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, Thailand
Email: chanwit_earth@hotmail.com, {sotarat.tha, supaporn.kit}@mahidol.ac.th

***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.**

***Index Terms*—text mining, economic indicators news, gold price, volatility prediction**

I. 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.

II. Backgrounds and Related Works

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

A. 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.

Manuscript received April 27, 2015; revised July 20, 2015.

This work was supported by the Faculty of Graduate Studies, Mahidol University grant to support graduate students in academic presentations in Singapore.

doi: 10.12720/jait.6.4.243-247

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.

B. 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.

C. 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.

D. K-Nearest Neighbour (K-NN)

K-Nearest Neighbor [6], illustrated in Fig. 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).

E. Na ïve 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.

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

F. 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.

G. 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.

H. 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.

III. 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.

A. Steps of Methodology

An overall methodology, as shown in Fig. 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.

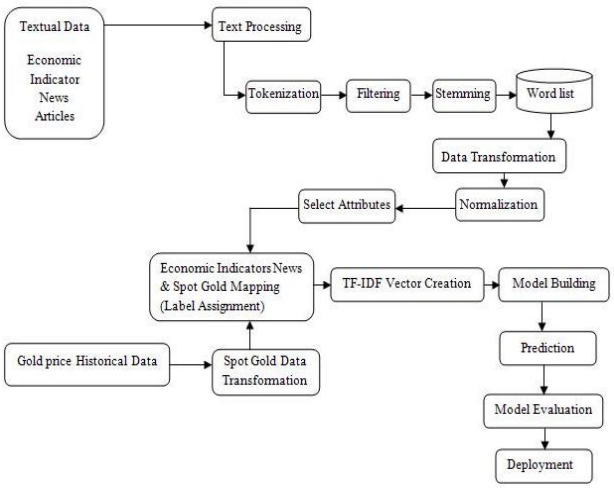

Figure 1. Methodology overview.

B. Economic Indicators News Articles

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

TABLE I. 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

IV. 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 Fig. 2.

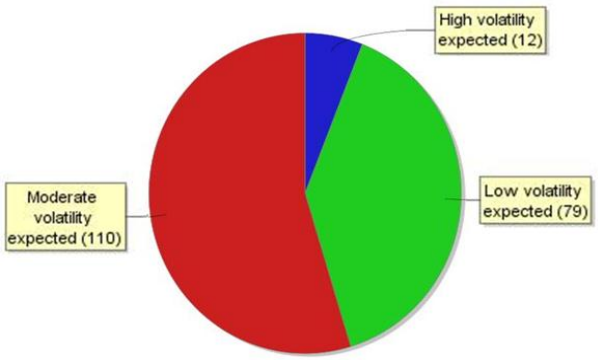

Figure 2. Data set summarization.

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 II and feature selection methods for the attribute weighting by SVM is shown in Fig. 3 and the feature selection methods for attribute weighting by Chi Squared Statistic is shown in Fig. 4.

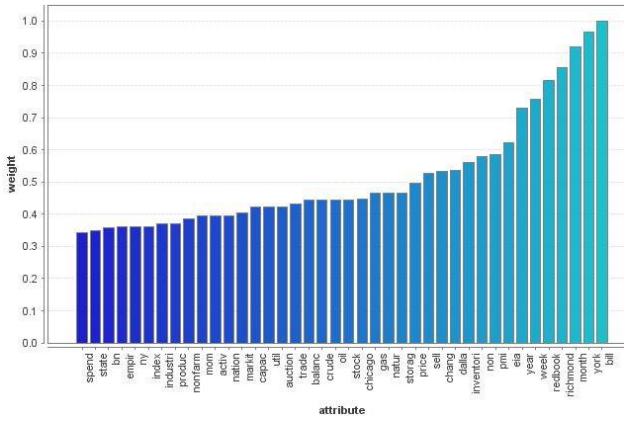

Figure 3. Attribute weights by SVM.

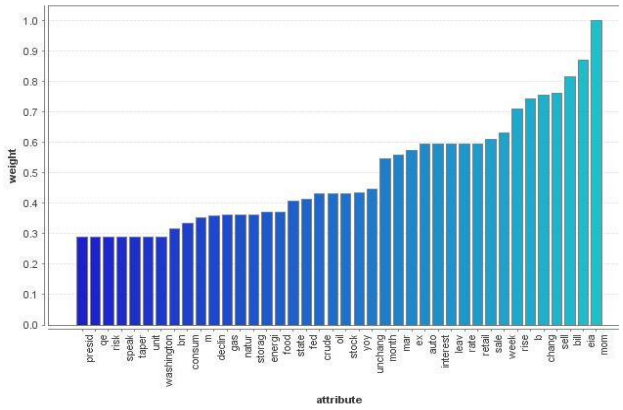

Figure 4. Attribute weights by Chi squared statistic.

Fig. 3 and Fig. 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.

TABLE II. 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 II 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.

V. 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] S. K. Roache and M. M. Rossi, *The Effects of Economic News on Commodity Prices: Is Gold Just Another Commodity?* International Monetary Fund, 2009.
[2] Y. H. Tseng, C. J. Lin, and Y. I. Lin, "Text mining techniques for patent analysis," *Information Processing & Management*, vol. 43, no. 5, pp. 1216-1247, 2007.
[3] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, United States of America: Morgan Kaufman Publishers, 2006.
[4] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, 2007.
[5] B. Ustun. Support Vector Machine. [Online]. Available: http://www.cac.science.ru.nl/people/ustun
[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, New York: Wiley-Interscience Publication, 2001.
[7] H. Zhang, *The Optimality of Naive Bayes*, FLAIRS Conference, 2004.
[8] H. Nguyen, K. Franke, and S. Petrovic, "Towards a generic feature-selection measure for intrusion detection," in *Proc. International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 1529-1532.
[9] M. Wang, *et al*., "Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition," *Protein Engineering Design and Selection*, vol. 17, no. 6, pp. 509-516, 2004.
[10] D. Holt, A. J. Scott, and P. D. Ewings, "Chi-squared tests with survey data," *Journal of the Royal Statistical Society*, pp. 303-320, 1980.
[11] H. C. Wu, R. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf–idf term weights as making relevance decisions," *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1–37, 2008.
[12] S. J. Rivera, *et al*., "A text mining framework for advancing sustainability indicators," *Environmental Modelling & Software*, vol. 62, pp. 128-138, 2014.
[13] A. K. Nassirtoussi, *et al*., Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics & sentiment," *Expert Systems with Applications*, vol. 42, no. 1, pp. 306-324, 2014.
[14] M. Liebmann and D. Neumann, "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

Chanwit Onsumran was born in Bangkok, Thailand in April 1991. He received his B.Sc. (Second-class honors) degree in Information Technology from Suan Dusit Rajabhat University, Thailand in 2012 and M.Sc. degree in Information Technology Management from Mahidol University, Thailand in 2015. His research interests are knowledge discovery, data mining in economic domain and text mining.

Sotarat Thammaboosadee was born in 1982 and received his B.Eng. and M.Sc. degrees in Computer Engineering and Technology of Information System Management from Mahidol University, Thailand, in 2003 and 2005 respectively. He also received a Ph.D. in Information Technology from King Mongkut's University Technology Thonburi, Thailand, in 2013. He is now a lecturer at Technology of Information System Management Division at Faculty of Engineering, Mahidol University, Thailand. His research interests include data mining in several domains such as the legal domain, healthcare domain and management domain. He also interests the research filed of technology valuation and business process improvement in CRM section.

Supaporn Kiattisin received her B.Eng. and M.Eng. degrees in Computer Engineering and Electrical Engineering from Chiang Mai University, Thailand, in 1996 and King Mongkut's University Technology Thonburi, Thailand, in 2000 respectively. She also received a Ph.D. in Electrical and Computer Engineering from King Mongkut's University Technology Thonburi, Thailand, in 2007. She is now a vice-director at Technology of Information System Management Division and program chair of Information Technology Management program at Faculty of Engineering, Mahidol University, Thailand. Her research interests include enterprise architecture, risk management, business process improvement, customer relationship management, and human resource management.