Stock Market Prediction Based on Term Frequency-Inverse Document Frequency

Mu'tasem Jarrah and Naomie Salim

***Abstract*—This paper presents a new method to predicting the change of stock prices by utilizing text mining news of the stock market. Term Frequency Inverse Document Frequency (TF- IDF) is one of the most useful and widely used concepts in information retrieval. The method can handle without difficulty unstructured news of Saudi stock market (Tadawul) through reading and analysis of the news and build a relationship between the contents of the news, and keywords (core phrases). This technique must identified by analysts and financial specialists that affect the direction of the share price up or down. The aim of this paper is to explore the possibility of using text mining to automate the identification of financial news articles. The empirical results show that the proposed techniques can predict the up and down on a stock price after the news announce or released. The proposed method presented in the study is straightforward, simple and valuable for the short-term investors.**

***Index Terms*—Stock market prediction, term frequency inverse document frequency (TF-IDF), text mining.**

I. Introduction

Investment in the stock market is one of the encouraging ways to get high rewards; on the other hand, it is also a risk for many investments. This paper examines the reaction of company news and events on stock market prices [1]. Different factors and events are generated within the company during the financial year, such as dividend announcements by the company's during the meeting, changes in the board of directors, new investments, layoffs, market scandals, firing of CEO or company officials, ect. These factors and events are treated as signals that are emitted by the company's information office and sometimes by the insider information in an unsure situation characterized by informational irregularity and interpreted as an effective and reliable by investors. The relationship between the stock market and financial news exists, nowadays, many news articles available on the Internet, and these articles are a type of unstructured data. In addition, the information contained in the news in a certain time period affect to a large degree on stock prices as well as the behavior of the market. [2], where this news rating to "good news", "normal" and "bad news" and its impacts on changes happen to the share price after the announcement of this news. [3].

The paper is organized as follows. Literature review and related works are explained in Section II. Section III discusses the term frequency and inverse term frequency (*tf-idf*). Results are presented in Section IV. Finally, the paper concludes in Section V.

Manuscript received August 8, 2014; revised July 29, 2015. This work was supported by Ministry of Higher Education (MOHE) and Research Management Center (RMC) at the Universiti Teknologi Makaysia (UTM) under Research Universiy Grant Category (VOT: R.J130000.7828.4F373).

M. Jarrah is with the King Abdulaziz University, Information Technology Department, Faculty of Computing and Information Technology, Jeddah, Saudi Arabia, and also with Universiti Teknologi Malaysia, Faculty of Computing, Johor Bahru, Malaysia (e-mail: mmjarrah@kau.edu.sa).

N. Salim is with the Universiti Teknologi Malaysia, Faculty of Computing, Johor Bahru, Malaysia (e-mail: naomie@utm.my).

II. Related Work

This paper explores the effect of news and events on a company's stock prices, the majority algorithms used in automatic text categorization (ATC) are universal from data mining applications. The data analyzed by data mining are numeric, which means they are already in the format necessary by the algorithms. These algorithms can be useful in ATC, but first it is necessary to transfer the content of the documents to a numeric representation. This step is called text preprocessing, and it is often divided into the activities feature extraction, feature selection, and document representation [3]. Feature extraction is the first step in text preprocessing and consists mainly in parsing the document collection. The goal is to generate a dictionary of words and phrases (i.e., features) that describes the document collection adequately. It is common to distinguish between local dictionaries, which means separate dictionaries for each category, and universal dictionaries, with a single dictionary for the whole document collection. The feature candidates are first compared against a list of stop words, and the dictionary is then usually free of "noise" (e.g., articles, prepositions, numbers). Word stemming techniques also can be applied so that features that differ only in the affix (suffix or prefix), i.e., words with the same stem, are treated as single features. Commonly applied word stemming techniques are affix removal, successor variety, n-grams, table lookup, peak & plateau, and Porter's algorithm [4].

In recent years, various techniques have been developed to reduce the size of the feature matrix, which is sometimes enormous. These techniques rely primarily on the assumption that a large number of features are close to being synonymous. Examples of these techniques are term clustering and latent semantic indexing [2]. Major approaches for ATC classifiers involve the use of decision trees, decision rules, k-nearest neighbors, Bayesian approaches, neural networks, regression-based methods, and vector based methods. At this point, only one representative of the vector-based methods, called "Support Vector Machines" (SVM), is briefly discussed, because NewsCATS is based on this classifier. The difference between SVM, first introduced by Cortes and Vapnik [10], and the other classifiers mentioned above is that in addition to positive training documents, SVM also needs a certain number of negative training documents, which are

untypical for the category considered. SVM then searches for the decision surface that best separates the positive from the negative examples in the n-dimensional space (determined by the n features) [4].

The document representatives closest to the decision surface are called support vectors. The result of the algorithm remains unchanged if documents that are not support vectors are removed from the set of training data. An advantage of SVM is its superior runtime behavior during the categorization of new documents: only one dot product per new document has to be computed. A disadvantage is the fact that a document could be assigned to several categories because the similarity is typically calculated individually for each category. Nevertheless, SVM is a very powerful method and has outperformed others in several studies [6], [7].

Feature extraction is followed by feature selection. The main objective of this phase is to eliminate those features that provide few or less important items of information. While there is a huge number of published articles about data mining, time series and other techniques in prediction of stock prices, the numbers of articles covering the application of text mining in stock market predictions are few. The reason for this is that this is still a new field. As mentioned earlier, there are many researches on the stock market variables. But the first system that scans resulting financial implications of news on stock prices and financial market goes back to Klein and Prestbo (1974) [5] some applications will be reviewed prior to which it is designed to perform this task to predict stock prices. It is analyzed many researches proposed methods to predict stock prices based on news articles. A system proposed called (Analyst to Predict the Stock price) during the day through the analysis of work the contents of the articles movements in real time [6]. System analyzer is based on news articles and time series. It is a complete system that combines two types of data and processes, and tries to find the relationships between them. The system used test algorithm split time (top to bottom) to determine the time-series trends. Uses agglomerative clustering algorithm for trends sectors based on the following characteristics including height and slope intersection. These characteristics serve as a basis for linking trends news articles. The advantage of this model is that it can learn a specific form of which affects each share [7], [8]. The main disadvantage of stock specific models is the small size of their training set, it means that companies that are rarely covered in the news articles are neglected [8]. The task is just to generate profitable action signal (buy and sell) dependent on time series.

Another model proposed by Gidofalvi in 2001 for predicting stock price movements using news articles.

1) Aims to show short-term forecasting and stock price movements statement using financial news articles.
2) This system is also not very successful as it tries to predict future prices based on past prices.
3) The marking articles, as are all rated news articles "up", "Down" or "unchanged" according to the movement of stocks associated at the time of publication of the article.

In (2004), another research in the area of stock prediction is conducted by Khare [12]. Khare developed Web news Mining based on stock movement, named "Stock Broker Prediction". The main components of this system include extracting opinions and sentiments of investors and news from Web sources using manipulation of web pages, and opinion mining. Then create an index based on the data available through the aggregating of sentiment in each new line and they will expect the direction of the stock market as a whole and particularly equities classified on the based on sentiment Index. This system worked to create the dictionary contains important words through a review of the various stock sites.

In (2010) The Arizona Financial Text System (AZFinText), is proposed for machine learning system. This system uses financial news articles and stock prices as inputs to predict characteristic affecting the share price movements:

1) The (AZFinText) system uses nouns as features, and determines the appropriate noun that occur to be included three or more times in the feature set.
2) The (AZFinText) system is characterized by using an algorithm support vector regression (SVR) rather than the classification because most other systems using.
3) Uses suitable nouns as features, and selects the suitable nouns that occurs three or more times.
4) Classifications the news as good, bad or neutral.
5) This means that he is trying to predict the value of deals when stock prices give financial news, and more specifically, it uses material news to predict what the price will be in 20 minutes [3].

TABLE I: SUMMERY OF PREVIOUS WORK

System Name	Year	Idea	Technique	Input	Output
Analyst to predict the stock price during the day	2000	Analysis of the contents of the articles movements in real time (during the day)	time-series	Articles	generate profitable action signal (buy and sell)
Prediction System Design	2001	Show short forecasting and stock price movements	time-series	Financial News Articles	"up", "Down" or "unchanged
Stock Broker Prediction	2004	Extracting opinions and sentiments of investors and news from web sources	Naïve Classifier	Web news Mining	Signals (good, bad or normal)
Arizona Financial Text System (AZFinText)	2010	Uses suitable nouns as features, and selects the suitable nouns that occurs three or more times	Support Vector Regression (SVR)	financial news articles and stock prices	Classifications the news as good, bad or neutral

III. METHODOLOGY

Indicators are commonly used to determine feature importance Term Frequency (TF), Inverse Document Frequency (IDF), and their product (TF×IDF).

When TF is used it is assumed that important terms occur in the document collection more often than unimportant ones [9]. The application of IDF presupposes that the rarest terms in the document collection have the highest explanatory power. With the combined procedure TF×IDF the two

measures are aggregated into one variable. Whatever metric is used, at the end of the feature selection process only the top n words with the highest scores are selected as features. While more sophisticated feature selection techniques, such as information gain, Chi-square, correlation coefficient, and relevance score, have been proposed, the above techniques (especially TF) have proved very efficient [2].

Document representation is the final task in text preprocessing. At this stage, the documents are represented in terms of the features to which the dictionary has been reduced in the preceding steps. Thus, the representation of a document is a feature vector of n elements, where n is the number of features remaining when the selection process is complete.

The whole document collection can therefore be seen as an $m \times n$ feature matrix F (with m as the number of documents), where the element f represents the frequency of occurrence of feature *j* in document *i*. Typical frequency measures are, again, TF, IDF, and TF×IDF, but a difference from the previous task *ij* is that these frequencies are now measured per document. Sometimes the frequency measure is limited to the values (0, 1), which indicate whether or not a certain feature appears at all in the document (binary representation). At the end, the feature vectors are usually cosine normalized, since some of the ATC classifiers require feature vectors of length 1 [5].

IV. Results

The scoring has hinged on whether or not a query term is present within a document. A document that mentions a query term more often has more to do with that query, therefore should receive a higher score. The scoring mechanism is to compute a score that is the sum, over the query terms, of the match scores between each query term and the document. Then combine the definitions of term frequency (*tf*) and inverse document frequency (*idf*), to produce a composite weight for each term in each document [11].

The *tf-idf* weighting scheme assigns to term *t* a weight in document *d* given by:

$$tf - idf\, t, d = tf\, t, d \,.\, idft \quad (1)$$

where *tf – idft, d* assigns to term *t* a weight in document *d* that is:

1) Highest when t occurs many times within a small number of documents (thus lending high discriminating power to those documents).
2) Lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal).
3) Lowest when the term occurs in virtually all documents.

- Score Measure:

The score of a document (*d*) is the sum, over all query terms (*qt*), of the number of times each of the query terms occurs in *d*, also the number of occurrences of each query *t* in *d*, but instead the *tf – idf* weight of each term in *d*.

$$Score\ (q, d) = tf - idf\, t, d \quad (2)$$

- Inverse Document Frequency (*idf*):

It is a measure of whether the term is common or rare across all documents, and it is obtain by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$idft = \text{Log}\ N / dft \quad (3)$$

where: *N* total documents

- Classification of keywords:

In this paper, keywords are extracted from the text of the obtained reports. As shown in Table I, extracted keyword are classify into twenty-two groups after the difference adjusting. We also classify the keywords into three news types: Good, Bad, and Neutral News.

TABLE II: The Keyword

keyword	News Type	keyword	News Type
Increase	Good News	Revenue	Good News
Strong	Good News	Future	Good News
Upward	Good New	Downward	Bad News
Earn	Good News	Sell	Bad News
Upgrade	Good News	Debt	Bad News
More	Good News	Low	Bad News
Great	Good News	Drop	Bad News
Buy	Good News	Risk	Bad News
Profit	Good News	Downgrade	Bad News
Grow	Good News	Normal	Neutral News
Excellent	Good News	unchanged	Neutral News

In addition, Table II has been applied to the following articles (Documents), available electronic links listed below:

1) http://www.albawaba.com/business/pr/axa-shops-484323 - Published April 14th, 2013.
2) http://www.gulfbase.com/news/axa-expands-presence-in-saudi-arabia/234445 - Published April 15th, 2013.
3) http://www.gulfbase.com/news/growing-insurance-awareness-reating-new-opportunities-/235161-Published April 28th, 2013

After searching for keywords in the above articles (Documents), and found a summary of the number of times to repeat each keyword and it shows in the Table III.

Now we want to determine the direction of the third articles based on keywords and unique, is it good news, bad news or Neutral News? After the implementation of this phase appeared to have, the following results are shown below Table IV-Table VI.

TABLE III: Matching Keyword

Term	Doc1			Doc2			Doc3		
	tf	*idf*	*tf.idf*	*tf*	*idf*	*tf.idf*	*tf*	*idf*	*tf.idf*
part	5	2.3	11.6	8	3.0	24.1	15	3.9	58.7
Expansion	8	3.0	24.1	22	4.4	98.2	11	3.4	38.1
Strategy	25	4.6	116.2	16	4.0	64.1	30	4.9	147.3
Aim	4	6.6	26.6	8	3.0	24.1	15	3.9	58.7
Increase	2	1	2	1	0	0	2	1	2
Company	2	1	2	7	2.8	19.7	10	3.3	33.3
geographical	12	3.5	43.1	9	3.1	28.6	3	1.5	4.8
Footprint	23	4.5	104.1	12	3.5	43.1	8	3.0	24.1
Across	30	4.9	147.3	26	4.7	122.3	18	4.1	75.1
Kingdom	4	6.6	26.6	7	2.8	19.7	2	1	2
Saudi	4	6.6	26.6	7	2.8	19.7	2	1	2

Arabia	4	6.6	26.6	7	2.8	19.7	2	1	2
AXA	3	1.5	4.8	7	2.8	19.7	2	1	2
Cooperative	6	3.5	21.6	3	1.5	4.8	2	1	2
One	7	2.8	19.7	10	3.3	33.3	3	1.5	4.8
Large	20	4.3	86.5	15	3.9	58.7	8	3	24.1
Lead	15	3.9	58.7	10	3.3	33.3	8	3	24.1
Life	20	4.3	86.5	11	3.4	38.1	9	3.1	28.6
International	10	3.3	33.3	12	3.5	43.1	20	4.3	86.5
Insurers	22	4.4	98.2	8	3	24.1	3	1.5	4.8
Region	5	2.3	11.7	8	3	24.1	2	1	2
Announced	7	2.8	19.7	3	1.5	4.8	9	3.1	28.6
Recently	7	2.8	19.7	9	3.1	28.6	5	2.3	11.6
Open	11	3.4	38.1	9	3.1	28.6	8	3	24.1
Great	0	0	0	0	0	0	2	1	2
Earn	0	0	0	0	0	0	1	0	0
Risk	0	0	0	0	0	0	1	0	0
Three	6	3.5	21.6	2	1	2	1	0	0
New	7	2.8	19.7	10	3.3	33.3	3	1.5	4.8
Shop	28	4.8	134.7	10	3.3	33.3	8	3	24.1
Access	3	1.5	4.8	5	2.3	11.7	8	3	24.1
Grow	2	1	2	2	1	2	12	3.5	43.1
Opportunity	2	1	2	0	0	0	0	0	0
Khobar	3	1.5	4.8	5	2.3	11.7	2	1	2
Two	3	1.5	4.8	4	6.6	26.6	1	0	0
Jeddah	4	6.6	26.6	2	1	2	4	6.6	26.6
Plans	3	1.5	4.8	6	3.5	21.6	7	2.6	19.7
Further	3	1.5	4.8	4	6.6	26.6	9	3.1	28.6
Conveniently	5	2.3	11.7	3	1.5	4.8	2	1	2
More	3	1.5	4.8	3	1.5	4.8	5	2.3	11.7
Located	6	3.5	21.6	7	2.8	19.7	2	1	2
Low	0	0	0	0	0	0	1	0	0
Throughout	9	3.1	28.6	3	1.5	4.8	8	3	24.1
Bringing	10	3.3	33.3	7	3.3	33.3	4	6.6	26.6
Total	5	2.3	11.7	6	3.5	21.6	2	1	2
Table	10	3.3	33.3	8	3	24.1	5	2.3	11.7
Multi	8	3	24.1	3	1.5	4.8	10	3.3	33.3
Strong	1	0	0	0	0	0	5	2.3	11.7
Profit	0	0	0	0	0	0	2	1	2

TABLE IV: RESULT OF GOOD NEWS

Good News	*df*	*idf*	Doc1-*tf*	*tf* . *idf*	Doc2-*tf*	*tf* . *idf*	Doc3-*tf*	*tf* . *idf*
Increase	3	0	2	0	1	0	2	0
Strong	2	0.58	1	0.58	0	0	5	2.9
Upward	0	0	0	0	0	0	0	0
Earn	1	1.58	0	0	0	0	1	1.58
Upgrade	0	0	0	0	0	0	0	0
More	3	0	3	0	3	0	5	0
Great	1	1.58	0	0	0	0	2	3.1
Buy	0	0	0	0	0	0	0	0
Profit	1	1.58	0	0	0	0	2	3.1
Grow	3	0	2	0	2	0	12	0
Excellent	0	0	0	0	0	0	0	0
Revenue	0	0	0	0	0	0	0	0

Score (q, d) = tf – idf t, d; tq = 11.435

TABLE IV: RESULT OF BAD NEWS

Bad News	*df*	*idf*	Doc1-*tf*	*tf* . *idf*	Doc2-*tf*	*tf* . *idf*	Doc 3-*tf*	*tf* . *idf*
Downward	0	0	0	0	0	0	0	0
Sell	0	0	0	0	0	0	0	0
Debt	0	0	0	0	0	0	0	0
Low	1	1.58	0	0	0	0	1	1.58
Drop	0	0	0	0	0	0	0	0
Risk	1	1.58	0	0	0	0	1	1.58
Downgrade	0	0	0	0	0	0	0	0

Score for bad news = 3.17

TABLE V: RESULT OF NEUTRAL NEWS

Neutral News	*df*	*idf*	Doc 1-*tf*	*tf* . *idf*	Doc 2-*tf*	*tf* . *idf*	Doc 3-*tf*	*tf* . *idf*
Normal	0	0	0	0	0	0	0	0
Unchanged	0	0	0	0	0	0	0	0

Score for Neutral news = 0

The following information has been brought from Saudi Stock Exchange (Tadawul) and note the impact of the announcement news on the activity of stocks and determine the following days 14, 15, 27, 28, 29/04/2013. As listed in Table VII and shown in Fig. 1.

TABLE VI: PERFORMANCE SUMMARY (TADAWUL)

Date	Close	Open	High	Low	Change	# of Trades
30/04/13	49	50	52.3	49	-0.9	3,053
*29/04/13	49.9	47.9	50.3	48	2.4	2,358
*28/04/13	47.5	46.2	49.6	46	1.2	2,551
*27/04/13	46.3	46.2	46.6	46	0.2	622
24/04/13	46.1	47.3	47.6	45	-1.5	793
23/04/13	47.6	47.2	48.2	46	0.4	1,072
22/04/13	47.2	46.4	47.8	46	1	1,356
21/04/13	46.2	46	47.2	46	-0.3	1,403
20/04/13	46.5	44.2	47.9	44	2.8	2,094
17/04/13	43.7	46.8	46.8	42	-3.2	1,090
16/04/13	46.9	47.6	49	47	-1	1,532
*15/04/13	47.9	49	51.5	48	1	4,980
*14/04/13	46.9	42.6	46.9	43	4.2	1,126
13/04/13	42.7	42.1	43.1	42	0.6	929
10/04/13	42.1	42.2	42.6	42	0	769

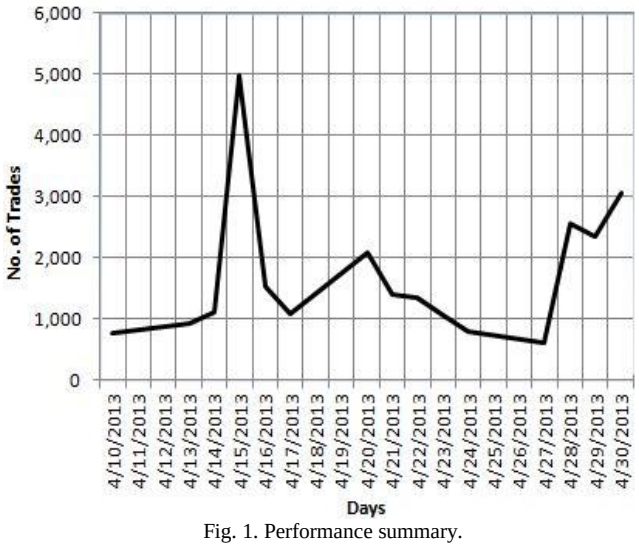

Fig. 1. Performance summary.

V. Conclusion and Future Work

The fundamental search term weighting formula known as IDF, planned by Sparck Jones on heuristic reason in 1972. Note that the use of this method has benefit. In this paper, we investigated the use of the financial news articles in Saudi Stock Exchange (Tadawul), where he managed to predict the stock price up or down through the analysis of some articles relating to one of stocks in the stock market in Saudi Arabia. Process analysis showed that articles if the stock target to study if any positive that the stock price will rise, and when the follow-up event of the arrow and found that the share price has risen as a result of the statements published by the authority responsible in the facility. IDF has also proved a magnet for researchers who feel inspired to replace what they perceive as an heuristic with some reasonably-constituted theoretical argument, in order to 'explain' why it is that IDF works so well.

This conclusion is the result of viewing words separately and does not take into account that words co-occur. While IDF is not expensive to compute, it must be modified when co-occurrence of words is taken into account. There is a common optimization framework where word co-occurrences can be taken into account and which produces the well-known IDF in the special case of a single feature. The general framework opens up the possibility of assigning weights to more sophisticated lexical questions that is consistent with the popular notion of IDF. Through the results that have emerged after analysis of articles three titles mentioned above, we conclude that the articles talk about the situation improved share price during the news. Therefore, it can rely on the news in the forecasting process at market prices up or down.

As a future study, we are working on diverse characteristic representations and feature selection methods to improve our classification performance.

Acknowledgements

The authors would like to express my deep gratitude to Professor Naomie Salim, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work.

References

[1] R. Schumaker and H. Chen, "Textual analysis of stock market prediction using financial news articles," in *Proc. the 12th Americas Conf. on Information Systems*, Acapulco, Mexico, Aug. 2006, pp. 1432-1440.

[2] S. Robertson. (January 2004). Understanding inverse document frequency: On theoretical arguments for IDF. *SIGIR*. [Online]. 60. pp. 503–520. Available: http://www.soi.city.ac.uk/~ser/idf.html

[3] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfintext system," *ACM Trans. on Information Systems*, vol. 27, no. 2, Feb. 2009.

[4] M. A. Mittermayer and G. F. Knolmayer, "News CATS: A news categorization and trading system," in *Proc. the 6th International Conf. on Data Mining*, Washington, DC, 2006, pp. 1002-1007.

[5] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing and Management*, vol 39, pp. 45-65, Aug. 2001.

[6] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, and W. Lam, "Daily stock market forecast from textual web data," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 1998, vol. 3, pp. 2720-2725.

[7] V. Lavrenko, P. Lawire, D. Jensen, and J. Allan, "Mining of concurrent text and time series," in *Proc. the 6th International Conf. on Knowledge Discovery and Data Mining Workshop on Text Mining*, Boston, MA, USA, 2000, pp. 37-44.

[8] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, "Language models for financial news recommendation," in *Proc. the 9th International Conf. on Information and Knowledge Management*, New York, NY, 2000, pp. 389-396.

[9] M. Butler and D. Kazakov, "A learning adaptive bollinger band system," in *Proc. the 2012 Computational Intelligence for Financial Engineering and Economics (CIFEr)*, 2012, pp. 1-8.

[10] P. Dinda, "Online prediction of the running time of tasks," *Cluster Computing*, vol. 5, no. 3, pp. 225-236, July 2002.

[11] P. Dinda and D. O'Hallaron, "Host load prediction using linear models," *Cluster Computing*, vol. 3, no. 4, pp. 256-280, 2000.

[12] R. Khare, N. Pathak, S. K. Gupta, and S. Sohi, "Stock broker p-Sentiment extraction for the stock market," in *Proc. the 5th International Conference on Data Mining, Text Mining and Their Business Applications*, Spain, Sep. 15-17, 2004, pp. 43-52.

Mu'tasem Jarrah received a first degree in computer science from University of Philadelphia, Jordan in 2002. He worked as a developer for the telecommunications sector programs in the sector of information systems in telecom (orange).

Then he got a master's degree from the Al-Balqa Applied University, Jordan with the subject of fuzzy logic technique and data mining. Currently he is holding the position of lecturer in the Department of Information Technology, Faculty of Computing and Information Technology University (KAU) Saudi Arabia. He has studied in several topics including introduction to information technology as well as databases advanced. In addition, He is a candidate PhD student in computer science–stock market predictions based on text mining and neurel network at University Teknologi Malaysia (UTM) presently.

Naomie Salim graduated with a bachelor degree of science (computer science) from Universiti Teknologi Malaysia. She obtained her master degree in computer science from Western Michigan University, USA and her PhD degree in information studies (chemoinformatics) from University of Sheffield. She was the deputy dean for research and post graduate studies at the Faculty of Science and Information Systems for six years since 2005.

In her academic career, Prof. Naomie has taught at both undergraduate and postgraduate level, mostly for subjects related to databases, and information systems. Her research interest includes information retrieval and cheminformatics. Prof. Naomie has been involved in 25 research projects, many in collaboration with colleagues within UTM or with external organizations, to a total value of RM 3 million. She has also authored over 150 journal articles and conference papers describing research into novel techniques for computerized information retrieval, with particular reference to textual, chemical and biological information.

Among the research and innovation awards received by Prof. Naomie are the PECIPTA 2011 Gold Medal award, the I-inova 2010 gold medal award, bio innovation 2011 bronze award, UTM 2010 best research award and the INATEX distinction award in 1998.