word “would”, “will”, “can”, “could”, “should”, “expect”, “may”, “might”, “believe”, “anticipate”, “plan”, “seek”, “project” “hope”, “intend”, “target”, “aim”, “forecast”, “objective”, “goal” are all defined as future prediction statements.

Then, I exclude sentences that consist of all capital letters. After that I stem words which means to reduce word to their common root form. For example, explore, explores, explored are all become explor. In this way we can reduce the miscellaneous words.

After that I do a stopword process which aims to suppress a collection of words, like: is, the, and, or etc. Since these words do not tell us any useful information.

APPENDIX B

TRAINING DATA

I classified selected sentences according to their contents and tones.

the repurcha program may suspend discontinu time without notice.

negative, financ, invest

as part resourc manag program announc januari 2009, addit reduct third quarter fiscal year 2009, will reduc employ headcount approxim 3,400 employ june 30, 2010, will elimin merit increa employ fiscal year 2010.

negative, employ

if actual result differ signif estimates, stock-ba compen expen result oper impacted.

negative, cost, oper

the impact propo adjust year current open examin signif impact financ statement no resolv favorably.

negative, regul

we no believ reason possibl total amount unrecogn tax benefit will signif increa decrea within next 12 months, no believ audit will conclud within next 12 months.

negative, profit

other requir addit disclosures, no anticip materi impact financ statement upon adoption.

negative, other

quarter annual revenu may impact deferr revenue.

negative, revenu

chang estim assumpt materi affect determin fair valu goodwil impair report unit.

negative,cost

if market, industry, and/or invest condit deteriorate, may incur futur impairments.

negative,profit

a major invest price price vendor general level 1 level 2 invest vendor either provid quot market price activ market use observ input price without appli signif adjustments.

neutral, invest

these event circumst includ signif chang busi climate, legal factors, oper perform indicators, competition, sale disposit signif portion report unit.

neutral, revenue

foreign currenc exchang rate no hav signif impact revenu quarter.

neutral, revenu

adopt fsp fas 157-4 fsp fas 115-2 fas 124-2 may financ statements.

neutral, regul

dure three month end march 31, 2009, record employ sever charg million expect reduct headcount.

neutral, cost, employ

under new guidanc arrang includ softwar elements, tangibl product softwar compon essenti function tangibl product will no long within scope softwar revenu recognit guidance, software-en product will now subject relev revenu recognit guidance.

neutral, revenu

the repurcha program expir septemb 30, 2013 may suspend discontinu time without notice.

neutral, financ

determin fair valu stock-ba award grant date requir judgment, includ estim expect dividends.

neutral, financ

the tax benefit recogn financ statement posit measur base largest benefit greater 50% likelihood realiz upon ultim settlement.

neutral, profits

addit properti equip will continue, includ new facilities, data centers, comput system research development, sale marketing, support, administr staff.

neutral,invest

chang factor materi impact financ statements.

neutral, other

the follow tabl outlin expect futur recognit unearn revenu decemb 31, 2009:

neutral, revenu

we believ adopt new guidanc will no hav materi impact financ statements.

neutral,other

the tax benefit recogn financ statement posit measur base largest benefit greater 50% likelihood realiz upon ultim settlement.

neutral, profit

addit properti equip will continue, includ new facilities, data centers, comput system research development, sale marketing, support, administr staff.

neutral, invest

determin fair valu stock-ba award grant date requir judgment, includ estim expect dividends.

neutral, financ

in addition, judgment also requir estim amount stock-ba award expect forfeited.

neutral, other

microsoft' window marketplac infrastructur will support new nokia-brand applic store.

neutral, other

the follow tabl outlin expect futur recognit unearn revenu march 31, 2011:

neutral, revenu

a portion revenu may record unearn due undeliv elements.

neutral, revenu

becaus convert debt may wholli partial settl cash, requir separ account liabil equiti compon note manner reflect nonconvert debt borrow rate interest cost recogn subsequ periods.

neutral, regul

our softwar hardwar platform invest can seen product like kinect, windows, window azure, window phone, window server, xbox.

neutral, invest

these server applic can host customer, partner, microsoft cloud.

neutral, other

if cost invest exceed fair value, evaluate, among factors, general market conditions, credit qualiti debt instrument issuers, durat extent fair valu less cost, equiti securities, intent abil hold, plan sell, investment.

neutral, cost

the transact expect complet end 2012, upon satisfact customari condit regulatori approv

neutral, other

for fixed-incom securities, also evalu whether plan sell secur like no that will requir sell secur recovery.

neutral, other

if entiti determin fair valu report unit less carri amount, two-step goodwil impair test no required.

neutral, oper

we invest signif resourc dynamics, exchange, lync, office, offic 365, sharepoint, window live.

neutral, invest

we believ exist cash cash equiv short-term investments, togeth fund generat operations, suffici meet oper requirements, regular quarter dividends, debt.

positive, investing, oper

expen will reduc cut travel expenditures, reduc spend vendor cont staff, reduc market spending, scale back capit

positive, cost

revenu relat window vista no subject similar deferr no signif undeliv elements.

positive, revenu

the guidanc will becom effect us report period begin july1, 2011.

positive, regul

the cap call transact expect reduc potenti dilut earn per share upon conver notes.

positive, revenu

the product servic bring market may develop internally, brought market part partnership

alliance, acquisition.

positive, other

our goal lead industri area long term, expect will translat sustain growth well future.

positive, other

they decid solut will make employ productive, collaborative, satisfied.

positive, employ

revenu relat window vista no subject similar deferr no signif undeliv elements.

positive, revenu

we believ exist cash cash equiv short-term investments, togeth fund generat operations, suffici meet oper requirements, regular quarter dividends, debt.

positive, investing, oper

we invest signif resourc dynamics, exchange, lync, office, offic 365, sharepoint, window live.

neutral, invest

if entiti determin fair valu report unit less carri amount, two-step goodwil impair test no required.

neutral, oper

our softwar hardwar platform invest can seen product like kinect, windows, window azure, window phone, window server, xbox.

neutral, invest

addit properti equip will continue, includ new facilities, data centers, comput system research development, sale marketing, support, administr staff.

neutral,invest

in addition, judgment also requir estim amount stock-ba award expect forfeited.

neutral, other

we no believ reason possibl total amount unrecogn tax benefit will signif increa decrea within next 12 months, no believ audit will conclud within next 12 months.

negative, profit

if actual result differ signif estimates, stock-ba compen expen result oper impacted.

negative, cost, oper

APPENDIX C

PERL CODE

Stem.pl

```
#!/usr/bin/perl

use 5.010;
use Lingua::Stem::Snowball qw/stem/;
use Lingua::StopWords qw( getStopWords );


open IN, "</Users/Agnes/Documents/2012-10-18result.TXT" or die "Cannot open file $!\n";
open OUT1, ">/Users/Agnes/Documents/2012-10-18stem.txt";

while(<IN>)
{

        $_=~s/\bno\b\s+/no\-/g;
        #no,none,nor,not
        $_=~s/\bnone\b\s+/no\-/g;
        $_=~s/\bnor\b\s+/no\-/g;
        $_=~s/\bnot\b\s+/no\-/g;



        my $stopwords = getStopWords('en');
        my @words=$_=~/(\S+)/g;



        @words=grep { !$stopwords->{$_} } @words;
        #print @words;
        my $stemmer = Lingua::Stem::Snowball->new( lang => 'en' );
      $stemmer->stem_in_place( \@words );
      my @stems = stem( 'en', \@words );
 map($_=~s/no\-/no /g,@stems);
      print join " ",@stems;
      print "\n";


       print OUT1 join " ",@stems;
      print    OUT1 "\n";}
```

Stopword.pl

```
#!/usr/bin/perl

use Lingua::StopWords qw( getStopWords );


open IN, "</Users/Agnes/Documents/FLS1.TXT" or die "Cannot open file $!\n";
open OUT1, ">/Users/Agnes/Documents/FLS1.1.txt";

while(<IN>)
{

        $_=~s/\bno\b\s+/no\-/g;
        #no,none,nor,not
        $_=~s/\bnone\b\s+/no\-/g;
        $_=~s/\bnor\b\s+/no\-/g;
        $_=~s/\bnot\b\s+/no\-/g;



        my $stopwords = getStopWords('en');
        my @words=$_=~/(\S+)/g;



        @words=grep { !$stopwords->{$_} } @words;
      print join " ",@words;
      print "\n";
      print OUT1 join " ",@words;
      print    OUT1 "\n";

}
```

Content.pl

```
#!/usr/bin/perl
use warnings;
use Algorithm::NaiveBayes;
local $/ = "\015";

my $liq_file = '/Users/Agnes/Documents/LIQUIDITY.TXT';
my $pro_file = '/Users/Agnes/Documents/PROFIT.txt';
my $oth_file = '/Users/Agnes/Documents/OTHER.txt';

my $categorizer = Algorithm::NaiveBayes->new;
my $fh;

open($fh,"<",$liq_file) or die "Could not open $liq_file: $!";
while (my $sentence = <$fh>) {
        chomp $sentence;
        my @words = split(' ',$sentence);
        my %LIQUIDITY;
        $LIQUIDITY{$_}++ for @words;
        $categorizer->add_instance(
        attributes => \%LIQUIDITY,
        label => 'LIQUIDITY');
}
close($fh);

open($fh,"<",$pro_file) or die "Could not open $pro_file: $!";
while (my $sentence = <$fh>) {
        chomp $sentence;
        my @words = split(' ',$sentence);
        my %PROFIT;
        $PROFIT{$_}++ for @words;
        $categorizer->add_instance(
        attributes => \%PROFIT,
        label => 'PROFIT');
}
close($fh);

open($fh,"<",$oth_file) or die "Could not open $oth_file: $!";
while (my $sentence = <$fh>) {
        chomp $sentence;
        my @words = split(' ',$sentence);
        my %OTHER;
        $OTHER{$_}++ for @words;
        $categorizer->add_instance(
        attributes => \%OTHER,
        label => 'OTHER');
}
close($fh);
```

```
$categorizer->train;

my $sentence_file = '/Users/Agnes/Documents/20121018_0.13043.txt';

open($fh,"<",$sentence_file) or die "Could not open $sentence_file: $!
+";
while (my $sentence = <$fh>) {
	chomp $sentence;
	my @words = split(' ',$sentence);
	my %test;
	$test{$_}++ for @words;
	my $probability = $categorizer->predict(attributes => \%test);
	if ($probability->{LIQUIDITY} > $probability->{PROFIT} and
	$probability->{LIQUIDITY} > $probability->{OTHER}) {
		print "LIQUIDITY: $sentence\n";
	}
	elsif ($probability->{PROFIT} > $probability->{LIQUIDITY} and
	$probability->{PROFIT} > $probability->{OTHER}) {
		print "PROFIT: $sentence\n";
	}
	elsif ($probability->{OTHER} > $probability->{LIQUIDITY} and
	$probability->{OTHER} > $probability->{PROFIT}) {
		print "OTHER: $sentence\n";
	}
}

close ($fh);
```

Computation.pl

```
#!/usr/bin/perl
use warnings;
use Algorithm::NaiveBayes;
local $/ = "\015";

my $pos_file = '/Users/Agnes/Documents/positive.TXT';
my $neg_file = '/Users/Agnes/Documents/negative.txt';
my $neu_file = '/Users/Agnes/Documents/neutral.txt';

my $categorizer = Algorithm::NaiveBayes->new;
my $fh;

open($fh,"<",$pos_file) or die "Could not open $pos_file: $!";
while (my $sentence = <$fh>) {
	chomp $sentence;
	my @words = split(' ',$sentence);
	my %positive;
	$positive{$_}++ for @words;
	$categorizer->add_instance(
	attributes => \%positive,
	label => 'positive');
}
close($fh);

open($fh,"<",$neg_file) or die "Could not open $neg_file: $!";
while (my $sentence = <$fh>) {
	chomp $sentence;
	my @words = split(' ',$sentence);
	my %negative;
	$negative{$_}++ for @words;
	$categorizer->add_instance(
	attributes => \%negative,
	label => 'negative');
}
close($fh);

open($fh,"<",$neu_file) or die "Could not open $neg_file: $!";
while (my $sentence = <$fh>) {
	chomp $sentence;
	my @words = split(' ',$sentence);
	my %neutral;
	$neutral{$_}++ for @words;
	$categorizer->add_instance(
	attributes => \%neutral,
	label => 'neutral');
}
close($fh);
```

```
$categorizer->train;

my $sentence_file = '/Users/Agnes/Documents/2012_10_18stem.txt';

open($fh,"<",$sentence_file) or die "Could not open $sentence_file: $!
+";
while (my $sentence = <$fh>) {
	chomp $sentence;
	my @words = split(' ',$sentence);
	my %test;
	$test{$_}++ for @words;
	my $probability = $categorizer->predict(attributes => \%test);
	if ($probability->{positive} > $probability->{negative} and
	$probability->{positive} > $probability->{neutral}) {
		print "positive: $sentence\n";
	}
	elsif ($probability->{negative} > $probability->{positive} and
	$probability->{negative} > $probability->{neutral}) {
		print "negative: $sentence\n";
	}
	elsif ($probability->{neutral} > $probability->{positive} and
	$probability->{neutral} > $probability->{negative}) {
		print "neutral: $sentence\n";
	}
}

close ($fh);
```

Parse_para.pl

```
#!/usr/bin/perl

use 5.010;


open IN, "</Users/Agnes/Documents/FLS.txt" or die "Cannot open file
$!\n";
open OUT1, ">/Users/Agnes/Documents/sentence.txt";
open OUT2, ">/Users/Agnes/Documents/words.txt";

sub split_to_sentences
{
    my @str = split(/[\.\?\!]/, $_);    #以.或者?或者!来分割段落成句子
    return @str;
}

sub split_to_words
{
    my @str = split(/ /, $_);   #以空格来分割句子成单词
    return @str;
}


sub main
{
    while(<IN>) {
        chomp($_);  #去除最后的换行符
        $_ = lc;    #大写字母转化为小写字母
        s/[,"]//g;  #去除,
        s/ +/ /g;   #将多个连续的空格转为一个空格
        if ($_) {
            my @sentences = split_to_sentences($_);
            foreach (@sentences)
            {
                $_ =~ s/^\s+|\s+$//g;   #去除头尾的空格
                #say $_;
                print OUT1 "$_\n";
                my @words = split_to_words($_);
                foreach (@words)
                {

                    #say $_;
                    print OUT2 "$_\n";

                }
            }
```

```
        }
    }
    close IN;
    close OUT;
}

main
```