GOLD PRICE VOLATILITY PREDICTION BY TEXT MINING IN ECONOMIC INDICATORS NEWS

CHANWIT ONSUMRAN

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(INFORMATION TECHNOLOGY MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2015**

Thesis
entitled

GOLD PRICE VOLATILITY PREDICTION BY TEXT MINING IN ECONOMIC INDICATORS NEWS

……
Mr. Chanwit Onsumran
Candidate

……
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Major advisor

……
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Co-advisor

……
Asst. Prof. Adisorn Leelasantitham,
Ph.D. (Electrical Engineering)
Co-advisor

……
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

……
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Program Director
Master of Science Program in Technology of Information System Management
Faculty of Engineering
Mahidol University

Thesis
entitled

GOLD PRICE VOLATILITY PREDICTION BY TEXT MINING IN ECONOMIC INDICATORS NEWS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Information Technology Management)
on
March 28, 2015

……
Mr. Chanwit Onsumran
Candidate

……
Lect. Taweesak Samanchuen,
Ph.D. (Electrical Engineering)
Chair

……
Lect. Sotarat Thammaboosadee,
Ph.D. (Information Technology)
Member

……
Asst. Prof. Supaporn Kiattisin,
Ph.D. (Electrical and Computer Engineering)
Member

……
Asst. Prof. Adisorn Leelasantitham,
Ph.D. (Electrical Engineering)
Member

……
Asst. Prof. Kairoek Choeychuen,
Ph.D. (Electrical and Computer Engineering)
Member

……
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

……
Lect. Worawit Isarangkul,
M.S. (Technical Management)
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my thesis advisor, Dr. Sotarat Thammaboosadee and co-advisor, Asst. Prof. Supaporn Kiattisin and Asst. Prof. Adisorn Leelasantitham for invaluable help and constant encouragement throughout the course of this research. I am most grateful for his teaching and advice, not only the research methodologies but also many other methodologies in life. I would not have achieved this far and this thesis would not have been completed without all the support that I have always received from his.

The indispensable group was my family and my friends who always stay beside and encourage me when I was under pressure and discouraged. Even they do not know deeply in this topic, but they tend to support me as much as possible. Finally, I most gratefully acknowledge my parents for all their support throughout the period of this research.

Chanwit Onsumran

GOLD PRICE VOLATILITY PREDICTION BY TEXT MINING IN ECONOMIC INDICATORS NEWS

CHANWIT ONSUMRAN 5636623 EGIT/M

M.Sc. (INFORMATION TECHNOLOGY MANAGEMENT)

THESIS ADVISORY COMMITTEE: SOTARAT THAMMABOOSADEE, Ph.D., SUPAPORN KIATTISIN, Ph.D., ADISORN LEELASANTITHAM, Ph.D.

ABSTRACT

This research focuses on the text mining approach of the gold prices volatility prediction model from the text of economic indicators news articles. The model is designed and developed to analyze how the news articles influence gold price volatility. The selected reliable source of news articles is provided by FXStreet, which offers several economic indicators. The data will be used to build text classifiers and news group affecting volatility price of gold. According to the fundamentals of the data mining process, each news article is firstly transformed in to a feature by the TF-IDF method. Then, a comparative experiment is set up to measure the accuracy of the combination of two attributes weighting approaches - which are Support Vector Machine (SVM) and Chi-Squared Statistic - and three classification algorithms - which are the k-Nearest Neighbors, SVM and Naive Bayes. The results show that the SVM classification algorithm, weighted by SVM, is the best among all tests with an accuracy of 87.52%. In the future, it can be developed to improve the classification system; it may be input to other economic news for more factors to increase the efficiency of analyzing consequences that affect the volatility of the gold price.

KEY WORDS: TEXT MINING/ ECONOMIC INDICATORS NEWS/ GOLD PRICE/ VOLATILITY PREDICTION

74 pages

การทำนายความผันผวนของราคาทองคำ โดยวิธีเหมืองข้อความในบทความข่าวชี้วัดทางเศรษฐกิจ
GOLD PRICE VOLATILITY PREDICTION BY TEXT MINING IN ECONOMIC INDICATORS NEWS

ชาญวิทย์ อ้นสำราญ 5636623 EGIT/M

วท.ม. (การจัดการเทคโนโลยีสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : โยฑศ์รัตต ธรรมบุษดี, Ph.D., สุภาภรณ์ เกียรติสิน, Ph.D., อดิศร ลีลาสันติธรรม, Ph.D.

บทคัดย่อ

งานวิจัยนี้มุ่งเน้นไปที่วิธีการทำเหมืองข้อความจากบทความข่าวชี้วัดทางเศรษฐกิจ เพื่อทำนายความผันผวนของราคาทองคำ โดยทำนายความผันผวนจากรูปแบบของข่าวซึ่งจะเป็นข้อความข่าวชี้วัดทางเศรษฐกิจ โดยจะออกแบบและพัฒนาขึ้นเพื่อวิเคราะห์บทความข่าวที่มีอิทธิพลต่อความผันผวนของราคาทองคำ โดยข้อมูลบทความข่าวที่นำมาใช้ในงานวิจัยนี้คือข้อมูลจากเว็บไซต์ FXStreet ซึ่งเป็นแหล่งข้อมูลที่เป็นสากลและมีความน่าเชื่อถือ โดยได้เลือกบทความข่าวทางเศรษฐกิจที่มีความเกี่ยวข้องกับงานวิจัยในครั้งนี้มาวิเคราะห์ ซึ่งข้อมูลจะถูกนำมาใช้ในการสร้างรูปแบบสำหรับการจัดหมวดหมู่ของข้อความ และกลุ่มข่าวที่มีผลต่อความผันผวนของราคาทองคำตามพื้นฐานของกระบวนการทำเหมืองข้อความ โดยใช้การให้ค่าน้ำหนักของคำในเอกสารด้วยวิธี TF/IDF จากนั้นทดลองเปรียบเทียบประเมินผลความถูกต้องด้วยการเลือกใช้ Attributes Weighting ซึ่งประกอบด้วยวิธีการของ Support Vector Machine และวิธี Chi-Squared Statistic โดยใช้เทคนิค Classification Algorithms ทั้งหมด 3 วิธีประกอบด้วย k-Nearest Neighbors, SVM และ Naive Bayes ซึ่งผลการศึกษาพบว่าวิธี SVM โดยการเลือกใช้ Attributes Weighting Support Vector Machine มีความถูกต้องมากที่สุดร้อยละ 87.52 ซึ่งการวิจัยในอนาคตสามารถที่จะพัฒนาและปรับปรุงรูปแบบการจำแนกข้อความโดยอาจจะเป็นการเพิ่มปัจจัยข้อมูลข่าวเศรษฐกิจด้านอื่น ๆ เพิ่มเติมเพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ผลกระทบที่ส่งผลต่อความผันผวนของราคาทองคำต่อไป

74 หน้า

CONTENTS

	Page
ACKNOWLEDGEMENTS	**iii**
ABSTRACT (ENGLISH)	**iv**
ABSTRACT (THAI)	**v**
LIST OF TABLES	**viii**
LIST OF FIGURES	**ix**
CHAPTER I INTRODUCTION	**1**
1.1 Background and significance of problems	1
1.2 Research objectives	2
1.3 Scopes of research	2
1.4 Expected results of research	3
CHAPTER II LITERATURE REVIEW	**4**
2.1 Economic News Categories	4
2.2 Data Mining	7
2.2.1 Data mining techniques	9
2.2.2 Data mining process (CRISP-DM)	10
2.3 Text Mining	12
2.4 Vector Space for Term Weighting	12
2.5 Related Algorithms for research	14
2.6 Relevant Research of Text Mining in Economic Domain	16
CHAPTER III RESEARCH METHODOLOGY	**18**
3.1 Processes of Research Methodology	18
3.2 Economic news indicators with articles	20
3.3 Model classification and validation	20
3.4 Research Schedule	22

CONTENTS (cont.)

	Page
CHAPTER IV RESULTS ANALYSIS AND DISCUSSION	**23**
4.1 Gold price matching with economic news indicators	23
4.2 Comparative results of nine experimental schemes	24
4.3 Attribute weighting by SVM and Chi Squared Statistic	25
4.4 Feature selection	26
4.5 Word list data	28
4.6 Result of 2014's unseen data sets with economic news indicators	29
4.7 Result of experimental schemes detail	30
CHAPTER V CONCLUSION AND FUTURE WORKS	**36**
5.1 Conclusion	36
5.2 Future works	36
REFERENCES	**38**
APPENDICES	**40**
Appendix A 2013 News historical data for generating model	41
Appendix B Economic news indicators of 2014's unseen data	54
Appendix C Gold prices historical data	57
Appendix D Technical paper of 2015 International Conference on Information Technology	66
BIOGRAPHY	**74**

LIST OF TABLES

Table		Page
3.1	Research Timeline	22
4.1	Example of gold price matching with economic news indicators	24
4.2	Comparative results	25
4.3	Feature selection with attribute weights by SVM	28
4.4	Feature selection with attribute weights by Chi Squared Statistic	28
4.5	Example of word list	29
4.6	Example of unseen data sets (2014 Economic news indicators)	29
4.7	Classification Results of SVM Classification with SVM Weighting	31
4.8	Classification Results of SVM Classification with Chi Squared Statistic Weighting	31
4.9	Classification Results of SVM Classification	32
4.10	Classification Results of k-NN Classification with SVM Weighting	32
4.11	Classification Results of k-NN Classification with Chi Squared Statistic Weighting	33
4.12	Classification Results of k-NN Classification	33
4.13	Classification Results of Naïve Bayes Classification with SVM Weighting	34
4.14	Classification Results of Naïve Bayes Classification with Chi Squared Statistic Weighting	34
4.15	Classification Results of Naïve Bayes	35
A :	Economic news indicators of 2013	41
B :	Economic news indicators of 2014	54
C :	Historical data of gold price during Jan 2013 - Dec 2013	57

LIST OF FIGURES

Figure		Page
2.1	The architecture of data mining system	8
2.2	The CRISP-DM process	10
2.3	Simple Input - Output model for text mining	12
2.4	Example of K-NN	14
2.5	Two dimensions SVM	15
3.1	Methodology overview	18
4.1	Data set summarization	23
4.2	Results of nine experiment schemes	24
4.3	Attribute weight results by SVM	25
4.4	Attribute weight results by Chi Squared Statistic	26
4.5	Selective results of attribute weights by SVM	27
4.6	Selective results of attribute weights by Chi Squared Statistic	27
4.7	Comparative results between training data set and unseen data set	30

CHAPTER I
INTRODUCTION

1.1 Background and significance of problems

Gold investment is now popular than ever. Hence, gold prices fluctuation is compatible with changing volume of both investment and speculation. The major factor for determining the gold price is the USD currency. If the other factors are stable, the gold prices will increase when USD is depreciated. Therefore, gold is used to hedge against the USD's depreciation, the declining value of USD. The central banks of many countries having USD reserves need to spread their risks by assets investments such as gold. This occurrence pushes up the gold price. With another major factor of the USD inflation rate, the gold price will increase, when inflation rate is higher. Gold price often increases during time of international political tension, when the world monetary system becomes unstable. Presently, demands of gold assets are declining as well as gold prices are dropping during the time periods of crisis. Additionally, demand and supply in the market are significant. With the stable factors, the gold prices will increase, when demand of gold is higher than the supply in the market [1].

Time series analysis is the fact that the data point's measurement taken from over time interval including the trend and seasonal changes. Time series analysis consists of a series of data analysis methods in time to extract meaningful statistics and other characteristics of data. Therefore, it is not suitable for research analysis using the text mining.

Economic news articles are also the indicators for gold price volatility apart of another factor as stated. Anyway, the economic news indicators are represented in text format. Hence, the text mining technique is appropriate for analysis in this research. The text mining methodology has been very popular, since more than 90% of the volume of data on the internet is unstructured. The large amounts of useful

information are often hidden such as: news articles, economic changes, variances of the circumstances, and events occurred at different times.

This research is focused on the relationships between the news articles and gold prices (gold spot). Gold spot price refers to the gold price for immediate delivery. Transactions for gold are mostly always priced using the spot price as a basis. The transactions of gold spot have trading almost 24 hours. It is actively taking orders for gold transactions [2]. It is designed and developed for methods to analyze how the news articles influence gold price volatility. News article by FXStreet is leader source for dependable news and real-time Forex analysis. FXStreet offers the real-time exchange rates and the economic calendar about Markit Manufacturing PMI, Bill Auction, Building Permits, ISM Manufacturing Index, Gross Domestic Product Annualized, Nonfarm Productivity, and EIA Natural Gas Storage change. This data will be used to train the text classifiers and news group affecting volatility price of gold.

1.2 Research objectives

- To match the news group and predict the gold price volatility of economic news indicators by text mining techniques.
- To analyze the suitable weighting of feature selection for level classification.
- To build the text data classifier for economic news indicators.

1.3 Scopes of research

- Economic indicators for news articles are referenced by FXStreet data [3], during January 2013 - December 2013. This can be divided into three groups, given as: low volatility, moderate volatility, and high volatility.
- In our experiment, historical data of Kitco.com's gold price (gold spot) is as the reference taken from Chang et al [4], during January 2013 - December 2013.

1.4 Expected results of research

- To get the suitable weighting of feature selection for prediction of volatility classification.
- To get the classifier model for the prediction of gold price volatility by economic news indicators.

The next chapter will be described the categories of economic news and relevant research of text mining. Data mining process will focus on the Cross Industry Standard Process for Data Mining (CRISP-DM). Furthermore, it also explains the concept of text mining, vector space for Term Frequency - Inverse Document Frequency (TF-IDF), and related algorithms for this research.

CHAPTER II
LITERATURE REVIEW

This chapter describes the categories of economic news and data mining process based on the Cross Industry Standard Process for Data Mining (CRISP-DM). The chapter also explains the concept of text mining, vector space of weighting term (TF-IDF), related algorithms, and relevant research of text mining in economic domain.

2.1 Economic News Categories

Economic news consists of the following 24 categories [5], given as:

1) Chicago Purchasing Managers Index (CPMI)

CPMI, an indicator of business trend and relations with ISM manufacturing index, widely used to indicate the overall economic situation in the United States.

2) Dallas Fed Manufacturing Business Index

The company data, requested by the Central bank of Dallas as well as the swing values of employment, export orders, prices and other indicators, still does not change from the previous month to explore the answer as the index for each individual identifier. Index is calculated by subtracting the percentage of respondents report decreased from the increase of the report.

3) EIA Natural Gas Storage change

The data estimations, delivered by computation, use both monthly survey data of energy and weekly survey data of American gas.

4) Durable Goods Orders (DGO)

DGO is issued by US demographic surveys, the cost of the purchase orders have been received by manufacturers for durable goods. It means that the product plan will be the end of year or more with the exception of the transport sector as a durable

product often involves a large investment with sensitivity of the US economy situation.

5) Month Bill Auction

Month Bill Auction average auction values are given by the United States Department of the Treasury. Treasury bills are short-term securities to maturity in one year or less. Yield values refer to the yield that investors can get from holding a bond until maturity. Investors examine the earnings fluctuations, and compare the average rate at auction for the previous auction rate security as well as an indicator of public sector debt situation.

6) GDP Price Index (GDP)

GDP is a measurement of the global economic movements. The GDP numbers change is implied the change of the economic growth rate tied by the inflation rate. The increase in overview GDP would make up the money by appreciation.

7) Philadelphia Fed Manufacturing Survey

This survey serves as an indicator of trends in production associated with the production index of Institute for Supply Management and the index of industrial production.

8) Building Permits

Building Permits, the number of licenses for a new construction project, is given by Department of Commerce. It represents the movement of the company's investment (economic development), it has a tendency to cause volatility in the USD.

9) Manufacturing Purchasing Managers Index (PMI)

PMI index is an indicator important for the overall economic conditions and business in the US.

10) Producer Price Index ex Food & energy

Given by the Department of labor, it is to measure the change of the average price in the main market of America by the in production all States.

11) Business inventories

Given by US Census Bureau, it is to measure the monthly changes of inventories in manufacturers including retailers and wholesalers.

12) Year Note Auction

It is the Yield record as investors will receive a return on bonds held to maturity. Investors monitor the output, and compare the average rate at auction.

13) Gross Domestic Product Annualized

GDP Annualized is that the value money of each item services and production structures of the country with the given period. It is used as an initial measurement of market activity, because it indicates that the country's economy beat growth or decline.

14) ISM New York Index

ISM New York Index is survey of the financial group. Survey results collected are the distributed index, which is calculated by the percentage of the respondents.

15) Richmond Fed Manufacturing Index

This survey includes the data on new orders, deliveries, order backlogs inventories and the Federal Reserve banks. It is operated by Richmond property to provide the information of the activity in the manufacturing sector.

16) Producer Price Index (PPI)

PPI is the average changes prices in the main markets of the United States by the commodity producers in all states of processing. Changes in the PPI are widely operating as an indicator of inflation commodities.

17) NY Empire State Manufacturing Index

This survey is conditions for New York manufacturers manage by the Federal Reserve Bank of New York business.

18) Nonfarm Productivity

Nonfarm Productivity, the output per hour of labor worked influencing on GDP, indicates the overall business in the US.

19) ISM Non-Manufacturing Index

This index shows business conditions in the US non-manufacturing sector. The ISM Non-Manufacturing index includes purchases of food, energy, and import purchases excluding crude oil.

20) Capacity Utilization

The capacity utilization, the percentage of the US production capacity, is actually used over the short-time period. It shows the overall growth and demand in the US economy.

21) Construction Spending

Construction spending is an indicator to measure the amount of spending in US of all types of construction elements. It is that housing construction will be useful for predicting the future national new home sales and mortgage origination volume.

22) American Petroleum Institute Monthly Report (API)

API is the analysis of recent developments for the production of primary products imports of oil refinery operations and inventories.

23) Housing Starts

Housing starts, the indicator of the survey of each house and each apartment, will only be counted as one of the early residential number including all agencies and private ownership. It shows the movement of the US housing market.

24) Industrial Production

Industrial Production shows the volume of US industry including factories and manufacturing more likely to be regarded as inflation expectations and interest rates.

2.2 Data Mining

Data mining [6] is a process working with the large amounts of data to find the patterns and the relationships of hidden data sets. Recently, data mining has been applied in various types of business to help the executive decision in science and medicine as well as in the economics and social aspects.

Data mining is to compare the evolution of storage with the interpretation of data. The existing simple storage in database can retrieve information to apply the data mining in order to discover the hidden knowledge in data.

Figure 2.1 is shown the data mining process containing the sub-workflows for transforming data into knowledge. The procedure is given as:

- Data Cleaning: screening out irrelevant information;
- Data Integration: combining multiple data sources into a data set;
- Data Selection: retrieving information from sources that are recorded for analysis;
- Data Transformation: converting the raw data to the suitable data for utilization;
- Modeling: searching for the benefit patterns from the existing data;
- Evaluation: evaluating forms of data mining;
- Knowledge Representation: knowledge discovery using the presented techniques for understanding.

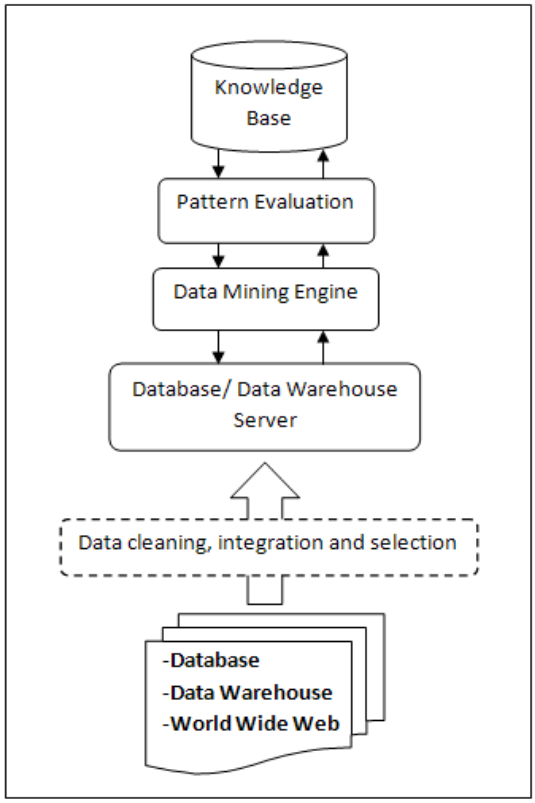

Figure 2.1 The architecture of data mining system.

The evolutions of data mining are given as:

- In 1960, Data Collection was the appropriate data storage with the reliable equipment in order to protect the data loss.

- In 1980, Data Access was to generate the relationships of data for the analysis and decision with quality.
- In 1990, Data Warehouse & Decision Support was to gather the stored data in a huge database covering all of the organization to support the decision.
- In 2000, Data Mining was to create the models with statistical relationships of the data analysis taken from database.

2.2.1 Data mining techniques

Statistical Analysis is method to classify the different types of variable on analysis. Forms of the knowledge are useful for system improvement. It enhances the security of the system, and it is easy to update the information and decision support.

Association Rule is to find the relationship information with the association among sets. For the significance of the measurement, it uses the values of support and confidence. The support value is a percentage of the operation of the rule for accuracy, whereas the confidence value is a number of cases where the rule is correct in relation. To find the association rules, the widely used algorithm is called Apriori algorithm.

Clustering is a grouping of data which looks similar but not identical. The classification of data analysis is based on a template. It is the main task of data mining research and general techniques in statistical data analysis used in many fields including pattern recognition, machine learning, information retrieval, bioinformatics, and image analysis.

Classification purposes can be used as templates for the prediction. This template is created from a set of data analysis training (Training data). There are several kinds of classifications, given as: Classification Rules, Decision Trees, Neural Networks, and so on.

Sequential Pattern mining is a topic of data mining involved for finding the related forms relevant statistics. It usually depends on the situation that the continuity and time series are related closely. It is a technique to search the event sequence. The relationship between transactional data and time are involved.

2.2.2 Data mining process (CRISP-DM)

Cross Industry Standard Process for Data Mining model (CRISP-DM) [7] consists of six phases intended as a cyclical process, as shown in Figure 2.2.

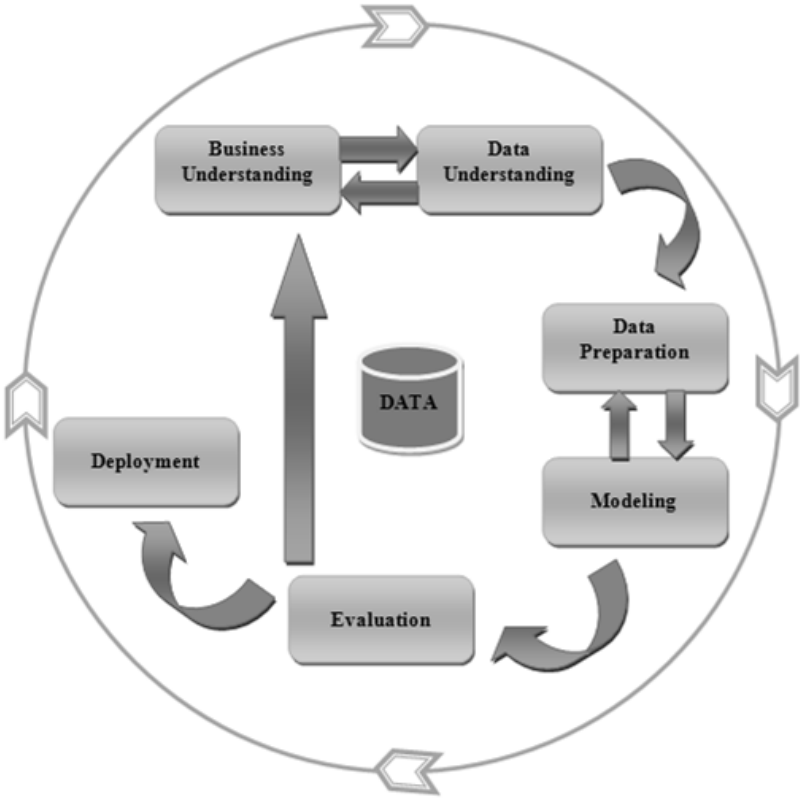

Figure 2.2 The CRISP-DM process.

1) Business Understanding

The objectives of the project needs from business perspective should be understood, and then transform this knowledge into problem definition and data mining basic plan designed to achieve the objectives.

2) Data Understanding

Data is the most important factor and indispensable in data mining process. In this step, it has to collect the relevant information to use in the analysis process with data mining techniques. Data collection should consider whether the information has come from reliable source of accurate information that has detailed enough to identify data quality problem and discover interesting subset.

3) Data Preparation

Data preparation also takes a long time, because it covers all activities needed to make the final data set. The preparation can be divided into three phases, given as:

- Data selection: It refers to the process of determining the types of appropriate information and sources, as well as the appropriate tools to collect data;

- Data cleaning: The solution to data cleaning by cross check with the data set that passed the examination. There is also increasing the efficiency of data. To practice the clean data, data would be more perfect by adding the relevant information.

- Data transformation: It converts a series of the data from the sources of information systems into the destination data system's data format.

4) Modeling

In this step, various modeling technique are chosen to be applied, and their parameters are calibrated to the best values. There are many kinds of problems in data mining technique is usually happened, and some specific techniques are needed the different form of data to be input.

5) Evaluation

In this phase, before the final deployment, this is important to be more thorough evaluation model and review process to build a model to ensure the corrected data. Business purpose is important to check whether there is a problem to find some important business issue that has not been considered adequately.

6) Deployment

Creation of the model is generally not the end of the project. Depending on the requirements, the applications can be as simple as creating a report or as complex as the reproduced data mining process. It is important to make customer understood information.

2.3 Text Mining

Text mining [8], called discovery in databases document, is a technique to search for patterns from a large amount of text or process unstructured (textual) information. It uses the algorithm from the department of pattern recognition and statistics machine learning. Text mining is the process performed on a message to find the guidelines and hidden relationships behind the text. For example, model for text mining as shown in Figure 2.3.

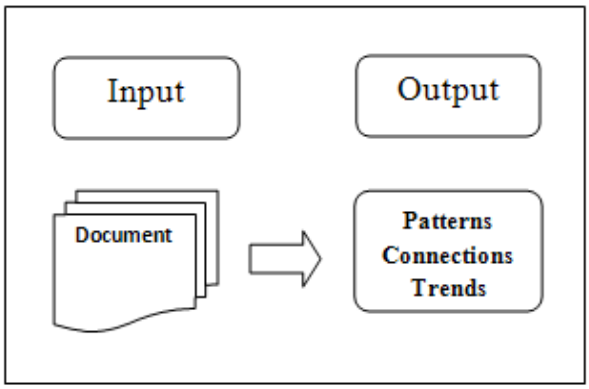

Figure 2.3 Simple Input - Output model for text mining.

2.4 Vector Space for Term Weighting

Term Frequency - Inverse Document Frequency (TF-IDF) [9] is a product of both statistical, inverse document frequency and frequency range. In the case of term frequency *tf*(*t*,*d*), the simplest choice is to use the raw frequency of a term in a document. The number of times *t* occurs in document *d,* where frequency of *t* is given by *f*(*t*,*d*); *tf* scheme is *tf*(*t*,*d*) = *f*(*t*,*d*); frequencies is *tf*(*t*,*d*) = 1 if *t* occurs in *d* and 0 otherwise; logarithmically scaled frequency: *tf*(*t*,*d*) = 1 + log *f*(*t*,*d*), or zero if *f*(*t*,*d*) is zero. It is possible to increase the frequency as well as other to prevent prejudice raw materials. For example, the frequency divided by the frequency of any of the words in most raw materials is shown in Eq. (2.1).

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{\max\{f(w,d) : w \in d\}}. \quad (2.1)$$

The inverse document frequency is a measure of how much data, whether it is a word that is common or rare in the entire document. It is algorithmically scale section of the document that contains the word get by dividing the total number of documents from a number of documents that contain the word. Then, the algorithm of the quotient is shown in Eq. (2.2).

$$idf(t, D) = \log \frac{N}{|\{d \in D: t \in d\}|}. \quad (2.2)$$

Where,

N: The total number of documents in the archive;

$|\{d \in D: t \in d\}|$: Number of documents where the term t appears.

Based on mathematical functions, a factor has to be multiplied by the operation, and TF-IDF is calculated as shown in Eq. (2.3).

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D). \quad (2.3)$$

The high weight in TF-IDF is up to the high frequency and low frequency of the term in the whole collection of all documents. The weight tends to filter words frequently. This is because the proportion of IDF inside the lock function is always greater than or equal to the value of the IDF 1 (and TF-IDF). Note: IDF 1 is also greater than or equal to 0. As a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the *idf* and *tf-idf* closer to 0.

2.5 Related Algorithms for research

1) K Nearest Neighbor (K-NN)

K-NN [10] is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The K-NN algorithm is among the simplest of all machine learning algorithms. With special handling step-by-step how to K-NN, a summary of procedure is given as follows:

- Determine the size of *K*.
- Calculate the distance of information to consider the sample data.
- Sort the pitch and consider the set of information that must be taken into consideration in accordance with point *K* of the set.
- Consider the information of a number of *K* series, and be noticed that the group (class) points to consider as many as possible.
- Define the class to the point that considers.

K - Nearest Neighbor is algorithm for data classification by the concept of data grouping based on the information that is mostly close to the value of the information. If classification uses *k* groups, it can be determined by Euclidean distance. It is called the K-NN (K Nearest Neighbor), as shown in Figure 2.4.

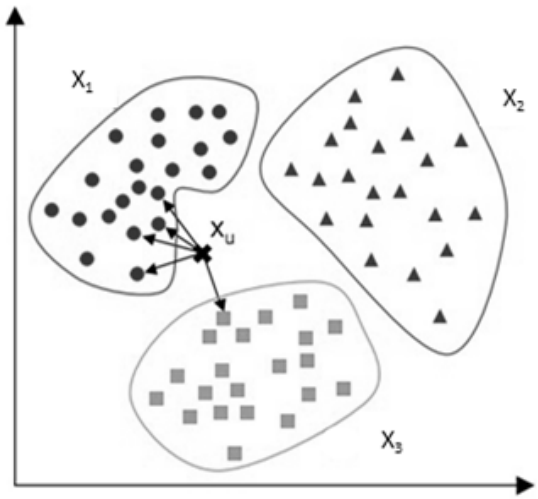

Figure 2.4 Example of K-NN.

2) Support Vector Machines (SVM)

SVM [11] is a concept in statistics and computer science for series of how to learn under the supervision of the relevant data analysis and recognition model used for classification and regression analysis standard. SVM uses a set of input data, and predicts each input, which two possible classes comprise the input, making the SVM a non-probabilistic binary linear classifier. SVM is an algorithm that can be used to solve the problem of data analysis and classification. In brief, its principle is to create a separate set of data that is entered into the system taught to learn by focusing on the best dividing line to distinguish data, as shown in Figure 2.5.

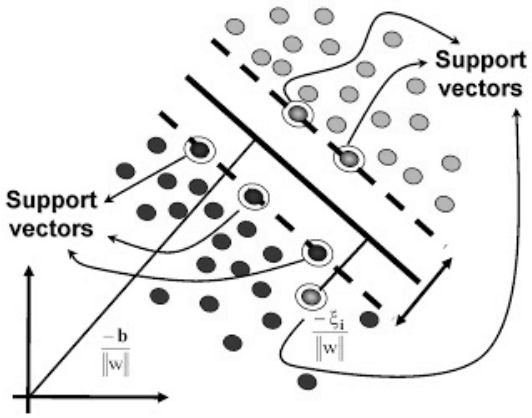

Figure 2.5 Two dimensions SVM.

3) Naive Bayes

Naive Bayes [12] is a type of screening model based on probability, which is based on Bayes' Theorem and assumptions. The occurrence of various events is independent. For Naive Bayes classifier, the advantage is that it requires a small amount of training data to estimate the parameters needed for the classification of independent variables to consider only the variance of each class. For some types of probability model, Naive Bayes classifier can be trained more effective in learning setting under the supervision. In practice, with a large number of parameter estimation for the Naive Bayes method of maximum likelihood, anyone could work with Naive Baye model without accepting Bayesian probability or using any Bayesian methods. Bayes' theorem can be written, as shown in Eq. (2.4).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (2.4)$$

Where, A and B are events;

$P(A)$ and $P(B)$ are the probabilities of A and B independent of each other;

$P(A|B)$ is a conditional probability is the probability of A given that B is true;

$P(B|A)$ is the probability of B given that A is true.

2.6 Relevant Research of Text Mining in Economic Domain

Rivera et al. [13] proposed text mining framework for advancing sustainability indicators. Typical pre-processing methods were used to convert the unstructured textual data to a word-document matrix including the tasks of tokenization and elimination of stop words. Two main techniques used to complete this task consist of: 1) the creation of frequent concept sets and association rules, and 2) Part-of-Speech (POS) tagging and a gazetteer for geo-referencing of news articles. The proposed techniques implemented in the system could highly provide the directional-accuracy result which is up to 88 %. However, the research process still has the delay because of no news classification process before processing.

Nassirtoussi et al. [14] proposed the text mining of news-headlines based on multi-layer dimension reduction algorithm to forecast the intraday FOREX market. The proposed system proved that the proposed assumption was absolutely correct. Therefore, results of this experimentation can completely demonstrate a predictive relationship between this specific market-type and the textual data of news. With the multi-layer algorithm, the proposed techniques implemented on the system could highly provide the directional-accuracy result of 83.33%. However, the weakness of research was the non-existence of selection feature causing the malfunctions of creating prototypes, the important part of analysis.

Hagenau et al. [15] proposed the stock price prediction based on financial news using context-capturing features. It was shown that the combinations of extraction methods, advance features, and feature selection suggestions could improve the accuracy of classification and sentiment analysis. The proposed techniques

implemented on the system could highly provide the directional-accuracy result of 76%. However, the data set for creating prototypes of research was not various; it could not be applied to other samples.

This study has concluded the related researches and the use of text mining technique to apply the classification of economic news. The related researches for the improvement and development model are suited to this research. The experiments of some algorithms with feature selection schemes for getting the higher accuracy is also challenge. The next chapter will be described the research methodology and example of economic news articles with indicators. It also explains the concept of classification model, validation, and schedule of this research.

CHAPTER III
RESEARCH METHODOLOGY

This chapter describes the processes of research and example for economic news. The chapter also explains the concept of model for classification, validation, and schedule of this research.

3.1 Processes of Research Methodology

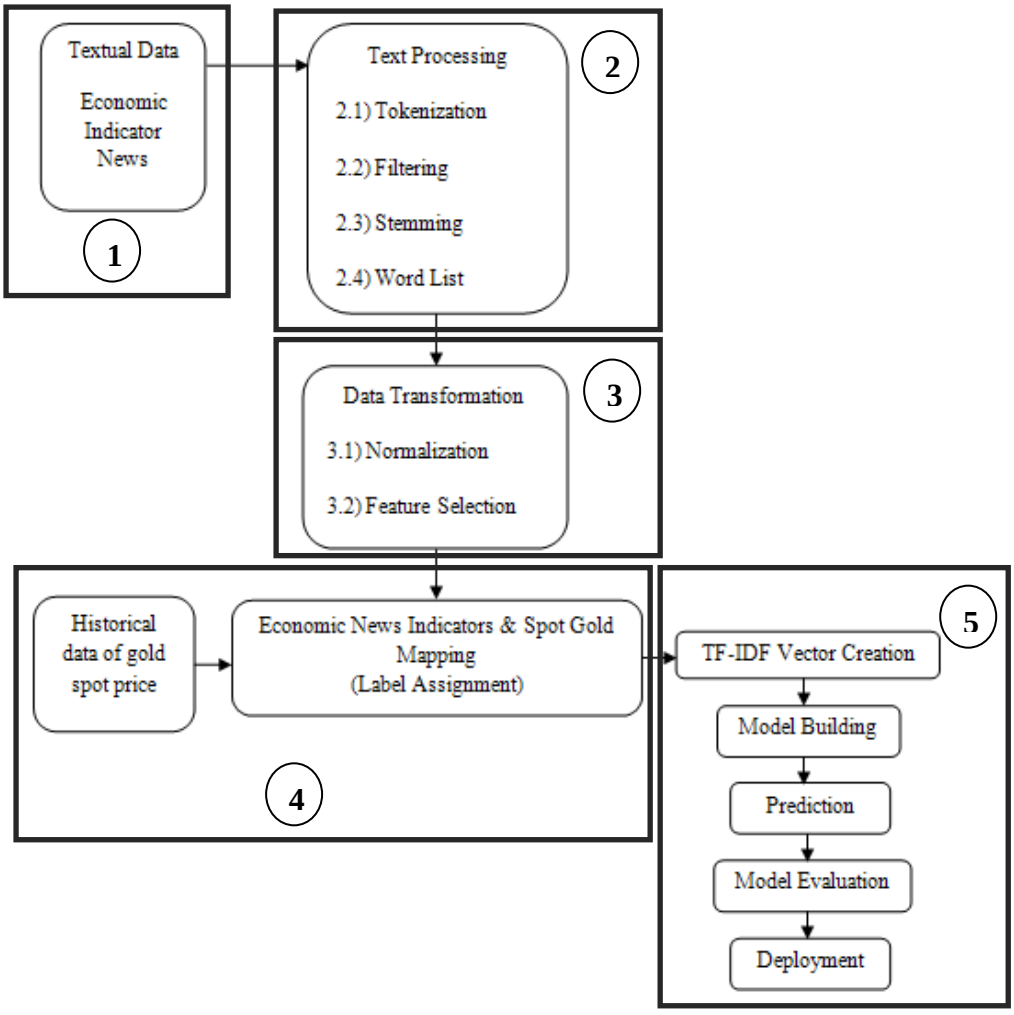

Figure 3.1 Methodology overview.

An overall methodology, as shown in Figure 3.1, is summarized as follows:

1) Implement the economic news data with indicators in text file for text processing.

2) Run the text processing with the sub-processes of the tokenization, filtering, stemming, and word list data, respectively.

3) Run the data transformation with the normalization, and select attributes with feature selection, respectively.

4) Map the gold price's historical data with economic news indicators and assign the label of news article as the pre-process for creating model.

5) Perform TF-IDF vector and create model for volatility prediction of gold price by economic news indicators.

Textual data of economic news indicators with article will be processed by typical pre-processing methods from unstructured textual data to a word-document matrix. It includes the tasks of tokenization, filtering stop words, and the calculation of an important metric. Tokenization is the process of splitting the text into individual words or tokens. Tokenization within the English language is often done by using blank spaces and punctuation marks as token delimiters. Next, the stemming is processes of linguistic normalization, in which the variant forms of word are reduced to a common form, for example, "connective", "connection", "connecting", and "connected" have a common word of "connect". After the process of stemming, word list data will be transformed by normalization process, and select attributes with feature selection by SVM's attributes weighting and Chi Squared Statistic's attributes weighting to be described in Section 3.3.

Consequently, the economic news indicator has been processed by mapping the gold spot price's historical data and assigning the label. The final pre-processing step involves the conversion of wording occurrence to metric document representing its relative importance. In this study, the binary representation of the occurrence of word and the term frequency-inverse document frequency (TF-IDF) are used as the metrics of importance. The binary representation is computed, given as: assigning by 1 if a word in word list is presented in the document, and 0 for otherwise. This metric is used to filter out the words which are appeared in one document. The

TF-IDF computes the frequency of word in document and other documents in a matrix of important wording score.

Then, the model will be created to predict the gold price volatility by SVM, K-NN and Naive Bayes methods. Model is evaluated by 10-fold cross-validation. It is that the basis of this technique is the re-sampling data series into sections, called fold. Some of data folds are used to test and train for the accuracy results of prediction model.

3.2 Economic news indicators with articles

The example of FXStreet.com's economic news indicators with articles [3], are given as:

- Low volatility expected : US Chicago Purchasing Managers' Index increases to 56.8 in Feb from 55.6 in Jan
- Moderate volatility expected : US Nov Factory Orders (MoM) down to 0% vs 0.8% (Oct)
- High volatility expected : Obama addresses congress at the SOTU speech, US GDP falls 0.1%

3.3 Model classification and validation

1) Feature Selection

Feature selection is the kind of attribute selection, considering the weight values of key words with the high percentage formats to create the model, whereas the attribute with low weight value will be rejected for consideration due to the less importance of classification. In the research, the several kinds of feature selection schemes would impact the efficient classification, given as: standard analysis forecast, neural network analysis, regression tree, general linear models, SVM, and Chi Squared Statistic weighting.

- Feature selection with SVM weighting: It is a linear attribute weighting by using the coefficient of normal vector of SVM. The coefficients of hyperplane

calculated by SVM (support vector machine) are set as attribute weights. However, this feature selection can be applied only on numerical data set.

- Feature selection with Chi Squared Statistic weighting: It is a calculation of the attribute using chi-square statistics. The higher weight of the attributes is more related. Attribute weighting operators is to check the key performance indicators (KPI) with the stated purpose, called the label variable.

2) Model building and validation

This process involves a consideration of various styles, and selects the best one based on the predicted performance. This may sound like a simple operation. However, in fact, it sometimes involves a very complex process. There are a variety of techniques to achieve the goal, which are based on the evaluation of the competition. That is different format for the same data set. The results of operations are compared to find the best choice of techniques, which would be the trend of data mining for prediction [17].

This research creates the model to classify the label of economic news indicators for volatility prediction of gold price, and evaluates the performance of the model by testing cross validation test (10-fold cross validation).

3.4 Research Schedule

Schedule of this research started on August 2014 with 5 months for thesis task in Table 3.1.

Table 3.1 Research Timeline.

Task Name	Timeline																			
	August				September				October				November				December			
	w 1	w 2	w 3	w 4	w 1	w 2	w 3	w 4	w 1	w 2	w 3	w 4	w 1	w 2	w 3	w 4	w 1	w 2	w 3	w 4
Issue identification																				
Business understanding																				
Selection of information sources																				
Data understanding																				
Data preparation																				
Modeling																				
Evaluation																				
Deployment																				

The next chapter describes the comparative results of nine experiments including SVM, K-NN, and Naive Bayes with feature selection by SVM weighting, Chi-Squared statistic weighting, and No-feature. The next chapter also explains the result of the test data sets of economic news indicators 2014.

CHAPTER IV
RESULTS ANALYSIS AND DISCUSSION

This chapter describes the comparative results of nine experiments including: SVM, K-NN, and Naive Bayes with feature selection by: SVM weighting, Chi-Squared statistic weighting, and No-feature. The chapter also explains the experimental results unseen data sets of 2014 economic news indicators.

In this work, using the data set of 201 examples are divided into three labels of volatility expectation including high volatility, moderate volatility, and low volatility, as shown in Figure 4.1.

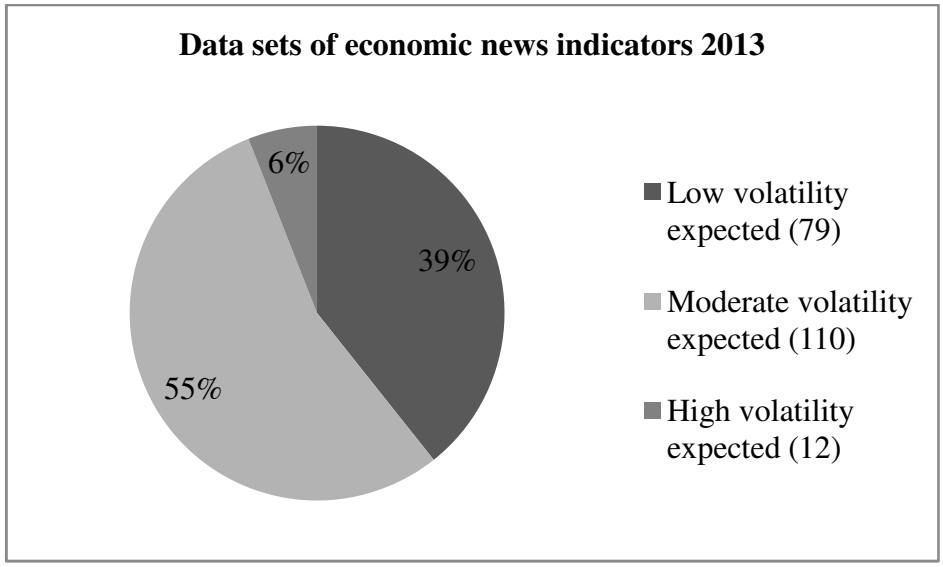

Figure 4.1 Data set summarization.

4.1 Gold price matching with economic news indicators

Gold price is matching with economic news indicators by date and the assigned volatility label, as shown in Table 4.1.

Table 4.1 Example of gold price matching with economic news indicators.

Date	% Gold price change	Economic indicators with news	Volatility Level
Dec 31, 2013	-0.10	Chicago Purchasing Managers' Index	Moderate
Dec 30, 2013	-1.07	Dallas Fed Manufacturing Business Index	Moderate
Dec 27, 2013	0.16	EIA Natural Gas Storage change	Low

4.2 Comparative results of nine experimental schemes

With the classification results of nine experiment schemes, it is found that the model with the SVM classification algorithm and feature selection with SVM weighting is the best among all tests with accuracy of 87.52%. For all methods, SVM can handle the complicate data better than the other methods with both contexts of feature selection and classification, as shown in Figure 4.2 and Table 4.2.

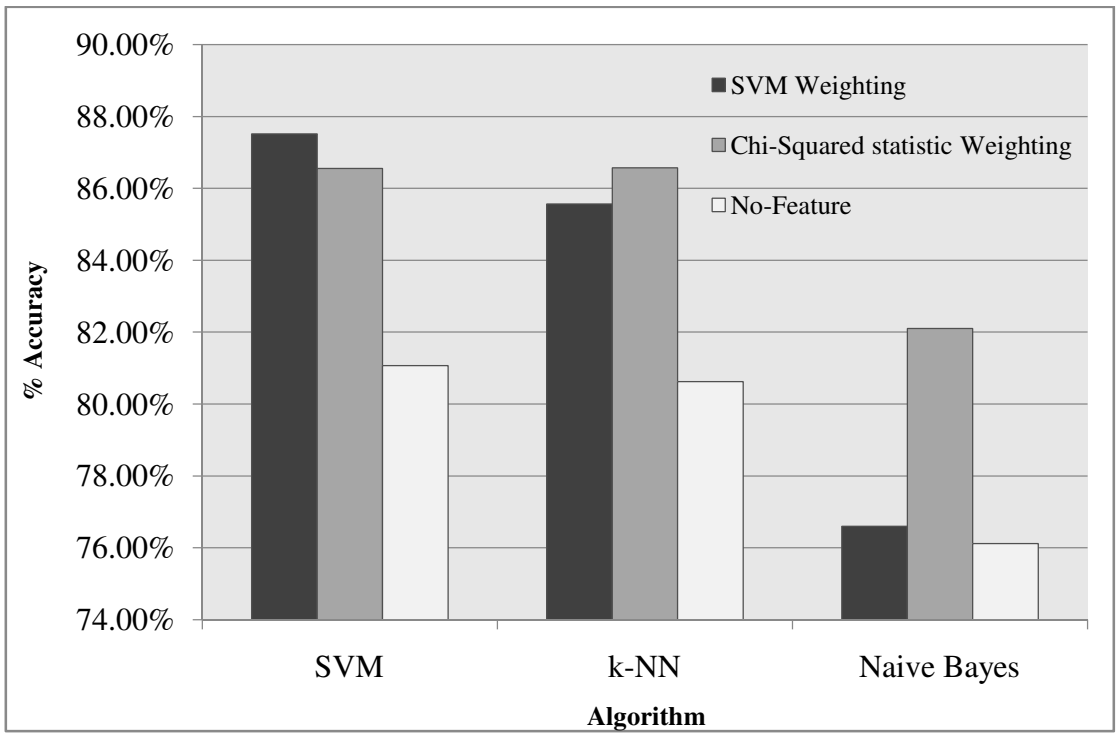

Figure 4.2 Results of nine experiment schemes.

Table 4.2 Comparative results.

Classification Methods	Feature Selection Methods	Accuracy (%)
SVM	SVM Weighting	87.52
	Chi-Squared statistic Weighting	86.55
	No-Feature	81.07
K-NN	SVM Weighting	85.57
	Chi-Squared statistic Weighting	86.57
	No-Feature	80.62
Naive Bayes	SVM Weighting	76.60
	Chi-Squared statistic Weighting	82.10
	No-Feature	76.12

4.3 Attribute weighting by SVM and Chi Squared Statistic

With the attribute weighting by SVM, the feature selection uses the coefficients of the normal vector of a linear SVM as attribute weights, as shown in Figure 4.3.

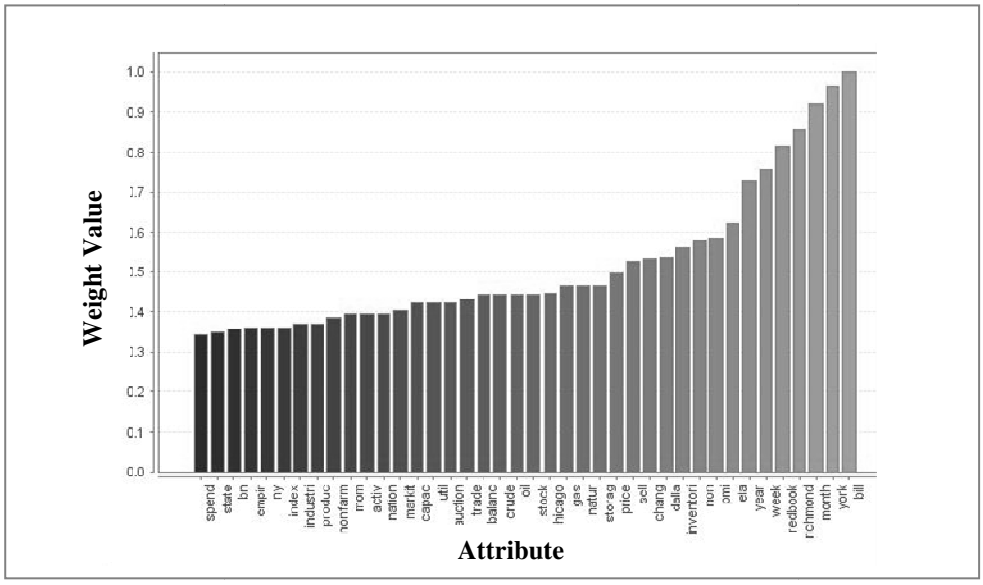

Figure 4.3 Attribute weight results by SVM.

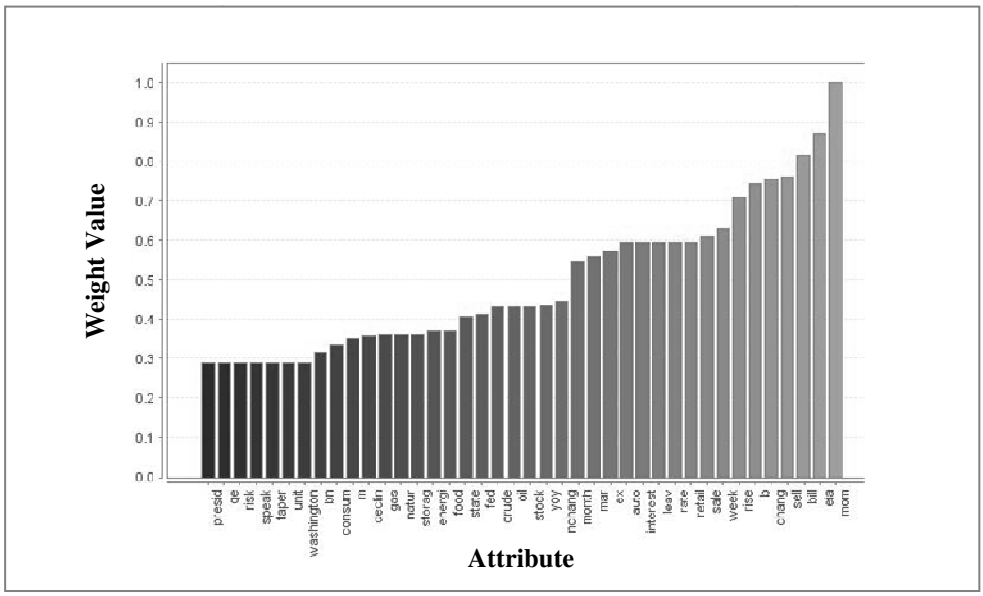

Figure 4.4 Attribute weight results by Chi Squared Statistic.

Figure 4.4 shows the weight by Chi Squared Statistic. The weight attributes are calculated by chi-squared statistic scheme.

4.4 Feature selection

By the selection criteria with top *P* of 0.5 percent, the feature selections are weighted by SVM and Chi Squared Statistic, respectively. The results are filtering for six attributes of SVM and eight attributes of Chi Squared Statistic, as shown in Figures 4.5 - 4.6.

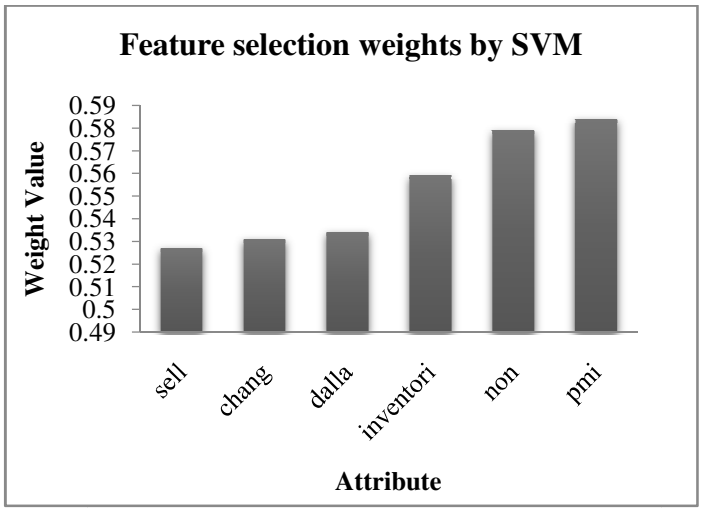

Figure 4.5 Selective results of attribute weights by SVM.

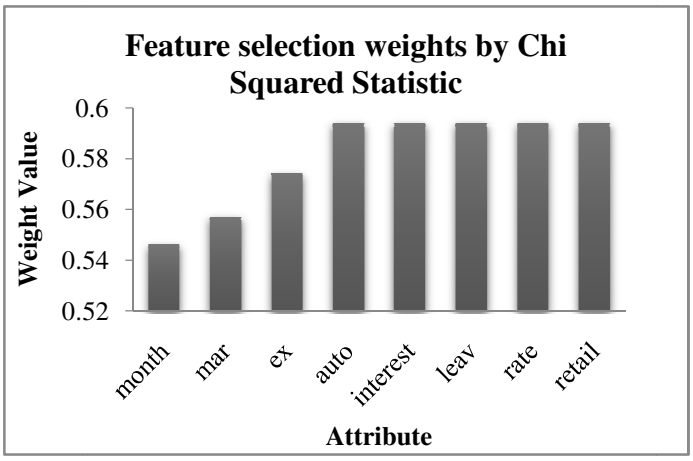

Figure 4.6 Selective results of attribute weights by Chi Squared Statistic.

The feature selection weights by SVM are filtering for six features (or attributes) including sell, chang, dalla, inventori, non, and pmi with selection criteria of top *P* percent. It is noted that the top *P* % is used to specify the percentage of attributes to select range, as shown in Table 4.3.

Table 4.3 Feature selection with attribute weights by SVM.

Feature Selection	Attribute	Weight	Selection Criteria
Attribute Weights by SVM	sell	0.527	Top *P* percent of 0.5
	chang	0.531	
	dalla	0.534	
	inventori	0.559	
	non	0.579	
	pmi	0.584	

The feature selection weights by Chi Squared Statistic are filtering for eight features (or attributes) including month, mar, ex, auto, interest, leav, rate, and retail selection criteria with top *P* percent, as shown in Table 4.4.

Table 4.4 Feature selection with attribute weights by Chi Squared Statistic.

Feature Selection	Attribute	Weight	Selection Criteria
Attribute Weights by Chi Squared Statistic	month	0.546	Top *P* percent of 0.5
	mar	0.557	
	ex	0.574	
	auto	0.594	
	interest	0.594	
	leav	0.594	
	rate	0.594	
	retail	0.594	

4.5 Word list data

By experiment the occurrences results of word list for the labels are shown in Table 4.5. For example of results, the product's attribute has total occurences number of 14, and index's attribute has total occurences number of 49.

Table 4.5 Example of word list.

Attribute Name	Total occurences	Low volatility	Moderate volatility	High volatility
yoy	17	4	11	2
sell	32	24	8	0
claim	6	0	6	0
fall	9	5	4	0
index	49	18	29	2
product	14	2	11	1

4.6 Result of 2014's unseen data sets with economic news indicators

The example of the 2014 economic news indicators (unseen data sets) are divided into three labels including high volatility, moderate volatility, and low volatility, as shown in Table 4.6. The unseen data of the 2014 economic news indicators is shown in Appendix B.

Table 4.6 Example of unseen data sets (2014 Economic news indicators).

Date	Economic indicators news	Volatility Level
January 28, 2014	US Durable Goods Orders decline 4.3% in December; ex-transportation orders fall 1.6%	High
March 03, 2014	US Personal Spending came in at 0.4% beating forecasts of 0.1% in January	Moderate
April 24, 2014	US EIA Natural Gas Storage change registered at 49B to beat expectations (40B) in April 18	Low

In our research, the application model of SVM classification with feature selection based on SVM weighting is tested on new data sets of the 2014 economic news indicators, there are 35 news divided into 3 major groups of volatility data set, given as: the 5 high volatility data sets, 15 low volatility data sets, and 15 moderate volatility data sets. The accuracy result is with 85.71 %, as shown in Figure 4.7.

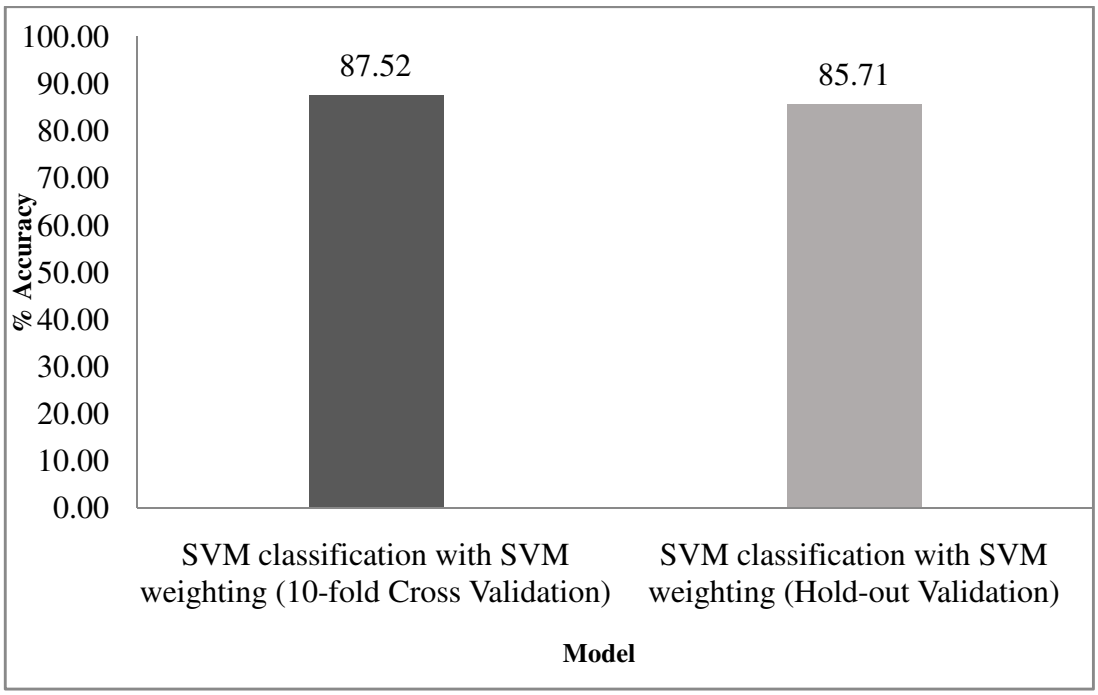

Figure 4.7 Comparative results between training data set and unseen data set.

4.7 Result of experimental schemes detail

In our experiment, the model schemes consists of three algorithms and two features selection with non-features selection, as shown in Tables 4.7 - 4.15.

For the results of SVM classification with SVM weighting, it is found that the accuracy is 87.52%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 2 data sets of economic news indicators. True low volatility prediction equals 71 data sets of economic news indicators. True moderate volatility prediction equals 103 data sets of economic news indicators, as shown in Table 4.7.

Table 4.7 Classification Results of SVM Classification with SVM Weighting.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	2	0	0	100.00
Low volatility	2	71	7	88.75
Moderate volatility	8	8	103	86.55

For the results of SVM classification with Chi Squared Statistic weighting, it found that the accuracy is 86.55%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 4 data sets of economic news indicators. True low volatility prediction equals 61 data sets of economic news indicators. True moderate volatility prediction equals 109 data sets of economic news indicators, as shown in Table 4.8.

Table 4.8 Classification Results of SVM Classification with Chi Squared Statistic Weighting.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	4	0	0	100.00
Low volatility	2	61	1	95.31
Moderate volatility	6	18	109	81.95

For the results of SVM classification, it found that the accuracy is 81.07%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 0 data sets of economic news indicators. True low volatility prediction equals 55 data sets of economic news indicators. True moderate volatility prediction equals 108 data sets of economic news indicators, as shown in Table 4.9.

Table 4.9 Classification Results of SVM Classification.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	0	0	1	0.00
Low volatility	0	55	1	98.21
Moderate volatility	12	24	108	75.00

For the results of k-NN classification with SVM weighting, it found that the accuracy is 85.57%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 5 data sets of economic news indicators. True low volatility prediction equals 71 data sets of economic news indicators. True moderate volatility prediction equals 96 data sets of economic news indicators, as shown in Table 4.10.

Table 4.10 Classification Results of k-NN Classification with SVM Weighting.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	5	1	3	55.56
Low volatility	1	71	11	85.54
Moderate volatility	6	7	96	88.07

For the results of k-NN classification with Chi Squared Statistic weighting, it found that the accuracy is 86.57%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 4 data sets of economic news indicators. True low volatility prediction equals 73 data sets of economic news indicators. True moderate volatility prediction equals 97 data sets of economic news indicators, as shown in Table 4.11.

Table 4.11 Classification Results of k-NN Classification with Chi Squared Statistic Weighting.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	4	1	2	57.14
Low volatility	0	73	11	86.90
Moderate volatility	8	5	97	88.18

For the results of k-NN classification, it found that the accuracy is 80.62%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 4 data sets of economic news indicators. True low volatility prediction equals 73 data sets of economic news indicators. True moderate volatility prediction equals 85 data sets of economic news indicators, as shown in Table 4.12.

Table 4.12 Classification Results of k-NN Classification.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	4	1	2	57.14
Low volatility	4	73	23	73.00
Moderate volatility	4	5	85	90.43

For the results of Naïve Bayes classification with SVM weighting, it found that the accuracy is 76.60%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 6 data sets of economic news indicators. True low volatility prediction equals 74 data sets of economic news indicators. True moderate volatility prediction equals 74 data sets of economic news indicators, as shown in Table 4.13.

Table 4.13 Classification Results of Naïve Bayes Classification with SVM Weighting.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	6	0	13	31.58
Low volatility	4	74	23	73.27
Moderate volatility	2	5	74	91.36

For the results of Naïve Bayes classification with Chi Squared Statistic weighting, it found that the accuracy is 82.10%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 4 data sets of economic news indicators. True low volatility prediction equals 61 data sets of economic news indicators. True moderate volatility prediction equals 100 data sets of economic news indicators, as shown in Table 4.14.

Table 4.14 Classification Results of Naïve Bayes Classification with Chi Squared Statistic Weighting.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	4	0	4	50.00
Low volatility	2	61	6	88.41
Moderate volatility	6	18	100	80.65

For the results of Naïve Bayes classification, it found that the accuracy is 76.12%. Comparisons of prediction effectiveness are found that true high volatility prediction equals 5 data sets of economic news indicators. True low volatility prediction equals 75 data sets of economic news indicators. True moderate volatility prediction equals 73 data sets of economic news indicators, as shown in Table 4.15.

Table 4.15 Classification Results of Naïve Bayes.

Prediction	True High volatility	True Low volatility	True Moderate volatility	% Class Precision
High volatility	5	0	10	33.33
Low volatility	5	75	27	70.09
Moderate volatility	2	4	73	92.41

The next chapter describes the conclusion of this research presents the framework for using text mining to create model for predict gold price volatility by economic news indicators and future works.

CHAPTER V
CONCLUSION AND FUTURE WORKS

This chapter describes about conclusion of this research presents the framework for using text mining to create model and future works.

5.1 Conclusion

In this research, with the feature selection, it is found that the attribute weighting of SVM is more accurate than Chi Squared Statistic feature for the prediction of gold price volatility. SVM delivers the weights of the attributes relating the label attribute. The attributes with higher weight are considered more relevant.

The research presents the framework of text mining for generating model of gold price volatility's prediction. The framework studies are the factors of economic news articles as indicators. Typical pre-processing methods used for transformation will have normalization process and attributes selection with classification algorithm for measuring the result. The effectiveness of SVM, K-NN, and Naive Bayes algorithms are compared for prediction. The best performance method is SVM algorithm with the attribute weight by SVM feature with accuracy of 87.52%.

However, this research is unable to predict the actual price of gold, but the model would be helpful to predict the volatility of economic news indicators affecting the gold price.

5.2 Future works

In the future, this research can be developed to enhance the performance and complexity of the prediction model, as follows:

- The sample data of economic news should have the diversity to create the prediction model.
- The weight relation parameters are adjusted in various forms to suit the data used for building model.
- The model can be developed to forecast the gold price affecting the economic news.
- The model in this research can be applied for stock markets, derivative markets, commodities markets, and forex markets.
- The applications of the numerical data are analyzed and recombined with text processing for the best performance.

In the future, text analysis would be related to text information retrieval, lexical analysis, information extraction techniques, data mining techniques including link, the distribution frequency of pattern recognition tagging, association analysis, predictive analytics, and visualization. Moreover, overarching goal of forecasting is essentially all the data analysis through the application of natural language processing (NLP) and analysis related the area of human-computer interaction. NLP, involving many challenges in natural language understanding, has to derive the meaning of human language input.

To improve the classification system, we need the diversity of economic news for more factors to improve the efficient analysis of consequences affecting the volatility of the gold price. The experiments of other algorithms and feature selection methods for the higher accuracy are also challenge.

REFERENCES

1 Roache SK, Rossi M. (2009). *The effects of economic news on commodity prices: Is gold just another commodity*. International Monetary Fund.

2 Liang Y, Xiaozhou Z. (2009). The Oscillation of China's Futures and Spot Gold Prices and The Price Determination Mechanism [J]. *Shanghai Finance*, *4*, 012.

3 FXStreet. (2014). *Economic news.* [Online]. available: http://www.fxstreet.com

4 Chang CL, Della Chang JC, Huang YW. (2013). Dynamic price integration in the global gold market. *The North American Journal of Economics and Finance*, *26*, 227-235.

5 FXStreet. (2014). *Economic Categories.* [Online]. available: http://www.fxstreet.com/economic-calendar

6 Bernstein A, Provost F, Hill S. (2005). Toward intelligent assistance for a data mining process: An ontology-based approach for cost-sensitive classification. *Knowledge and Data Engineering, IEEE Transactions on*, *17*(4), 503-518.

7 Wirth R, Hipp J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29-39).

8 Tseng YH, Lin CJ, Lin YI. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216-1247.

9 Ramos J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*.

10 Chou KC, Shen HB. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers.*Journal of proteome research*, *5*(8), 1888-1897.

11 Joachims T. (1998). *Text categorization with support vector machines: Learning with many relevant features* (pp. 137-142). Springer Berlin Heidelberg.

12 McCallum A, Nigam K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41-48).

13 Rivera SJ, Minsker BS, Work DB, Roth D. (2014). A text mining framework for advancing sustainability indicators. *Environmental Modelling & Software*, *62*, 128-138.

14 Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL. (2015). Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. *Expert Systems with Applications*, *42*(1), 306-324.

15 Hagenau M, Liebmann M, Neumann D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features.*Decision Support Systems*, *55*(3), 685-697.

16 Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. (2000, December). Feature selection for SVMs. In *NIPS* (Vol. 12, pp. 668-674).

17 Gnanapriya S, Suganya R, Devi GS, Kumar MS. (2010). Data Mining Concepts and Techniques. *Data Mining and Knowledge Engineering*, *2*(9), 256-263.

APPENDICES

APPENDIX A
2013 NEWS HISTORICAL DATA FOR GENERATING MODEL

Table A Economic news indicators of 2013. Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Dec 31, 2013	US Redbook index up to -0.7% from -1%	Low
Dec 30, 2013	Dallas Fed manufacturing output index Dec +7.1 vs +16.9 prior	Moderate
Dec 27, 2013	US December 20 EIA Natural Gas Storage change improves to -177B from -285B	Low
Dec 24, 2013	US durable goods orders Nov +3.5% vs +2.0% exp	Moderate
Dec 23, 2013	US November Personal Income (MoM) up to 0.2% vs -0.1% (October)	Moderate
Dec 20, 2013	US 3Q Gross Domestic Product Price Index improves to 2% vs 0.6% (2Q)	Moderate
Dec 19, 2013	The Fed leaves interest rate unchanged at 0.25% and announces $10B taper	High
Dec 18, 2013	US EIA Crude Oil Stocks change up to -2.941M in December 13 from -10.585M	Low
Dec 17, 2013	US Consumer Price Index Core s.a increases to 235.24 in November from 234.88 in Oct.	Moderate
Dec 16, 2013	US Markit Manufacturing PMI falls to 54.4 in December from 54.7 in November	Moderate
Dec 13, 2013	US Business Inventories up to 0.7% in October from 0.6%	Moderate
Dec 12, 2013	US November Retail Sales rises 0.7% MoM; Ex-autos +0.4%	High

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Dec 11, 2013	US EIA Crude Oil Stocks change: -10.585M in December 6 from -5.585M	Low
Dec 09, 2013	US sells 3-month bills by $32B at 0.07%	Low
Dec 05, 2013	US Gross Domestic Product Price Index up to 2% in 3Q from 0.6% (2Q)	Moderate
Dec 03, 2013	US ISM New York index up to 69.5 in November from 59.3 in October	Low
Dec 02, 2013	Investors risk averse ahead of very important week	High
Nov 27, 2013	US EIA Crude Oil Stocks change (November 22): 2.953M vs 0.375M	Low
Nov 26, 2013	US Richmond Fed Manufacturing Index up to 13 in November from 1 in October	Low
Nov 25, 2013	US sells $32B in 3-month bills at 0.080%	Low
Nov 21, 2013	US Philadelphia Fed Manufacturing Survey decreases to 6.5 in November from 19.8	Moderate
Nov 20, 2013	US Business Inventories (September): 0.6% vs 0.4% (August)	Moderate
Nov 19, 2013	US sells $45B in 4-week bills at 0.06%	Low
Nov 15, 2013	US NY Empire State Manufacturing Index down to -2.21 in November	Low
Nov 14, 2013	US Initial Jobless Claims: 339K in November 8 from 336K	Moderate
Nov 13, 2013	US sells $45B in 4-week bills at 0.06%	Low
Nov 12, 2013	US Chicago Fed National Activity Index up to 0.14 in September from 0.13	Moderate
Nov 07, 2013	US 3Q Gross Domestic Product Price Index up to 1.9% vs 0.6% (2Q)	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Nov 06, 2013	US 3Q Nonfarm Productivity rises to 3% vs 1.8% (2Q)	Moderate
Nov 05, 2013	US October ISM Non-Manufacturing PMI increase to 55.4 vs 54.4 in September	Moderate
Nov 04, 2013	US ISM New York index (October): 59.3 vs 53.6 (September)	Low
Nov 01, 2013	US Markit Manufacturing PMI declines to 51.8 in October.	Moderate
Oct 31, 2013	US October Chicago PMI jumps to 65.9	Moderate
Oct 30, 2013	US EIA Crude Oil Stocks change (October 25): 4.087M vs 5.246M	Low
Oct 29, 2013	US September Producer Price Index: MoM -0.1% ; 0.3% YoY	Moderate
Oct 28, 2013	US Capacity Utilization up78.3% in September vs 77.9% in August	Low
Oct 25, 2013	US September 20 EIA Crude Oil Stocks change increase to 2.635M vs -4.368M	Low
Oct 24, 2013	US October Markit Manufacturing PMI decreases to 51.1 vs 52.8 in September	Moderate
Oct 21, 2013	US 3-month bills at 0.035%	Low
Oct 17, 2013	US: Consumer Price Index Ex Food & Energy: 0.1% MoM; 1.8% YoY	High
Oct 16, 2013	US Industrial Production (MoM) (August): 0.4% vs 0% (July)	Moderate
Oct 15, 2013	US October NY Empire State Manufacturing Index falls to 1.52 vs 6.29	Low
Oct 08, 2013	US sells $30 bn in 4-week bills at 0.35%	Low
Oct 07, 2013	US Consumer Credit rises to $13.63B	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Oct 03, 2013	US September 27 EIA Natural Gas Storage change increase to 101B vs 87B	Low
Oct 01, 2013	US ISM Manufacturing PMI improves to 56.2 in September from 55.7 In August	High
Sep 30, 2013	US Chicago Purchasing Managers' Index improves to 55.7 in September from 53	Moderate
Sep 26, 2013	US Gross Domestic Product Price Index down to 0.6% in 2Q	Moderate
Sep 25, 2013	US Durable Goods Orders ex Transportation: -0.1% in August	Moderate
Sep 24, 2013	US: Housing Price Index climbes 1% in July	Moderate
Sep 23, 2013	US sells $30Bn of 3-month bills at 0.020%	Low
Sep 19, 2013	The Fed leaves unchanged its interest rate at 0.25%	High
Sep 18, 2013	US: EIA Crude Oil Stocks change (September 13): -4.368M vs -0.219M	Low
Sep 17, 2013	US: Consumer Price Index Ex Food & Energy: 0.1% MoM; 1.8% YoY	Moderate
Sep 16, 2013	US Industrial Production (MoM) (August): 0.4% vs 0% (July)	Moderate
Sep 13, 2013	US Producer Price Index (YoY) up 1.4% in August; (MoM) grow 0.3%	Moderate
Sep 12, 2013	The president of the United States, Barack Obama, will speak about the economy, in Washington DC.	High
Sep 11, 2013	US EIA Crude Oil Stocks change: -0.219M in September 6 from -1.836M	Low
Sep 09, 2013	US sells $30 Bn in 3-month bills at 0.020%	Low

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Sep 05, 2013	US ISM Non-Manufacturing PMI increase to 58.6 in August from 56 in July	Moderate
Sep 04, 2013	US: Trade deficit widened to $39.15 billion in July	Moderate
Sep 03, 2013	US: July Construction Spending up 0.6%	Moderate
Aug 30, 2013	August Chicago PMI 53.0 vs 53.0 expected	Moderate
Aug 29, 2013	US EIA Natural Gas Storage change increase to 96B in August 2 from 59B	Low
Aug 28, 2013	10-Year Notes at 1.624%	Low
Aug 27, 2013	US: Consumer confidence index rises to 81.5 in August	Moderate
Aug 26, 2013	US sells 3-month bills by $30Bn at 0.040%	Low
Aug 22, 2013	US EIA Natural Gas Storage falls by 57B in August 16	Low
Aug 20, 2013	US Chicago Fed National Activity Index improves to -0.15 in July from revised -0.23 in June	Moderate
Aug 19, 2013	US sells $30Bn in 3-month bills at 0.05%	Low
Aug 16, 2013	US Housing Starts (MoM) improves to 0.896M in July from 0.846M in June	Moderate
Aug 15, 2013	US July Consumer Price Index Ex Food & Energy (YoY) rises 1.7%; grows 0.2%	High
Aug 14, 2013	US Producer Price Index (YoY) down to 2.1% in July; 0% (MoM)	Moderate
Aug 13, 2013	US Business Inventories increase to 0% in June from -0.1% in May	Moderate
Aug 12, 2013	US July monthly budget deficit $97.6 billion vs $94.5 billion expected	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Aug 08, 2013	US July 27 Continuing Jobless Claims increase to 3.018M vs 2.951M	Moderate
Aug 07, 2013	US August 2 EIA Crude Oil Stocks change: -1.32M vs 0.431M	Low
Aug 06, 2013	US Trade Balance improves to $-34.22B in June from $-44.1B.	Moderate
Aug 05, 2013	US ISM Non-Manufacturing PMI up to 56 in July from 52.2 in June	Moderate
Aug 02, 2013	US: ISM New York Index rises to 67.8	Low
Aug 01, 2013	US July 26 EIA Natural Gas Storage: 59B	Low
Jul 31, 2013	US Chicago Purchasing Managers' Index: 52.3 in July	Moderate
Jul 30, 2013	US sells 45B of 4-Week Bill at 0.02%	Low
Jul 29, 2013	US sells $30Bn in 3-month bills at 0.03%	Low
Jul 25, 2013	US June Durable Goods Orders increase 4.2%; ex Transportation remain unchanged	Moderate
Jul 24, 2013	US Markit Manufacturing PMI (July): 53.2 vs 51.9	Moderate
Jul 23, 2013	US July Richmond Fed Manufacturing Index declines to -11 from 8 in June	Low
Jul 18, 2013	US July Philadelphia Fed Manufacturing Survey rises to 19.8 vs 12.5 in June	Moderate
Jul 17, 2013	US Building Permits (MoM) decreases to 0.911M in June from 0.985M in May	Moderate
Jul 16, 2013	US June Industrial Production up to 0.3%	Moderate
Jul 15, 2013	US sells $30Bn of 3-month bills at 0.04%	Low
Jul 12, 2013	US June Producer Price Index ex Food & Energy (MoM) up to 0.2%; 1.7% (YoY)	Low

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Jul 11, 2013	FOMC Minutes: About half of members see QE3 ending this year	High
Jul 10, 2013	US EIA Crude Oil Stocks change: -9.874M in July 5 from -10.347M	Low
Jul 09, 2013	IMF cuts global growth forecast, sees US growing just 1.7% this year	Moderate
Jul 08, 2013	US Consumer Credit rises by $19.6B in May	Moderate
Jul 03, 2013	US ISM Non-Manufacturing PMI declines to 52.2 in June from 53.7 in May	Moderate
Jul 02, 2013	US May Factory Orders (MoM) up to 2.1% vs 1.3%	Moderate
Jun 28, 2013	US June Chicago Purchasing Managers' Index falls to 51.6 vs 58.7 in May	Moderate
Jun 27, 2013	US Personal Spending up to 0.3% in May from -0.3% in April.	Moderate
Jun 26, 2013	US 1Q Gross Domestic Product Annualized revised down to 1.8%	High
Jun 25, 2013	US Redbook index (MoM): -0.5% in June 16; 2.8% (YoY)	Low
Jun 24, 2013	US Chicago Fed National Activity Index: -0.3 in May	Moderate
Jun 18, 2013	US Housing Starts (MoM) increase to 0.914M in May from 0.856M	Moderate
Jun 17, 2013	US: NY Empire State index rose to 7.84 in June	Low
Jun 14, 2013	US Producer Price Index (MoM) increase to 0.5% in May; 1.7% (YoY)	Moderate
Jun 13, 2013	EUR/USD testing 1.3300 barrier	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Jun 12, 2013	US sells $21 billion of 10-Year Note at 2.209%	Moderate
Jun 10, 2013	US sells $30 Bn of 3-month bills at 0.045%	Low
Jun 06, 2013	US Initial Jobless Claims add 346K in Jun 1 and Continuing obtains 2.952M	Moderate
Jun 05, 2013	US: EIA Crude Oil Stocks change (May 31): -6.267M vs 3M	Low
Jun 04, 2013	US Trade Balance declines to $-40.29B in Apr from $-37.13B in Mar	Moderate
Jun 03, 2013	US: Apr Construction Spending (MoM) increases 0.4% vs -0.8% in Mar	Moderate
May 31, 2013	US Chicago Purchasing Managers' Index improves to 58.7 in May from 49 in Apr	Moderate
May 30, 2013	US: GDP expanded 2.4% YoY in Q1	Moderate
May 29, 2013	U.S. sells $35 Bn of 5-year notes at 1.045%	Moderate
May 28, 2013	US sells $35 Bn of 2-year notes at 0.283%	Moderate
May 24, 2013	US Durable Goods Orders up to 3.3% in Apr; ex Transport: 1.3%	Low
May 23, 2013	US Markit Manufacturing PMI decreases to 51.9 in May from 52.1 in Apr	Moderate
May 21, 2013	US sells $45 Bn of 4-week bills at 0.035%	Low
May 20, 2013	US Apr Chicago Fed National Activity Index: -0.53 vs -0.23 in Mar	Moderate
May 16, 2013	US Apr Building Permits (MoM) rises to 1.017M vs 0.89M	Moderate
May 15, 2013	US Producer Price Index (YoY) decreases to 0.6% in Apr; 0.1% (MoM)	Moderate
May 14, 2013	US Business Inventories stay unchanged 0%	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
May 13, 2013	US sells $29 Bn of 3-month notes at 0.045%	Low
May 09, 2013	US EIA Natural Gas Storage change up to 88B in May 3 from 43B	Low
May 08, 2013	US sells 24B of 10-Year Notes at 1.810%	Moderate
May 07, 2013	US: Consumer Credit rises by $7.97B	Moderate
May 06, 2013	US sells $29 Bn of 3-month bills at 0.040%	Low
May 03, 2013	US: EIA Natural Gas Storage rose 43 Bcf in April-26 week	Low
May 02, 2013	US Trade Balance rises to $-38.8B in Mar from $-43.6B in Feb	Moderate
May 01, 2013	US Construction Spending declines 1.7%	Moderate
Apr 30, 2013	US sells $30 Bn of 4-week bills at 0.025%	Low
Apr 29, 2013	US Personal Spending declines to 0.2% in Mar from 0.7% in Feb	Moderate
Apr 26, 2013	US 1Q Gross Domestic Product Annualized grows 2.5% and Price Index increases 1.2%	Moderate
Apr 25, 2013	US Durable Goods Orders: -5.7% in Mar from 4.3% in Feb	High
Apr 24, 2013	US Durable Goods Orders ex Transportation increase to -1.4% in Mar from -1.7% in Feb	Moderate
Apr 23, 2013	US Markit Manufacturing PMI falls to 52 in Apr from 54.6 in Mar	Moderate
Apr 22, 2013	US: Chicago Fed National Activity index down to -0.23 in March	Moderate
Apr 18, 2013	US: Philadelphia Fed Manufacturing Survey disappoints at 1.3 in April	Moderate
Apr 16, 2013	US: Building Permits (MoM) (Mar): 0.902M vs 0.968M	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Apr 12, 2013	US Mar Retail Sales (MoM) declines to -0.4% vs 1% in Feb. Ex-autos -0.4%	High
Apr 11, 2013	US EIA Natural Gas Storage change rises to -14B in Apr 5 from -94B	Low
Apr 10, 2013	US sells $21Bn of 10-year bills at 1.795%	Moderate
Apr 09, 2013	US Trade Balance rises to $-43B in Feb from $-44.45B	Moderate
Apr 08, 2013	US sells $30Bn of 6-month bills at 0.095%	Low
Apr 04, 2013	US EIA Natural Gas Storage change (Mar 29): -94B vs -95B	Low
Apr 03, 2013	US ISM Non-Manufacturing PMI falls to 54.4 in Mar from 56 in Feb	Moderate
Apr 02, 2013	US Factory Orders (MoM) up to 3% in Feb from -1% in Jan	Moderate
Apr 01, 2013	US: Mar Markit Manufacturing PMI up to 54.6	Low
Mar 28, 2013	US: Initial jobless claims disappoint at 357K in March-24	Moderate
Mar 27, 2013	US Pending Home Sales (MoM): -0.4% in Feb; 8.4% (YoY)	Moderate
Mar 26, 2013	US Feb Durable Goods Orders ex Transportation: -0.5% vs 2.9% (Jan)	Moderate
Mar 25, 2013	US Dallas Fed Manufacturing Business Index up to 7.4 in Mar from 2.2 in Feb	Low
Mar 21, 2013	US Mar Markit Manufacturing PMI increase to 54.9 vs 54.3 in Feb.	Low
Mar 19, 2013	US 4-Week Bill Auction decreases to 0.08% vs 0.1%	Low

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Mar 18, 2013	US Feb Industrial Production (MoM) improves to 0.7% vs 0% in Jan.	Moderate
Mar 15, 2013	US Mar NY Empire State Manufacturing Index down to 9.24 vs 10.04 in Feb.	Low
Mar 14, 2013	US Mar 8 EIA Natural Gas Storage change: -145B vs -146B	Low
Mar 13, 2013	US EIA Crude Oil Stocks change decreases to 2.624M in Mar 8 from 3.833M	Low
Mar 12, 2013	US sells $32B of 3-year notes at 0.411%	Moderate
Mar 11, 2013	US 6-Month Bill Auction falls to 0.115%; 3-month: 0.095%	Low
Mar 07, 2013	US Nonfarm Productivity down to -1.9% in 4Q from 3.2% in 3Q.	Low
Mar 06, 2013	US EIA Crude Oil Stocks change rises to 3.833M in Mar 1 from 1.13M	Low
Mar 05, 2013	US sells $45B of 4-week bills at 0.085%	Low
Mar 04, 2013	US ISM New York index up to 58.8 in Feb from 56.7 in Jan	Low
Mar 01, 2013	US EIA Natural Gas Storage change: -171B in Feb 22 from -127B	Low
Feb 28, 2013	US Feb 24 Initial Jobless Claims add 344K and Continuing obtain 3.074M	Moderate
Feb 27, 2013	US Jan Durable Goods Orders: -5.2% vs 3.7% (Dec)	Moderate
Feb 26, 2013	US Housing Price Index (MoM) rises to 0.6% in Dec from 0.4% in Nov	Moderate
Feb 25, 2013	US Jan Chicago Fed National Activity Index: -0.32 vs 0.25 (Dec)	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Feb 21, 2013	US Feb 15 EIA Natural Gas Storage change increase to -127B vs -157B	Low
Feb 20, 2013	US Producer Price Index (MoM) up to 0.2% in Jan from -0.2% in Dec. (YoY) 1.4%	Moderate
Feb 19, 2013	US sells $35B 3-month bills at 0.115%	Low
Feb 15, 2013	US Jan Industrial Production -0.1% vs 0.4%	Moderate
Feb 14, 2013	US: Initial Jobless Claims (Feb 10) add 341K and Continued obtain 3.114M	Moderate
Feb 13, 2013	US EIA Crude Oil Stocks change down to 0.56M in Feb 8 from 2.623M	Low
Feb 12, 2013	US sells $45B of 4-week bills at 0.080%	Low
Feb 11, 2013	US: Trade Balance (Dec): $-38.54B vs $-48.61B (Nov).	Moderate
Feb 07, 2013	US 4Q Nonfarm Productivity falls to -2%	Low
Feb 05, 2013	US Redbook index (YoY) grows 1.5% in Jan 27 and (MoM) decreases 0.6%	Low
Feb 04, 2013	US sells $32B 3-month bills at 0.070%	Low
Feb 01, 2013	US Dec Construction Spending (MoM) rises to 0.9% vs -0.3%	Moderate
Jan 31, 2013	US Chicago Purchasing Managers' Index up to 55.6 in Jan from 50 in Dec	Moderate
Jan 30, 2013	US: GDP contracted 0.1% YoY in Q4	Moderate
Jan 29, 2013	US Dec Durable Goods Orders rises to 4.6% vs 0.8% (Nov)	Moderate
Jan 28, 2013	US Jan Dallas Fed Manufacturing Business Index falls to 5.5 vs 2.5 (Dec)	Low
Jan 24, 2013	US Jan Markit Manufacturing PMI improves to 56.1 vs 54 (Dec)	Moderate

Table A Economic news indicators of 2013 (cont.). Source: FXStreet, 2013 [3].

Date	Description of Economic News Indicators	Volatility Level
Jan 23, 2013	US sells $30B 4-week bills at 0.060%	Low
Jan 22, 2013	US Jan Richmond Fed Manufacturing Index: -12 vs 5 (Dec)	Low
Jan 17, 2013	US EIA Natural Gas Storage change: -148B in Jan 11 from -201B	Low
Jan 16, 2013	US Capacity Utilization improves to 78.8% in Dec from 78.7% in Nov	Low
Jan 15, 2013	US Producer Price Index (YoY) declines to 1.3% in Dec; -0.2% (MoM)	Moderate
Jan 14, 2013	US Trade Balance: $-48.73B in Nov from $-42.06B in Oct	Moderate
Jan 10, 2013	US sells $13B of 30Y bonds at 1.863%	Moderate
Jan 09, 2013	US Dec 29 EIA Crude Oil Stocks change up to 1.314M vs -11.1M	Low
Jan 08, 2013	US sells 3-year notes at 0.385%, as expected	Moderate
Jan 07, 2013	US sells $35B 3-month bills at 0.065%	Low
Jan 04, 2013	US Nov Factory Orders (MoM) down to 0% vs 0.8% (Oct)	Moderate
Jan 03, 2013	US ISM Non-Manufacturing PMI improves to 56.1 in Dec from 54.7 (Nov.)	Moderate
Jan 02, 2013	US Construction Spending (MoM): -0.3%	Moderate

APPENDIX B
ECONOMIC NEWS INDICATORS OF 2014'S UNSEEN DATA

Table B Economic news indicators of 2014. Source: FXStreet, 2014 [3].

Date	Description of Economic News Indicators	Volatility Level
Jan 28, 2014	US Durable Goods Orders decline 4.3% in December; ex-transportation orders fall	High
Jan 29, 2014	The Fed leaves interest rate unchanged at 0.25%; another taper of $10bn announced	High
Feb 28, 2014	US Gross Domestic Product Annualized missed expectations (2.5%) in 4Q: 2.4%	High
Mar 18, 2014	US Consumer Price Index Ex Food & Energy (YoY) in line with forecasts (1.6%) in February	High
Apr 24, 2014	US: Durable Goods Orders (Mar) rose 2.6%n line with forecasts (1.6%) in February	High
Jan 02, 2014	US December 27 Initial Jobless Claims decreases to 339K vs 341K	Moderate
Jan 09, 2014	US Initial Jobless Claims declines to 330K in January 3	Moderate
Jan 14, 2014	November 2013 US business inventories 0.4% vs 0.3% exp m/m	Moderate
Jan 28, 2014	US: Consumer Confidence improves to 80.7 in January	Moderate
Feb 05, 2014	US ISM Non-Manufacturing PMI up to 54 in January from 53 (December)	Moderate

Table B Economic news indicators of 2014 (cont.). Source: FXStreet, 2014 [3].

Date	Description of Economic News Indicators	Volatility Level
Mar 03, 2014	US Personal Spending came in at 0.4% beating forecasts of 0.1% in January	Moderate
Mar 13, 2014	US Initial Jobless Claims in March 7: 315K	Moderate
Apr 01, 2014	US Markit Manufacturing PMI down to 55.5 in March from previous 57.1	Moderate
Apr 03, 2014	US Trade Balance registered at $-42.3B in February	Moderate
May 01, 2014	US Initial Jobless Claims above expectations(319K) in April 25: Actual (344K)	Moderate
May 27, 2014	US Durable Goods Orders ex Transportation above forecasts (0%) in April Actual (0.1%)	Moderate
May 29, 2014	US Initial Jobless Claims below expectations (318K) in May 23: Actual (300K)	Moderate
Jun 19, 2014	US Initial Jobless Claims below forecasts (314K) in June 13: Actual (312K)	Moderate
Jun 24, 2014	US Consumer Confidence above expectations (83.5) in June: Actual (85.2)	Moderate
Jul 03, 2014	US: ISM Non-Manufacturing PMI slides to 56 in June	Moderate
Jan 03, 2014	US December 27 EIA Natural Gas Storage change increase to -97B vs -177B	Low
Jan 07, 2014	US December 29 Redbook index (MoM); -0.6%; 4.1% (YoY)	Low
Jan 09, 2014	US January 3 EIA Natural Gas Storage change falls to -157B vs -97B	Low

Table B Economic news indicators of 2014 (cont.). Source: FXStreet, 2014 [3].

Date	Description of Economic News Indicators	Volatility Level
Jan 13, 2014	US sells $28 Bn in 3-month bills at 0.035%	Low
Feb 04, 2014	US sells $8Bn in 4-week bills at 0.130%	Low
Feb 13, 2014	US February 7 EIA Natural Gas Storage change up to -237B vs -262B	Low
Mar 03, 2014	US sells $25 Bn in 3-month bills at 0.050%	Low
Mar 07, 2014	US 3-Month Bill Auction remains at 0.05%	Low
Mar 12, 2014	US EIA Crude Oil Stocks change beat expectations (2.1M) in March 7: Actual (6.18M)	Low
Mar 31, 2014	US sells $25 Bn in 3-Month bills at 0.045%	Low
Apr 24, 2014	United States EIA Natural Gas Storage change registered at 49B to beat expectations (40B) in April 18	Low
May 02, 2014	United States ISM New York index down to 50.6 in April from previous 52	Low
May 13, 2014	United States Redbook index (MoM) climbed from previous 0% to 1.1% in May 9	Low
May 21, 2014	United States EIA Crude Oil Stocks change below expectations (-0.1M) in May 16: Actual (-7.226M)	Low
Jun 11, 2014	United States EIA Crude Oil Stocks change came in at -2.596M below forecasts (-1.3M) in June 6	Low

APPENDIX C
GOLD PRICES HISTORICAL DATA

Table C Historical data of gold price during Jan 2013 – Dec 2013.

Date	Last	Open	High	Low	Change (%)
Dec 31, 2013	1201.9	1196.4	1212.4	1182	-0.10%
Dec 30, 2013	1203.1	1215	1215.5	1194.4	-1.07%
Dec 27, 2013	1216.1	1213.4	1218.5	1212.9	0.16%
Dec 24, 2013	1205.1	1199.8	1205.6	1197.7	0.56%
Dec 23, 2013	1198.4	1205.2	1205.2	1195.9	-0.56%
Dec 20, 2013	1205.1	1190.9	1207.8	1190.9	0.85%
Dec 19, 2013	1195	1226.6	1226.6	1188.7	-3.32%
Dec 18, 2013	1236.1	1231.5	1245.1	1217	0.40%
Dec 17, 2013	1231.2	1237	1240.4	1231.2	-1.15%
Dec 16, 2013	1245.5	1238.4	1250.6	1228.9	0.79%
Dec 13, 2013	1235.7	1226.7	1239	1222.2	0.79%
Dec 12, 2013	1226	1253.5	1255.4	1225.9	-2.58%
Dec 11, 2013	1258.5	1261.9	1261.9	1252.1	-0.31%
Dec 09, 2013	1235.3	1230.2	1243.2	1229	0.41%
Dec 05, 2013	1233.2	1240	1241.5	1218.6	-1.20%
Dec 03, 2013	1221.7	1219.2	1225.7	1215.6	-0.05%
Dec 02, 2013	1222.3	1251.4	1251.5	1217.8	-2.26%
Nov 27, 2013	1237.8	1241.6	1254.8	1235.5	-0.29%
Nov 26, 2013	1241.4	1243	1243	1241.4	0.02%
Nov 25, 2013	1241.1	1231.1	1241.1	1229	-0.23%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
Nov 21, 2013	1243.5	1245.7	1247.1	1236.5	-1.14%
Nov 20, 2013	1257.9	1269	1269	1241.9	-1.22%
Nov 19, 2013	1273.4	1274.2	1275.2	1271	0.09%
Nov 15, 2013	1287.3	1287.7	1288.5	1281.5	0.09%
Nov 14, 2013	1286.2	1280.2	1293.3	1278.1	1.41%
Nov 13, 2013	1268.3	1274	1280.5	1268.3	-0.22%
Nov 12, 2013	1271.1	1282.3	1282.3	1268.5	-0.77%
Nov 07, 2013	1308.4	1323.6	1323.6	1298	-0.71%
Nov 06, 2013	1317.7	1313.5	1318.6	1312.5	0.74%
Nov 05, 2013	1308	1313.3	1317.5	1306.9	-0.50%
Nov 04, 2013	1314.6	1314.6	1320.5	1312.1	0.11%
Nov 01, 2013	1313.1	1325.7	1325.7	1308	-0.79%
Oct 31, 2013	1323.6	1334	1334	1321.7	-1.88%
Oct 30, 2013	1349	1340.8	1355.8	1335.6	0.28%
Oct 29, 2013	1345.2	1359.6	1359.7	1345.2	-0.50%
Oct 28, 2013	1352	1351.4	1359.2	1349	-0.03%
Oct 25, 2013	1352.4	1346.2	1353	1346.2	0.16%
Oct 24, 2013	1350.2	1334.5	1350.2	1334.5	1.22%
Oct 21, 2013	1315.7	1316.8	1317.7	1315.7	0.10%
Oct 17, 2013	1322.7	1307	1322.7	1307	3.17%
Oct 16, 2013	1282	1282.8	1284	1269.2	0.71%
Oct 15, 2013	1273	1270.5	1284.8	1254.1	-0.27%
Oct 08, 2013	1324.2	1326.5	1330	1321.2	-0.05%
Oct 07, 2013	1324.8	1313.8	1325.4	1311.8	1.15%
Oct 03, 2013	1317.4	1312	1319.5	1302.5	-0.24%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
Oct 01, 2013	1286	1328	1334.2	1283.8	-3.05%
Sep 30, 2013	1326.5	1345.8	1350.9	1322	-0.89%
Sep 26, 2013	1323.6	1335.2	1338.3	1323.6	-0.92%
Sep 25, 2013	1335.9	1322.7	1335.9	1322.7	1.51%
Sep 24, 2013	1316	1321.2	1324.9	1309.4	-0.82%
Sep 23, 2013	1326.9	1327.1	1327.8	1317.6	-0.42%
Sep 19, 2013	1369.4	1362.9	1373	1362.9	4.71%
Sep 18, 2013	1307.8	1292.1	1347.7	1292.1	-0.13%
Sep 17, 2013	1309.5	1313	1313.1	1309.5	-0.64%
Sep 16, 2013	1317.9	1328.4	1329.6	1310.8	0.73%
Sep 13, 2013	1308.4	1322.9	1323	1305.3	-1.65%
Sep 12, 2013	1330.4	1360	1360	1322.4	-2.46%
Sep 11, 2013	1363.9	1365.8	1365.8	1356.7	-0.01%
Sep 09, 2013	1386.8	1389.8	1389.8	1386.5	0.01%
Sep 05, 2013	1373.1	1393.5	1397.5	1367.1	-1.21%
Sep 04, 2013	1389.9	1413	1413	1387.1	-1.57%
Sep 03, 2013	1412	1393	1414.4	1382.3	1.31%
Aug 30, 2013	1396.1	1407.2	1409.8	1392.3	-1.19%
Aug 29, 2013	1412.9	1417.4	1417.4	1404.5	-0.43%
Aug 28, 2013	1419	1417.3	1428	1415.1	-0.11%
Aug 27, 2013	1420.6	1403.8	1421.1	1398.7	1.98%
Aug 26, 2013	1393	1398.6	1403	1391.3	-0.19%
Aug 22, 2013	1371.2	1362	1380.4	1362	0.04%
Aug 20, 2013	1373.1	1355	1377	1355	0.51%
Aug 19, 2013	1366.2	1379	1384.4	1363.9	-0.40%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
Aug 16, 2013	1371.7	1366.8	1378.9	1361	0.74%
Aug 15, 2013	1361.6	1340.1	1369.6	1319.9	2.07%
Aug 14, 2013	1334	1317.7	1336.3	1317.7	0.97%
Aug 13, 2013	1321.2	1331.9	1339.1	1320.2	-1.01%
Aug 12, 2013	1334.7	1317.2	1342.4	1317.2	1.66%
Aug 08, 2013	1310.7	1283.9	1312.8	1283.7	1.91%
Aug 07, 2013	1286.1	1282.5	1288.4	1274	0.23%
Aug 06, 2013	1283.2	1303	1305.9	1279.3	-1.49%
Aug 05, 2013	1302.6	1311.2	1318.3	1297.5	-0.61%
Aug 02, 2013	1310.6	1306.7	1317.6	1283	-0.03%
Aug 01, 2013	1311	1323	1328.5	1306.9	-0.11%
Jul 31, 2013	1312.4	1325.5	1338.4	1304.9	-0.88%
Jul 30, 2013	1324	1327.2	1329.6	1316.2	-0.19%
Jul 29, 2013	1326.55	1330.95	1338.45	1323.15	0.37%
Jul 25, 2013	1329	1318.5	1336.8	1318.5	0.70%
Jul 24, 2013	1319.7	1342.2	1345.1	1318.6	-1.15%
Jul 23, 2013	1335.1	1334.8	1342.3	1327.9	-0.10%
Jul 18, 2013	1284.6	1275.3	1286.4	1275.2	0.52%
Jul 17, 2013	1277.9	1286.2	1300	1272.9	-1.00%
Jul 16, 2013	1290.8	1279.4	1291.4	1279.4	0.55%
Jul 15, 2013	1283.8	1292.1	1292.4	1281.6	0.47%
Jul 12, 2013	1277.8	1282.3	1282.3	1273.3	-0.18%
Jul 11, 2013	1280.1	1284.6	1285	1278.9	2.62%
Jul 10, 2013	1247.4	1244.6	1263.9	1244.6	0.12%
Jul 09, 2013	1245.9	1234.7	1254.5	1234.4	0.89%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
Jul 08, 2013	1234.9	1218.8	1235.2	1215.4	1.81%
Jul 03, 2013	1252.1	1243.8	1258.2	1240.3	0.68%
Jul 02, 2013	1243.6	1254.1	1266.3	1241.2	-0.98%
Jun 28, 2013	1223.8	1200.9	1232.3	1183.2	1.02%
Jun 27, 2013	1211.4	1224.7	1241.6	1198.3	-1.48%
Jun 26, 2013	1229.6	1273.9	1273.9	1228.8	-3.55%
Jun 25, 2013	1274.8	1281.3	1286	1273.2	-0.16%
Jun 24, 2013	1276.8	1297.3	1297.3	1275.7	-1.15%
Jun 18, 2013	1366.6	1383.1	1383.5	1363	-1.17%
Jun 17, 2013	1382.8	1390.3	1390.5	1380	-0.32%
Jun 14, 2013	1387.3	1383.9	1389.6	1377.8	0.70%
Jun 13, 2013	1377.6	1386.8	1390.9	1373.4	-1.02%
Jun 12, 2013	1391.8	1375	1393.5	1373.7	1.07%
Jun 10, 2013	1386.2	1378.8	1386.9	1375.6	0.23%
Jun 06, 2013	1415.7	1401.7	1422.7	1391.6	1.24%
Jun 05, 2013	1398.4	1397.5	1410	1395.6	0.09%
Jun 04, 2013	1397.1	1410.9	1414.1	1389	-1.03%
Jun 03, 2013	1411.7	1389.1	1416.3	1388.3	1.37%
May 31, 2013	1392.6	1412.4	1421.1	1384.2	-1.34%
May 30, 2013	1411.5	1392.1	1417.5	1388.3	1.39%
May 29, 2013	1392.15	1380.15	1394.95	1379.85	0.95%
May 28, 2013	1379.1	1380.8	1392.5	1379.1	-1.21%
May 24, 2013	1385.25	1387.05	1396.85	1381.25	-0.48%
May 23, 2013	1392	1380.6	1394.8	1377.8	1.78%
May 21, 2013	1377.8	1386.5	1392.3	1358.5	-0.47%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
May 20, 2013	1384.3	1342.9	1391.9	1342.9	1.42%
May 16, 2013	1387.1	1390.4	1395.1	1370.6	-0.67%
May 15, 2013	1396.5	1425.7	1425.7	1390.1	-1.98%
May 14, 2013	1424.7	1430.1	1444.4	1422	-0.68%
May 13, 2013	1434.5	1440	1440.6	1427.8	-0.16%
May 09, 2013	1468.8	1471.2	1473.4	1453	-0.35%
May 08, 2013	1473.9	1455.5	1474.7	1450.7	1.72%
May 07, 2013	1449	1462	1466.2	1441.2	-1.30%
May 06, 2013	1468.1	1468.1	1477.2	1466.8	0.26%
May 03, 2013	1464.3	1465.5	1487.1	1461.4	-0.23%
May 02, 2013	1467.7	1459	1471.2	1453.5	1.48%
May 01, 2013	1446.3	1475.8	1477.1	1440	-1.76%
Apr 30, 2013	1472.2	1471.1	1476.6	1461	0.33%
Apr 29, 2013	1467.4	1462.7	1475.5	1462.3	0.95%
Apr 26, 2013	1453.6	1467.4	1480	1451.4	-0.56%
Apr 25, 2013	1461.8	1433	1461.8	1433	2.70%
Apr 24, 2013	1423.4	1423.2	1430.1	1420.4	1.05%
Apr 23, 2013	1408.6	1429.5	1429.5	1405.8	-0.87%
Apr 22, 2013	1421	1406.8	1434.9	1406.8	1.84%
Apr 18, 2013	1392	1347.5	1396.4	1347.5	0.71%
Apr 16, 2013	1386.8	1358	1401	1323	1.93%
Apr 12, 2013	1501	1560.8	1560.8	1480.2	-4.05%
Apr 11, 2013	1564.3	1557.1	1566.7	1554.8	0.39%
Apr 10, 2013	1558.3	1584.4	1586	1557	-1.76%
Apr 09, 2013	1586.2	1571.6	1589.3	1570	0.90%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
Apr 08, 2013	1572	1575.8	1580.6	1567.6	-0.22%
Apr 04, 2013	1551.8	1557.3	1557.3	1539.1	-0.06%
Apr 03, 2013	1552.8	1575.7	1576.4	1549.2	-1.42%
Apr 02, 2013	1575.1	1599.2	1602.6	1574.1	-1.56%
Apr 01, 2013	1600	1596.8	1600.5	1594.8	0.33%
Mar 28, 2013	1594.8	1604.7	1606.6	1593.6	-0.71%
Mar 27, 2013	1606.2	1598.5	1608.1	1590.3	0.65%
Mar 26, 2013	1595.8	1595.8	1595.8	1595.8	-0.55%
Mar 25, 2013	1604.6	1607.1	1607.8	1596.1	-0.10%
Mar 21, 2013	1613.8	1612.8	1613.8	1611.9	0.39%
Mar 19, 2013	1611.3	1613.1	1613.1	1611.3	0.42%
Mar 18, 2013	1604.6	1605.9	1607.7	1601.6	0.76%
Mar 15, 2013	1592.5	1590.5	1594.4	1590.2	0.12%
Mar 14, 2013	1590.6	1585.7	1590.6	1575.7	0.14%
Mar 13, 2013	1588.3	1592.3	1596	1585	-0.20%
Mar 12, 2013	1591.5	1579.8	1595.6	1579.8	0.87%
Mar 11, 2013	1577.8	1576	1579.7	1576	0.08%
Mar 07, 2013	1574.8	1580	1583.5	1573.9	0.01%
Mar 06, 2013	1574.6	1577.6	1582.6	1566.8	0.00%
Mar 05, 2013	1574.6	1577.6	1585.5	1571.4	0.16%
Mar 04, 2013	1572.1	1580.8	1580.8	1569.7	0.01%
Mar 01, 2013	1571.9	1579.6	1585.6	1566	-0.37%
Feb 28, 2013	1577.7	1596	1601.6	1575.1	-1.10%
Feb 27, 2013	1595.2	1612.8	1612.8	1591.8	-1.24%
Feb 26, 2013	1615.2	1593.3	1615.2	1589	1.83%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
Feb 25, 2013	1586.2	1579.4	1596	1577.7	0.88%
Feb 21, 2013	1578.2	1568.5	1581.2	1554.5	0.04%
Feb 20, 2013	1577.6	1604.7	1606.9	1561.8	-1.62%
Feb 19, 2013	1603.6	1610	1617	1599.7	-0.32%
Feb 15, 2013	1608.8	1633.2	1633.9	1599.5	-1.58%
Feb 14, 2013	1634.7	1641.1	1646.9	1632.7	-0.58%
Feb 13, 2013	1644.2	1652.2	1652.3	1640.2	-0.27%
Feb 12, 2013	1648.7	1643.8	1650.4	1642.4	0.03%
Feb 11, 2013	1648.2	1666.5	1668	1643.7	-1.07%
Feb 07, 2013	1670.4	1677	1682.1	1663.1	-0.44%
Feb 05, 2013	1672.4	1674.2	1683.4	1666.5	-0.17%
Feb 04, 2013	1675.3	1668	1676.5	1661.6	0.35%
Feb 01, 2013	1669.4	1663.9	1681.2	1659.9	0.53%
Jan 31, 2013	1660.6	1675.4	1680	1657.4	-1.15%
Jan 30, 2013	1679.9	1662.4	1683.2	1661.8	1.16%
Jan 29, 2013	1660.7	1661.2	1661.5	1660.7	0.50%
Jan 28, 2013	1652.4	1651.6	1655.5	1651.6	-0.24%
Jan 24, 2013	1669.5	1680.8	1681.7	1668	-1.00%
Jan 23, 2013	1686.3	1691.3	1693.7	1685.2	-0.38%
Jan 22, 2013	1692.8	1690.3	1694.7	1687.8	0.37%
Jan 17, 2013	1690.4	1676.4	1694.6	1671	0.46%
Jan 16, 2013	1682.7	1674.4	1682.7	1674.4	-0.04%
Jan 15, 2013	1683.4	1671.6	1684.3	1671.6	0.87%
Jan 14, 2013	1668.9	1665.4	1673.8	1664.1	0.54%
Jan 10, 2013	1677.3	1672.5	1677.3	1672.5	1.36%

Table C Historical data of gold price during Jan 2013 – Dec 2013 (cont.).

Date	Last	Open	High	Low	Change (%)
Jan 09, 2013	1654.8	1658.6	1662.1	1651.2	-0.40%
Jan 08, 2013	1661.5	1647.7	1661.5	1647.7	0.97%
Jan 07, 2013	1645.5	1656.5	1659.9	1643.8	-0.16%
Jan 04, 2013	1648.1	1647	1658.3	1625.7	-1.53%
Jan 03, 2013	1673.7	1686.1	1686.8	1662	-0.84%
Jan 02, 2013	1687.9	1672.8	1693.8	1670	0.78%

APPENDIX D

TECHNICAL PAPER OF 2015 INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY

Gold Price Volatility Prediction by Text Mining in Economic Indicators News

Chanwit Onsumran, Sotarat Thammaboosadee and Supaporn Kiattisin
Technology of Information System Management Division, Faculty of Engineering, Mahidol University, Nakhon Pathom, 73170, Thailand

Abstract

This paper focuses on the text mining approach of the gold prices volatility prediction model from the textual of economic indicators news articles. The model is designed and developed to analyze how the news articles influence gold price volatility. The selected reliable source of news articles is provided by FXStreet which offers several economic indicators such as Economic Activity, Markit Manufacturing PMI, Bill Auction, Building Permits, ISM Manufacturing Index, Redbook index, Retail Sales, Durable Goods Orders, etc. The data will be used to build text classifiers and news group affecting volatility price of gold. According to the fundamental of data mining process, each news article is firstly transformed in to feature by TF-IDF method. Then, the comparative experiment is set up to measure the accuracy of combination of two attributes weighting approaches, which are Support Vector Machine (SVM) and Chi-Squared Statistic, and three classification algorithms, which are the k-Nearest Neighbour, SVM and Naive Bayes. The results show that the SVM method is the most superior to other methods in both attributes weighting and classifier viewpoint.

Keywords: text mining, economic indicators news, gold price, volatility prediction.

1. Introduction

Gold transaction and investment is now popular than ever which leads gold prices fluctuation is compatible with changing volume investment and speculation. The major influent factor which is used to determine the price of gold is the USD (United States Dollar) currency [1]. If the other factors are stable, the gold prices will increase when USD depreciates as gold provides a hedge against a lower USD. Since the USD is the international main currency, when it depreciates, central banks of many USD countries reserve spread their risk by investing in other assets, such as gold. This occurrence usually pushes up the gold price. Another major factor is the USD inflation rate. The gold prices increases when inflation rates are higher. Gold prices often increase during time of international political tension or the period of world monetary system becomes less stable. During these periods, assets are usually sold and gold is

bought as asset which causes its prices may drop during the times of instability or crisis. Additionally, demand and supply in the market are significant. If other factors are stable gold prices will increase when demand for gold is higher than the supply in the market.

Economic news articles are also an indicator for gold price volatility apart of another factor as stated [1]. Anyway, the economic indicators news is represented in text format. So the text mining techniques [2] is appropriate for analysis in this research. The text mining methodology has been very popular at present since, more than 90% of the volume of data on the internet is unstructured [2]. The large amounts of useful information are often hidden, such as news articles and economic change and vary according to the circumstances and events occurred at different times.

According to the statement of problems and technology described above, this paper focused matching news group and predict gold price, as known as spot gold, volatility of economic indicators news article by text mining techniques.

2. Backgrounds and related works

This section provides the background theory on data mining and the related works to this paper.

2.1 Data Mining

Data mining [3] is a process dealing with large amounts of data to find patterns and relationships hidden in the data set. At the present, data mining has been applied in various applications, for examples, business decisions made by executives, science and medicine, as well as the economic and social development. Data mining is like the evolution of the collection and interpretation of data from the existing simple storage into a database that can be used to retrieve information from data mining to discover knowledge hidden in the data.

The data mining process contains sub-workflows, for transforming data into knowledge, consists of the following steps [3]:

Step1: Data Cleaning: screening out irrelevant information.

Step2: Data Integration: combining multiple data sources into a data set.

Step3: Data Selection: retrieving information from sources that are recorded for analysis.

Step4: Data Transformation: converting data to be suitable for use.

Step5: Modeling: searching for patterns that benefit from the existing data.

Step6: Evaluation: evaluating forms of data mining.

Step7: Knowledge Representation: knowledge discovery using the techniques presented for understanding.

2.2 Text Mining

Text mining [4], or it may be called: knowledge discovery in document databases, is a technique to discover patterns of enormous amounts of text

automatically by using the data mining algorithm. The text mining, a process operates with textual data, is to find the patterns and relationships hidden in the text. Recognition based on statistical machine learning, document processing, text processing and the natural language processing.

2.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) [5] is algorithm that can be used to help solve the problem of data analysis and classification. In brief, its principle is to create a separate set of data that is entered into the system taught to learn by focusing on the best dividing line to distinguish data.

2.4 K-Nearest Neighbour (K-NN)

K-Nearest Neighbor [6], illustrated in Figure 4, is algorithm for data classification by the concept of data grouping based on the information that is mostly close to the value of the information. If classification using k groups, determined by Euclidean distance [6], it is called the k-NN (k Nearest Neighbor).

2.5 Naïve Bayes

Naive Bayes model [7] is the separation of data using probability, which is based on Bayes' Theorem [7] and the assumption that the occurrence of events independent. Bayes' theorem can be written as shown in equation 1.

$$P(T|E) = \frac{P(E|T) \times P(T)}{P(E|T) \times P(T) + P(E|\neg T) \times P(\neg T)} \quad (1)$$

2.6 Feature Selection

Feature selection [8] is the process of selection subset of the terms occurring in the training set and using only this subset as features in text classification.

Weighting by SVM [9] uses the coefficients of the normal vector of a linear SVM as attribute weights. The attribute values still have to be numerical. This operator can be applied only on example sets with numerical label. This operator calculates the relevance of the attributes by computing for each attribute of the input for the weight with respect to the class attribute. The coefficients of a hyperplane calculated by an SVM are set as attribute weights.

Weighting by Chi Squared Statistic [10] calculates the weight of attributes with respect to the class attribute by using the chi-squared statistic. The higher the weight of an attribute, the more relevant it is considered. The chi-squared statistic can only be calculated for nominal labels. This operator calculates the relevance of the attributes by computing for each attribute of the input Example Set the value of the chi-squared statistic with respect to the class attribute.

2.7 TF-IDF Vector Space

The term frequency–inverse document frequency (TF-IDF) [11] takes into account the frequency of a term (word) in a document and all the other documents in the set to calculate a metric of importance of a term in a document relative to all the other words and documents.

2.8 Related works

Samuel J. Rivera et al. proposed a text mining framework for advancing sustainability indicators [12]. This study applies document classification algorithm and the incorporation of several information retrieval techniques, the analysis demonstrated that mining the growing amount of digitized news media can provide useful information for identifying, tracking, and reporting sustainability indicators. This work is similar to the proposed research in the terms of pre-processing and methods transformation of unstructured textual data and algorithm K-NN.

Arman et al. proposed the text mining of news-headlines for FOREX market prediction [13]. This work context by bringing together natural language processing and statistical pattern recognition as well as sentiment analysis to propose a system that predicts directional-movement of a currency-pair in the foreign exchange market based on the words used in adjacent news-headlines. The system succeeds in doing so with an accuracy level of 83.33% in some cases. This work is similar to the proposed research in term of pre-processing and news-mapping (label assignment).

Michael Hagenau proposed automatic news reading, cases study on stock price prediction based on financial news using context-capturing features [14]. In summary, research shows that the combination of advanced feature extraction methods and feedback-based feature selection boosts classification accuracy and allows improved sentiment analytics.

3. Methodology

The data used in this research is economic indicators news articles and gold price historical data. The label of economic indicators news article assigned to each groups is volatility of gold price. Economic indicators news articles reference by FXStreet data [15] in the period of 1st January 2013 to 31st December 2013 forms the study news group economic for classification and the spot gold reference by investing website. This experiment totally as 201 sets.

3.1 Steps of Methodology

An overall methodology, as shown in Figure 1, is summarized as follow:

Textual data of economic indicators news article will be processed by typical pre-processing methods from unstructured textual data to a word-document matrix. It includes the tasks of tokenization and filtering stop words, and the calculation of an importance metric. Tokenization is the process of splitting the

text into individual words or tokens. Tokenization within the English language is often done by using blank spaces and punctuation marks as token delimiters. Next, the stemming is a process of linguistic normalization, in which the variant forms of a word are reduced to a common form, for example, "connection", "connective", "connected" and "connecting" have a common word as "connect". After process of stemming, word list data will be transformed by normalization process and select attributes with feature selection by attributes weighting by SVM and attributes weighting by Chi Squared Statistic which were described in the previous section.

Consequently, the economic indicators news has been processed by mapping gold price historical transformed data, as known as spot gold, and assigning the label. The final pre-processing step involves converting the occurrence of words within a document to a metric that represents its relative importance. In this study the binary representation of the occurrence of a word and the term frequency-inverse document frequency (TF-IDF) [11] were used as metrics of importance. The binary representation is computed by assigning a 1 if a word is present in the document and 0 otherwise. This metric is used to filter out words that only appear in one document in the set, a necessary step for the correct implementation of the generalized discriminant analysis. The TF-IDF computes the frequency of a word in a document and all the other documents in the set as a metric of the importance of a word in a document relative to all of the other words and documents.

Then, the model will be created to predict gold price volatility by SVM, K-NN and Naive Bayes methods. Model evaluation by 10-Fold Cross-validation [4] the predictive ability of the model examples. The basis of this technique is the re-sampling by the start of the series divided into sections called fold and some of data set to test the results of test data and the model prediction.

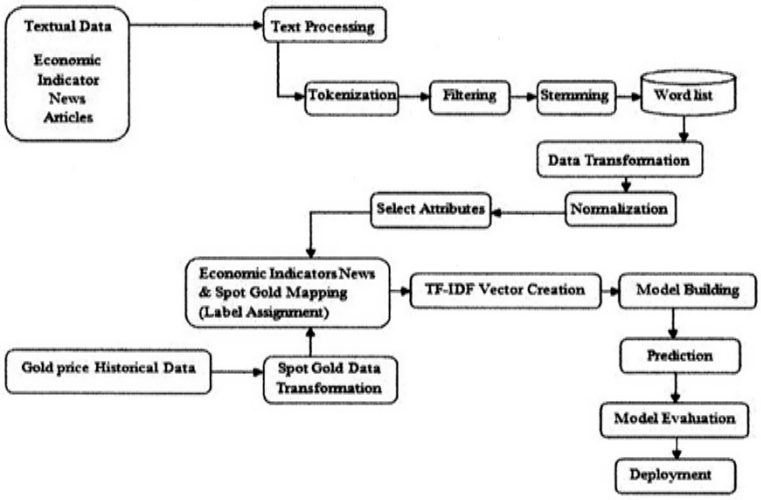

Fig.1: Methodology overview

3.2 Economic indicators news articles

Economic indicators news articles reference by FXStreet [15]. Examples of news article and its labeled volatility are shown in Table 1.

Table 1: Example of economic indicators news articles

Date	Name	Description	Volatility
Dec 31, 2013	Redbook index	US Redbook index up to -0.7% from -1%	Low volatility
Dec 24, 2013	Durable Goods Orders	US durable goods orders Nov +3.5% vs +2.0%	Moderate volatility
Dec 19, 2013	Fed Interest Rate Decision	The Fed leaves interest rate unchanged at 0.25%	High volatility

4. Experimental results

This study using example set of 201 examples were divided into three label by include High volatility expected, Moderate volatility expected and Low volatility expected, as shown in Figure 2.

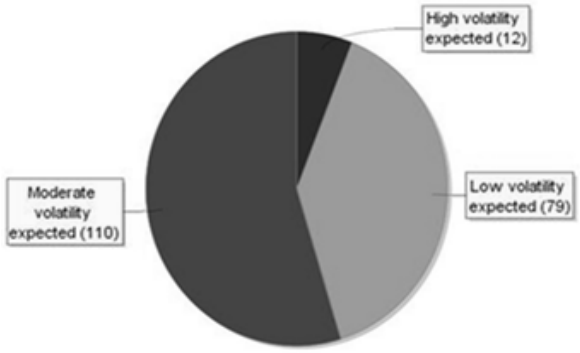

Fig.2: Data set summarization

In all experimental schemes, consists of three classification algorithms and two features selection methods. For summary, the comparative results of six experiments are shown in Table 2 and feature selection methods for the attribute weighting by SVM is shown in Figure 3 and the feature selection methods for attribute weighting by Chi Squared Statistic is shown in Figure 4.

Figure 3 and Figure 4 shows the selected features weighted by SVM and Chi Squared Statistic respectively with the selection criteria with top P percent. The top P percent attributes with highest weights are selected. In this paper, the top P percent by P equals 0.5. The SVM can filter six features and eight features by Chi Squared Statistic.

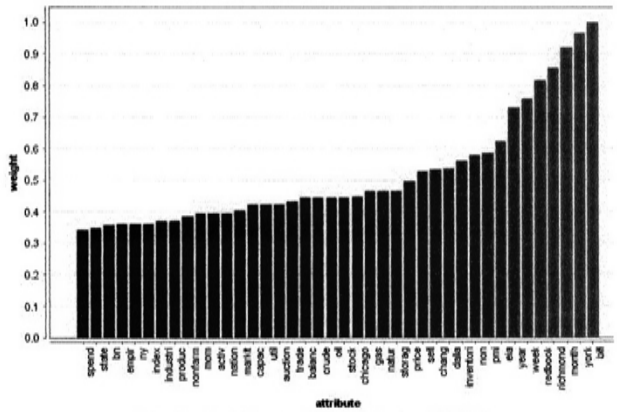

Fig.3: Attribute Weights by SVM

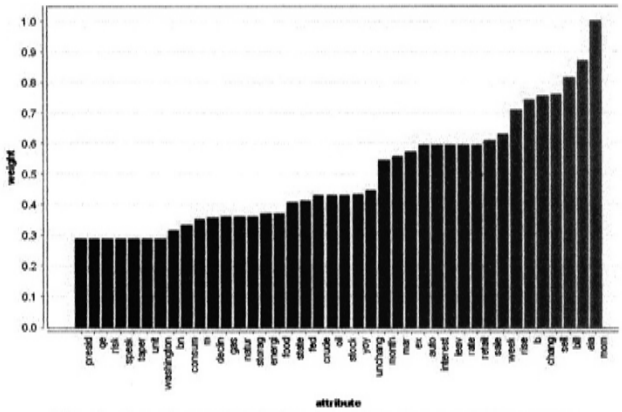

Fig.4: Attribute Weights by Chi Squared Statistic

Table 2: Comparative results

Classification Methods	Feature Selection Methods	Accuracy
SVM	SVM Weighting	87.52%
	Chi-Squared statistic Weighting	86.55%
K-NN	SVM Weighting	85.57%
	Chi-Squared statistic Weighting	86.57%
Naive Bayes	SVM Weighting	76.60%
	Chi-Squared statistic Weighting	82.10%

Table 2 shows the classification results of six experiment schemes. It is found that the SVM classification algorithm, weighted by SVM is the best among all tests with accuracy as 87.52%. In all methods SVM can handle the complicate data better than the other methods in both contexts of feature selection and classification.

5. Conclusion and future works

This work presents the framework for using text mining to create model for predict gold price volatility. The framework studies factors of economic indicators news article. Typical pre-processing methods used for the transformation will have procedure normalization and select attributes with classification algorithm for measuring the result. Compare the effectiveness of SVM, K-NN and Naive Bayes algorithm for prediction. The best in all tests is support vector machine algorithm, weight by SVM with accuracy as 87.52%.

To improve the classification system, it may be input other economic news more factors to increase efficiency analyze consequences that affect the volatility of the gold price. An experiment of some of other algorithms and feature selection methods to get the higher accuracy is also challenge.

References

[1] Shaun,K. & Marco,R., The Effects of Economic News on Commodity Prices Is Gold Just Another Commodity?. International Monetary Fund, 2009.

[2] Hsien,T.,Y., Lin,C., J. & Lin,Y.,I., Text mining techniques for patent analysis. Information Processing & Management, pp.1216-1247, 2007.

[3] Han,J. & Kamber,M., Data mining concepts and techniques. United States of America: Morgan Kaufman Publishers, 2006.

[4] Feldman,R. & Sanger,J., The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.

[5] Ustun,B., Support Vector Machine. [Online]. Access 12 October 2014. Available from http://www.cac.science.ru.nl/people/ustun

[6] Duda,R.,O., Hart,P.,E. & Stork,D.,G., Pattern Classification. New York: Wiley-Interscience Publication, 2001.

[7] Zhang,H., The Optimality of Naive Bayes. *FLAIRS conference*, 2004.

[8] Nguyen,H., Franke,K. & Petrovic,S., Towards a Generic Feature-Selection Measure for Intrusion Detection. *International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010.

[9] Wang,M., Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Engineering Design and Selection, pp.509-516, 2004.

[10] Holt,D., Scott,A.,J. & Ewings,P.,D., Chi-squared tests with survey data. *Journal of the Royal Statistical Society*, pp.303-320, 1980.

[11] Luk,W.,H., Wong,K.,F. & Kwok,K.,L., Interpreting tf–idf term weights as making relevance decisions. ACM Transactions on Information Systems, pp.1–37, 2008.

[12] Samuel,J., A text mining framework for advancing sustainability indicators. Environmental Modelling & Software, pp.128-138, 2014.

[13] Arman,K., Text Mining of News-Headlines for FOREX Market Prediction: A Multi-layer Dimension Reduction Algorithm with Semantics & Sentiment. Expert Systems with Applications, 2014.

[14] Liebmann,M. & Neumann,D., Automated news reading: Stock price prediction based on financial news using context-capturing features. Decision Support Systems, pp.685-697, 2013.

[15] FXStreet, Economic indicators articles. [Online]. Access 2 October 2014. Available from http://www.fxstreet.com

BIOGRAPHY

NAME	Mr. Chanwit Onsumran
DATE OF BIRTH	9 April 1991
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Suan Dusit Rajabhat University, 2009-2012 Bachelor of Science Second Class Honours (Information Technology) Mahidol University, 2013-2015 Master of Science (Information Technology Management)
RESEARCH GRANTS	Grant to Support Graduate Students in Academic Presentations in Singapore Academic Year 2015
HOME ADDRESS	61/5 Moo 3 Bangkruai-Jongthanom Road, Maha Sawat, Bangkruai, Nonthaburi, Thailand, 11130 Tel. 090-992-6455 E-mail: chanwit_earth@hotmail.com
PUBLICATION / PRESENTATION	Onsumran, C., Thammaboosadee, S., and Kiattisin, S. (2015), Gold Price Volatility Prediction by Text Mining in Economic Indicators News, The Proceedings of the 2015 International Conference on Information Technology (ICIT 2015), pp. 305-312, ISSN: 1743-3517.