Record: 1

Title: An enhanced feature representation based on linear regression model for stock market prediction.

Authors: Ihlayyel, Hani A.K.[1] *hani-ihlayyel@hotmail.com.*
Sharef, Nurfadhlina Mohd[1]
Nazri, Mohd Zakree Ahmed[1]
bakar, Azuraliza Abu[1]

Source: Intelligent Data Analysis. 2018, Vol. 22 Issue 1, p45-76. 32p.

Document Type: Article

Subject Terms: *Stock price forecasting
*Stock prices
*Financial market reaction
*Investors
Chi-squared test

Author-Supplied Keywords: Financial news
linear regression
statistical metric and feature representation
stock market prediction

Abstract: Stock price prediction has been an attractive research domain for both investors and computer scientists for more than a decade. Reaction prediction to the stock market, especially based on released financial news articles and published stock prices, still poses a great challenge to researchers because the prediction accuracy is relatively low. For prediction purposes, linear regression is a popular method. Statistical metrics, such as the Document Frequency (DF), term frequency-invert document frequency (TF-IDF) and information gain (IG), are used for feature selection to extract the most expressive features to reduce the high dimensionality of the data. However, the effectivenesses of the available metrics have not been explored in identifying important financial feature representations that have dependable and strong relations with the stock price. The objective of this study are (i) to investigate the performance of five statistical metrics, namely, DF, TF-IDF, IG, Chi-square Statistics (Chi-Sqr) and occurrence in identifying important features that can represent the news and have a strong relationship with the stock price; (ii) to introduce feedback variables, namely, the prediction accuracy (PA), directional accuracy (DA) and closeness accuracy (CA), to capture the interaction between the released news and the published stock prices; and (iii) to introduce a prediction model that integrates features from financial news and a stock price value series based on a 20-minute time lag using linear regression. The experiment used the ELR-BoW method to build a number of 330 datasets with five statistical metrics to select different feature sizes of 50, 100, 150, 200, 250, 300, 400, 500, 600, 700 and 800. The performance of ELR-BoW is observed based on three parameters, namely, PA, DA and CA, and is compared against Naïve Bayes (NB) as the benchmark approach and the Support Vector Machine (SVM). The proposed ELR-BoW-SVM obtained a higher accuracy compared to ELR-BoW-NB, where the best feedback measure is PA, which has an F-measure value of 0.842. In addition, the best number of features is 300 features and using document frequency DF statistical metric. The identification of the top feature representations for financial news is highly promising for automatic news processing for stock prediction. This study demonstrates that the identification of the top feature representations for financial news is highly promising for news article processing in stock prediction. [ABSTRACT FROM AUTHOR]

Author Affiliations: [1]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor Darul Ehsan, Malaysia

Full Text Word Count: 15581

ISSN: 1088-467X

DOI: 10.3233/IDA-163316

Accession Number: 128309043

Database: Business Source Complete

An enhanced feature representation based on linear regression model for stock market prediction

Stock price prediction has been an attractive research domain for both investors and computer scientists for more than a decade. Reaction prediction to the stock market, especially based on released financial news articles and published stock prices, still poses a great challenge to researchers because the prediction accuracy is relatively low. For prediction purposes, linear regression is a popular method. Statistical metrics, such as the Document Frequency (DF), term frequency-invert document frequency (TF-IDF) and information gain (IG), are used for feature selection to extract the most expressive features to reduce the high dimensionality of the data. However, the effectivenesses of the available metrics have not been explored in identifying important financial feature representations that have dependable and strong relations with the stock price. The objective of this study are (i) to investigate the performance of five statistical metrics, namely, DF, TF-IDF, IG, Chi-square Statistics (Chi-Sqr) and occurrence in identifying important features that can represent the news and have a strong relationship with the stock price; (ii) to introduce feedback variables, namely, the prediction accuracy (PA), directional accuracy (DA) and closeness accuracy (CA), to capture the interaction between the released news and the published stock prices; and (iii) to introduce a prediction model that integrates features from financial news and a stock price value series based on a 20-minute time lag using linear regression. The experiment used the ELR-BoW method to build a number of 330 datasets with five statistical metrics to select different feature sizes of 50, 100, 150, 200, 250, 300, 400, 500, 600, 700 and 800. The performance of ELR-BoW is observed based on three parameters, namely, PA, DA and CA, and is compared against Naïve Bayes (NB) as the benchmark approach and the Support Vector Machine (SVM). The proposed ELR-BoW-SVM obtained a higher accuracy compared to ELR-BoW-NB, where the best feedback measure is PA, which has an F-measure value of 0.842. In addition, the best number of features is 300 features and using document frequency DF statistical metric. The identification of the top feature

representations for financial news is highly promising for automatic news processing for stock prediction. This study demonstrates that the identification of the top feature representations for financial news is highly promising for news article processing in stock prediction.

Financial news; linear regression; stock market prediction; statistical metric and feature representation

1. Introduction

Stock market prediction continually draws the attention of researchers and financial investors because mastering the nuances of the market promise the ability to gain surplus profits. The rapid growth of online textual data such as financial news poses a challenge in extracting valuable information and determining its relationship to the stock market. In this respect, the limitation of the stock prediction models is mainly in transferring unstructured data to a structured format to model the stock market dynamicity accurately [[16]].

The investors are interested in getting the highest profits from the market, therefore, identifying the future trend of the stock is important, and this is termed as forecasting the stock prices. Predictions of stock prices can be performed using structured data (i.e., stock price records) and unstructured data (i.e., financial news with regard to the stocks). The structured data are categorized into two types, namely, fundamental analysis and technical analysis. The fundamental analysis evaluates the stock security by examining the related economic, financial and other qualitative and quantitative factors, whereas technical analysis utilizes statistics on the stock market, such as the past prices and volumes, which are modeled using mathematical tools to predict trends in the future values [[10] , [27]].

On the other hand, the success of the analysis methods that use unstructured data has gained more attention in stock price prediction. Among the popular methods is the text mining approach, which aims to explore and exploit the relationship between the news articles and the time-stamped stock prices [[28]]. Several studies have demonstrated the influences of the news articles on the stock market price where there is a strong relationship between the time of the stock price fluctuation and the time of the released news articles. The provided information in the news articles includes a number of terms that have a direct effect on the stock price [[12]]. Most of the previous studies extract a set of features such as the top financial terms published in the news and the used machine learning techniques in the prediction model [[1] , [34] , [42]]. These studies assign weights to these features to predict the stock market movements. However, these methods have obtained very weak stock price prediction performance mainly because of the relationships between the structured and unstructured data, which indicate the stock fluctuation behaviors. However, stock market prediction based on time series data might be not sufficient, due to the existing of a huge number of factors that affect the stock market movements that could be political, economic and psychological, which are inherently noisy, non-stationary and non-deterministically [[8]].

According to Nassirtoussi et al. [[28]], there is a strong correlation between the news articles and stock price. Several studies have demonstrated the influences of the news articles on the stock market price where there is a strong relationship between the time of the stock price fluctuation and the time of the released news articles.

The previous studies have confirmed that the news article has a positive and negative impact on the stock price movement, these news articles effect the measurement of return volatility and return volatility [[9] , [17]]. The strong efficient market hypothesis (EMH) states that the stock market is influenced by all kind of information. This hypothesis has motivated us to investigate all the possible of information that has an impact on the stock price movements [[4]]. Therefore, It is important to process all the available information that are related to the stock market to extract the most useful time series patterns and increase the performance of stock price prediction [[21]].

Recently, the combination of structured and unstructured data is assumed to provide better stock price prediction by combining the features that are extracted from both data modalities. Several techniques have been investigated to build more representative features for the stock market fluctuations. The bag-of-words technique is implemented [[9] , [30] , [38]] to denote the binary representation of terms, but the frequencies of these features are ignored [[9]]. Additionally, different techniques have been investigated, for example, noun phrases and named entities [[34]] are implemented to extract the occurrences of the named entities.

Other studies have explored the impact of statistical metrics on the prediction accuracy, such as the TF-IDF method, which captures the distribution of features inside the documents [[13] , [16] , [30]]. An attempt to select the features using the mutual information (MI), balanced mutual information (BMI) and chi-sqr to predict the directions of the stock prices has been made [[14]]. However, the existing statistical-based approaches still have a weak ability to capture the relationship between the news articles and the stock prices, to model all of the relative movement and fluctuations of the stocks accurately [[42]]. Moreover, there is no available research that has investigated the best statistical metrics to decide on the most representative features for the prediction modeling of a fluctuating stock price. A short-timeline-based prediction has an added value compared with the existing methods, which have commonly depended on the intra-day rate [[30]].

A few studies capture the impact of correlation features to explore more relationship between the unstructured data and stock price [[11] , [13] , [34]]. However, these methods have obtained very weak performance to capture correlation features, mainly due to two reasons (i) they ignore the temporal effect of the stock price for the short timeline, and (ii) the limitation of existing techniques to represent expressive features that affect the stock price movements.

Due to the limitations of the existing techniques to extract a correlation features that affect the stock price from a staggering amount of textual data. In this study, we intend to develop an algorithm for feature representation using time series data for short timeline prediction that implements a technique to discover series correlation features based on temporal events to predict the stock market movements. Therefore, this study addresses the investigation of the performance of statistical metrics and introduces feedback variables to build an Enhanced Linear Regression Based Bag-of-Word Model for Feature Representation (ELR-BoW) algorithm. The ELR-BoW utilizes the relationship between the news articles and stock prices based on bag-of-words for a short-timeline stock prediction. The ELR-Bow algorithm is based on heuristic using statistical measures to speed up the search process to find the best solution for the search space [[20] , [22]]. The heuristic search aims to discover series of correlation between the features for short timeline prediction [[44]]. The contributions of the study are three-fold; (i) identifying the best feature extraction model using five statistical measures, namely, DF, TF-IDF, IG, Chi-square Statistics (Chi-Sqr) and occurrences, (ii) introducing feedback variables, namely, closeness, directional movement and prediction, as indicative measures for the interaction between the financial news and stock prices, and (iii) proposing stock price predictions based on linear regression. The S&P500 index close prices dataset is used.

Our study shows that feature representation using the ELR-BoW algorithm has the ability to discover the relationship and represent the direct effect of news articles on the stock price. The implementation of the proposed feedback measure (PA) pushed the F-measure value up to 0.842 when the features are incorporated with SVM. The analysis of different feature sizes has different feature selection methods demonstrate that the best feature size is 300 when using the DF selection method.

This paper is organized as follows. Section 2 introduces the background of the study. Section 3 presents the an enhanced-linear regression based bag-of-word model for feature representation (ELR-BOW) for stock price prediction, which is based on short timeline stock information for stock price prediction. Section 4 presents the effectiveness evaluation. Section 5 presents the experimental results and the findings of the paper. Section 6 provides the conclusions of the paper.

2. Related studies

The financial time series facilitate the effective extraction of positive patterns of the stock market and predict its movements. The topic of stock market prediction continually draws the attention of researchers and financial investors, that whosoever capable of mastering the nuances of the market can beat the market and able to gain surplus profits. Generally, the investors are unaware of their stocks behavior; hence, they face difficulty in trading stocks. The investors mostly fail to gain more profit in trading stocks, as they are uncertain about the nature of the stock market and unsure of which stock to buy or sell. Nevertheless, it is crucial for them to be able to predict the future behaviour of the stock prices in order to gain more insights for trading.

This has further encouraged academic researchers and business practitioners to develop more time series prediction models by implementing artificial intelligence (AI) techniques, such as an artificial neural network (ANN), that are extensively used to accurately forecast the stock index and direction of its change [[19] , [29]]. Meanwhile, excellent performance from the Support Vector Machine applications has been obtained in investigating the issue of forecasting the stock index futures market [[7] , [8]]. However, the main challenge in stock price prediction is the price fluctuations [[6] , [16]].

According to the strong efficient market hypothesis (EMH), the stock market price data fluctuation reflects the all the information available about the stock market [[24]]. Furthermore, the efficient-market hypothesis (EMH) elucidate a link between the published information and the market price movements. The investors cannot guarantee that they will always achieve consistent returns even if they have a prior knowledge of the stock information before the investment [[5]]. The existence of an enormous amount of financial news generated from different sources has a direct effect on the market movement [[39]]. Therefore, understanding the news content and combining it with the stock price data can contribute to increasing the accuracy of the stock price prediction model.

One of the main issues of handling the textual data remains a sophisticated due to a large amount of information and the availability of different sources. In order to analysis this information and figure out the relationship Natural Language Processing (NLP) techniques need to be used to identify the most significant terms that might causes changes on the security prices [[35]]. So that, the analysis of the textual information are a great chance to know if the news article consists of good or bad news and attempt to predict the direction of the stock price in the future.

The idea of trading (buy/sell) the stock when there is a good or bad information. The unexpected good and bad news in the stock markets always occur, these cases make the stock price unpredictable due to the high volatility [[37]]. The news articles contain trustworthy information that leads to moving the stock prices. According to Zhang and Skiena [[43]], the news articles considered a reliable source that can be important as much as the commodity. Therefore, text mining pre-processing an important to analysis the text information and extract the most significant feature that has an impact on the stock price movements [[32]]. Although, there are several studies address the stock market movements. However, the investors are still interested to know more about stock market movements.

One of the most challenging aspects is to predict stock market movements from textual data due to the difficulty to capture correlation features between the stock price and news articles [[6] , [16]]. A few studies attempt to addresses the problem by proposed systems to capture the impact of correlation features based on bag-of-words (BoW) to explore more relationship between the unstructured data and stock price to predict the stock price in specific periods [[11] , [13] , [28] , [34]]. However, such a prediction models suffer from providing an accurate performance due to sudden changes in the stock market and a huge price fluctuation per minute in the stock market [[34]]. The investigated approaches are still in early stages and there is a need to dive more deep to examine inclusion the extracted features with the stock price that demonstrate the impact of price fluctuation for short timeline prediction [[28]].

Studies that model the relationship between the released news and the market movement have grown over the years. These include the investigation of representative features from the news and from the stock data as well as machine learning algorithms such as Support Vector Machines (SVM) [[15] , [28] , [34]], Naïve Bayes [[14] , [41]] and decision tree [[30]]. To represent the relationship between the news articles and the stock prices, there are several studies that map the news with the stock price time stamp to predict the stock price for specific periods. Table [1] shows a comparison of the pre-processing steps and machine learning methods used in various studies on modeling stock prices, which span from 1998 to 2015.

Table 1 Pre-processing steps and machine learning methods for stock price modeling

Reference	**Pre-processing steps**				**Machine learning**	
	Feature selection	**Dimensionality reduction**	**Representation**	**Timeline**	**Forecast type**	**Classifier**
[38]	Bag-of-words	Word sequence by expert	Binary	Daily	Up, down and steady	-N/A-
[30]	Bag of words	Keyword list	TFIDF TFCDF	1, 2 and 3 hours	Up, down and steady	SVM with Gaussian RBF kernel
[25]	Bag of words	Selecting 1000 terms	TFIDF	Daily	Good or bad	Naive Bayes, k-NN, SVM
[42]	Bag of words	Word net dictionary + top 30 concept	Binary TFIDF	Daily	Up or down	SVM
[33]	Bag of words, noun phrases, named entities	Minimum occurrence per document (3 times)	Binary	20 minutes	Discrete numeric	SVM
[6]	Character n-	Minimum	Frequency	Yearly	Up or	SVM-light

	Grams, three readability scores	occurrence per document			down	
[13]	Bag of words	Feature scoring methods using both Information Gain and Chi-Squared metrics	TFIDF	20 minutes	Positive or negative	CNG distance measure & SVM & combined
[34]	Opinion finder, overall tone and Polarity	Minimum occurrence per document (3 times)	Binary	20 minutes	Regression	NB
[14]	Bag-of-words	Series of best keywords 100, 200..1000	Binary	Daily	Up, down, error	Naïve Bayes
[16]	Bag-of-words, noun phrases, word combinations, n-grams	Frequency, Chi2 bi-normal separation (BNS) for exogenous-feedback based feature selection, dictionary	TFIDF	Daily	Positive or negative	Naive Bayes, k-NN, ANN, SVM
[11]	Bag of words	-N/A-	Bag-of-words, simple item count, plain, Piecewise Linear and technical indicator	5 minutes	Up or down	NN, DT and Stepwise Logistic regression
Nassirtoussi et al. [28]	Bag of words	Using wordnet to replace words	TFIDF	2 hours	Up, down and steady	SVM

The pre-processing steps are divided into feature selection, dimensionality reduction, data representation technique and timeline used. The bag-of-words technique is the most commonly used technique for feature selection, which is mainly due to its advantage of retaining the occurrence multiplicity [[13]]. For dimensionality reduction purposes, several methods have been used, such as filtering according to certain occurrence thresholds, expert-based keyword determinations, and scoring-based methods. At the same time, the binary method and the term frequency-inverse document frequency (TFIDF) method are the most used representation techniques because they indicate the weight of the selected terms in representing the documents. For the purpose of correlating the financial documents with the released stock price, several time granularities have been used. The time-line for the 20-minute stock close value has achieved a remarkable explanation with regard to the news impact on the stock market [[31]].

The developed machine learning-based prediction models can be explained according to the forecasting type and the classifiers. Various forecasting types have been applied, such as binary class, multiple class and discrete. However, only the discrete type forecasting through a regression-based technique can allow numerically based estimation of a stock price [[23]]. Several classifier types have also been explored, and the SVM is the most popular [[3]].

However, none of the existing approaches covered in the literature have provided a method for feedback measurement to capture the interaction between the fluctuating stock price and the released news. This technique has a low prediction accuracy because by depending on the latest stock price only, the stock fluctuation is ignored. The relationship between the stock price and the related messages in the released financial news is also vague. Linear regression is a machine learning-based approach that has the capability of capturing the relations from the financial news. The linear regression approach requires identification of strong features that can represent the direction of the stock price [[26]].

Although the TFIDF is the widely used representation approach, the performance of other statistically based feature representation methods on improving stock predictions is unknown. Therefore, this research fills this gap and addresses the investigation of effective feature representations through statistical metrics-based evaluation and through introducing feedback variables into the linear regression models, toward achieving high stock prediction accuracy.

3. An enhanced-linear regression-based bag-of-word model for feature representation ...

The primary goal of this paper is to introduce an enhancement to the conventional bag-of-words representation that will be able to capture the temporal events that effect the stock price for time-series data. The proposed model is based on the integration of statistical measurements with linear regression for short timeline prediction (within a 20-minute context) published financial news and the stock price. The proposed model map and represent the most relevant features that will increase the classification accuracy. Figure [NaN] presents general architecture of ELR-BoW implementation to discover the temporal effect from each feature vector.

The general architecture of the model building is composed of three phases. The first phase is called bag-of-words representation that is used 5000 news articles to build a lexicon and apply pre-processing steps, the second phase is stock price pre-processing and the third phase is feature selection technique. The next subsections discuss in details the description of each phase. The ELR-BoW algorithm is designed to tackle the limitations of feature representation using time series data for short timeline prediction. Which aims to discover series correlationfeatures based on temporal events to predict the stock market movements. The ELR-BoW implements different statistical metrics and introduces feedback variables to build an effective linear regression model, which utilizes the relationship between for a short timeline stock prediction.

The dataset consists of a total of 46674 news articles (saved into a Table called News) and stock price information (saved into Tables named Quote and Ticker) on the S&P500 gathered from an online financial news corpus such as from noodle and Reuters. The dataset consists of three Tables, namely, the ticker, quote and news Tables. Only the Quote and News Tables are used in this experiment. The quote Table records the stock information, and the attributes involved are the quote symbol for the stock names, quote time, quote close, quote high, quote low, quote open and quote volume. Only data on the quote close are used in this experiment.

We implemented the pre-processing steps of text mining such as tokenization, stemming and stop removal to extract a pattern from the structured or unstructured data. the main aim of these steps is to clean the text data by eliminating all the irrelevant characters (such as stop words, conjunctions prepositions, etc.) to reduce the dimensionality of term space [[36]]. The importance of the text processing is to remove all the characters that do not carry any significant meaning to the text, these characters are noisy and irrelevant data, those words are not measured as features in text mining application [[2]].

This research utilizes both the structured data (released stock price) and the unstructured data (financial corpus) for building the prediction model. The financial news corpus consists of 8500 articles, which are collected between 6th November 2013 and 25th March 2014. A total of 5000 news articles are used for training and building a lexicon, while 3500 news articles are used for evaluation purposes

3.1 Phase 1: Bag-of-words representation

The main process in the first phase is to represent the news articles (unstructured data) using bag-of-words technique N×1 feature vector D=[i1,i2,…,iN]. For the training data, the documents are represented as a set of f unique features or terms; perform queries to these functions to retrieve feature from the documents. In this step, five feature vectors VI=[i1,i2,…,i5] have been built, namely, TFIDFVector, IGVector, ChiVector, CoVector, and DFVector, for each of the statistical measures. These vectors are used to identify the representative words as features according to the statistical measure's ranking. Figure [NaN] presents the Enhanced BoW (eBoW) Representation Algorithm.

For the evaluation document, features set are formed in binary format (0, 1) to represents the absence or presence of terms for each document [[18]]. The binary representation of the words can be expressed as X(i)=[f1i,f2i,f1i,…,fni], where f1i= 1 (if the word f1 appears in the dth document) or f1i= 0 (if the word f1 is absent in the dth document). Figure [NaN] shows the bag-of-words representation using statistical measures.

The first phase utilizes the bag-of-words technique to list all of the words in the financial news corpus. Then, for each of the words, their scores according to the five statistical metrics, namely, TFIDF, Occurrence, Chi-Square, IG, and DF, are calculated. Next, the words stored in each vector are determined based on the top score. Figure [NaN] shows the steps in the first phase.

The formula for the calculation of the statistical metrics is as follows:

a) Term frequency invert document frequency (TFIDF): TFIDF evaluates the significance of a single word inside a document (1)&#xd835;&#xdc13;&#xd835;&#xdc05;&#xd835;&#xdc08;&#xd835;&#xdc03;&#xd835;&#xdc05;&#xd835;&#xdc98;=&#xd835;&#xdc61;&#xd835;&#xdc53;&#xd835;&#xdc8b;×l where tf is the number of times that the word w appears in document j N denotes the number of words w that appears in document j

b) Occurrence: Occurrence measures the number of words that occur in all of the documents, to indicate how relevant the word is to the domain. (2)Occ=(wd)

c) Chi-square: The Chi-square value is a statistical metric that is used to compare the independence between two random variables using the following equation: (3)Chi-square=D×P(w)-(D-P(w))2(D*P(w))*(D-P(w)) where D is the total number of documents P(w) is the percentage of times that the word w appears in a document based on the total number of documents

d) Information gain (IG) This metric is used to measure the expected reduction in entropy by assuming the presence and absence of a term in the document. The expected reduction in entropy is caused by partitioning the examples according to a given attribute. (4)IG=∑p(w)&#xd835;&#xdc59;&#xd835;&#xdc5c;&#xd835;&#xdc54;DP(w) where D is the number of documents in the selected range P(w) is the percentage of times that the word w appears in a document based on the total number of documents

e) Document frequency (DF): The document frequency depends on a very simple idea, to calculate the number of appearances of a single term in the existing documents, which is aimed at measuring how often the term is used. (5)DF=w(dj) where w represents the word, and dj is the number of the document that this word appears in.

3.2 Phase 2: Stock price pre-processing

In the second phase, a time series for stock price pre-processing (structured data) is conducted and incorporated into the feature vectors that were built in the first phase. The preparation of stock price information and the feedback parameters aims to compose the prediction model's features. The details of the process in each phase are described. Figure [NaN] shows the algorithm for the stock price pre-processing in phase 2.

The algorithm mechanism selects the document symbol DS to pick the stock symbol names that represent one stock market index. For each stock symbol, there are two variables are picked up which are (document date Dd and document time Dt). In the past research, time-line for 20-minute close value has considered achieving remarkable explanation of news impact on the stock market [[31]]. The main reason behind using 20 minutes time-line is to develop a method that able to captures the rapid stock price fluctuation and model all the relative movement of the stock accurately. Figure [NaN] shows the mapping process between the news articles and stock price. The presentation of the data in time-series format according to the following filter specifications:

(i) news date: specific intervals from Nov 6, 2013, to Mar 25 (35 days) to build data that is comparable to data in previous studies [[12] , [25] , [34]]

(ii) news time: from 9:00 am to 4:00 pm. These intervals are extremely important to restrict the news articles that highly affect the stock price to market hours, to reduce the impact of overnight news and allow for market prediction.

(iii) 20-minute lag-time: to remove redundant news and ensure that only news that appears within 20 minutes is retained [[28] , [34]].

At the end of the filtering process, the remaining 1887 financial news is left. Figure [NaN] represents the Close values for the stock price based on 20 minutes. For each document, three stock price value are obtained, these values are the stock close time st based on 20 minutes, which are the current close value Ct, the previous 20 minutes close values, (Ct-20) and the post-20-minutes close values, (Ct+20). Next, linear regression is applied to measure the next future value after 20 minutes, Cp. The linear regression requires Ct and Ct-20 as the explanatory variables and Ct+20 as the dependent variable. The total of document number D and the feedback measures PA, CA, DA are 1887. The bag-of-words for time series data are constructed by incorporating the documents with the feedback measures.

(6) $Cp = a + bCr$ (7) $b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$ (8) $a = \bar{y} - b\bar{x}$

where

C represents the current close value mentioned in the news article

Cp: the predicted (next 20 minutes close value)

b: The slope of the line

a: The intercept (the value of y when Cr= 0)

x= represents Cr+20

y= represents Cr-20

$\bar{x}$= represents the average Cr+20

$\bar{y}$= represents the average Cr-20

For the purpose of feedback measures that enable the relation between the predicted and actual close values to be captured, three parameters are used. These parameters evaluate the strength of the linear regression prediction technique, namely, the Prediction Accuracy (PA), Directional Accuracy (DA), and Closeness Accuracy (CA).

• Prediction Accuracy, PA: measures how close the Cr+20 is to Cp (9) $PA = \sum(Cp - Cr{+}20)$: $\{IF(Cp > Cr{+}20) \rightarrow up\ IF(Cp < Cr{+}20) \rightarrow$ &#xd835;&#xdc51;&#xd835;&#xdc5c;&#xd835;&#xdc64;&#xd835;&#xdc5b;

• Directional Accuracy, DA: measures how close the Cr+20 movement direction is to Cp (10) $DA = \sum(Cr - C20)$: $\{IF(Cr > C20) \rightarrow up\ IF(Cr < C20) \rightarrow$ &#xd835;&#xdc51;&#xd835;&#xdc5c;&#xd835;&#xdc64;&#xd835;&#xdc5b;

• Closeness Accuracy, CA: measures how close the Cr is to Cr+20 (11) $CA = \sum(Cp - C20)$: $\{IF(Cp > C20) \rightarrow up\ IF(Cp < C20) \rightarrow$ &#xd835;&#xdc51;&#xd835;&#xdc5c;&#xd835;&#xdc64;&#xd835;&#xdc5b;

The value of the feedback parameters is incorporated into the earlier prepared features and is used as prepared data for building the NB and SVM classifiers. The next section presents the heuristic Feature selection technique to discover temporal information.

3.3 Phase 3: Heuristic feature selection technique to discover temporal information

In order to select the most useful feature set $f=[f_1,f_2,\ldots,f_n]$ that capture the temporal events, a feature selection search technique is used to select a set number of features. The features are selected based on each statistical measures. Previously, the statistical measure form five feature vectors TFIDFVector, IGVector, ChiVector, CoVector, and DFVector. The proposed feature selection method aims to select set f from each vector. The value of the feedback parameters PA, CA, DA is incorporated into prepared features and used as prepared data for feature selection process. For building the NB and SVM classifiers, the NB and SVM are built based on 10-fold cross-validation.

Figure [NaN] shows the pseudo code for the feature selection technique based on heuristics, which describes the procedures of the selecting the best set of features for temporal data. The process of feature selection technique begins with input the top feature set for unique words, sort the features number according to the score and insert a set of best features number.

The process starts with identifying a set of features in the search space. At the first step, a feature number is selected according to feature vectors VI= ChiVector, DFVector, TF-IDFVector, IGVector, OccVector and sort it in decreasing order. For each feature vector, it selects a number of features (50, 100, 150, 200, 250, 300, 400, 500, 600, 700 and 800), class label (CA, PA, DA) and classification method (SVM and NB) to evaluate the performance based on two criteria (weighted accuracy and F-measure). Then, the feature selection technique calculate the initial value, the initial value assigned as the best value. In the second iteration, a different feature number is selected and the obtained another initial value are compared with the best value (initial value>Best value) to select the best feature number. This process is repeated until the identified number of features is reached.

4. Effectiveness evaluation

The classification effectiveness can be evaluated based on three evaluation measures which accuracy, F-measure and weighted accuracy. These evaluation measures are used to evaluate the effectiveness of the binary classification of document categorization. The classification process labels the binary data into two different categories either positive or negative, the classification is represented in confusion matrix according to the confusion the two class problem.

The confusion matrix consists of four categories: false positive (FP) indicates the negative instances and incorrectly labelled instances as positive, true positive (TP) the instances that correctly labelled as positive, true negative (TN) refer to the instances that are correctly labelled as negative and false negative (FN) indicates the negative instances that incorrectly labelled as negative. These are the content of the confusion matrix, these four categories are used to calculate the precision, recall, and F-measure.

i. Average of precision for the d class label: (12) $Precision = \sum_{i=1}^{d} \frac{TP_i}{TP_i + FP_i}$

ii. Average of recall for d class label: (13) $Recall = \sum_{i=1}^{d} \frac{TP_i}{PT_i + FP_i}$

iii. Average of F-measure for d class variables: (14) $F\text{-}meausre = \sum_{i=1}^{d} \frac{2 \times precision \times Recall}{Precision + Recall}$

i. Weighted Accuracy for F-measure value for x and y calsses Weighted Accurcy (15) $= \frac{\sum_{x}^{d} F\text{-}measure \times = Num\ of\ x\ class + \sum_{y}^{d} F\text{-}measure \times Num.\ of\ y\ class}{\sum Total\ instances}$

i. Accuracy for positive predictive value: (16) $Accurcy=\frac{\sum Ture\ Positve}{\sum Total\ of\ positive\ outcome}$

The evaluation is performed by measured the weighted average F-measure values for the classified classes. The macro F-measure score is computed by calculating the total performance for all categories. Then, the total score is used to calculate the performance for each category in the table.

5. Experimental results and evaluation

As described before, the main aim of this study is to enhance bag-of-words representation mechanism to capture the effect of news article features on the stock price. The representations of bag-of-words are

Table 2 Details of the PA datasets space numbers with different feature sizes

No of features	PA					DA					CA				
	CHI	DF	TFIDF	IG	OCC	CHI	DF	TFIDF	IG	OCC	CHI	DF	TFIDF	IG	OCC
50	DS1	DS2	DS3	DS4	DS5	DS56	DS57	DS58	DS59	DS60	DS111	DS112	DS113	DS114	DS115
100	DS6	DS7	DS8	DS9	DS10	DS61	DS62	DS63	DS64	DS65	DS116	DS117	DS118	DS119	DS120
150	DS11	DS12	DS13	DS14	DS15	DS66	DS67	DS68	DS69	DS70	DS121	DS122	DS123	DS124	DS125
200	DS16	DS17	DS18	DS19	DS20	DS71	DS72	DS73	DS74	DS75	DS126	DS127	DS128	DS129	DS130
250	DS21	DS22	DS23	DS24	DS25	DS76	DS77	DS78	DS79	DS80	DS131	DS132	DS133	DS134	DS135
300	DS26	DS27	DS28	DS29	DS30	DS81	DS82	DS83	DS84	DS85	DS136	DS137	DS138	DS139	DS140
400	DS31	DS32	DS33	DS34	DS35	DS86	DS87	DS88	DS89	DS90	DS141	DS142	DS143	DS144	DS145
500	DS36	DS37	DS38	DS39	DS40	DS91	DS92	DS93	DS94	DS95	DS146	DS147	DS148	DS149	DS150
600	DS41	DS42	DS43	DS44	DS45	DS96	DS97	DS98	DS99	DS100	DS151	DS152	DS153	DS154	DS155
700	DS46	DS47	DS48	DS49	DS50	DS101	DS102	DS103	DS104	DS105	DS156	DS157	DS158	DS159	DS160
800	DS51	DS52	DS53	DS54	DS55	DS106	DS107	DS108	DS109	DS110	DS161	DS162	DS163	DS164	DS165

composed in time-series forms, this kind of representation allows predicting the temporal effect within 20 minutes time-line. In the past studies, it is indicated that incorporated bag-of-words with the temporal effect of stock price lead to discovering more pattern for the stock price [[40]]. This makes a logical sense, the proper representation of the document in a time-series format with the stock price allows any model to provide more accurate prediction accuracy.

We performed experiments on our dataset with a Bag-of-Words representation, which contained 1887 news articles. To compare the performances of the different feature selection methods (Chi-Square, DF, TF-IDF, IG and occurrence), we allowed each feature selection method to select the most relevant 50, 100, 150, 200, 250, 300, 500, 600, 700 and 800 features from the 1887 articles and to represent each news article in the feature vector with respect to the number of selected features. For each feature vector, a binary representation is used (0, 1); these values indicate the absence or presence of the features inside each news article. We extracted 165 datasets (Table [2]) to cover all of the features sizes, and then, we labeled the data in two directions, namely, up and down, using three different class labels, namely, PA, DA and CA, as explained in the previous section.

In order to distinguish the variety of the datasets, a unique number has been added to each data space (DS). Table [2] shows the Name of Data Spaces (DS) numbers used for each PA, DA, CA feedback measures. The numbers of datasets (DS) are used to indicate the feature size and the statistical measure respectively.

An experiment has been conducted to observe the effectiveness of the feedback measures and statistical metrics as the representative features for the stock price modeling. This evaluation testifies to the ability of the features (which consist of the news ID, news publication time, the top selected expressive words determined by each of the statistical metrics, Cr, Cr+20, Cr-20, and the feedback measure parameters) to capture the strong relationship between the financial news and the stock prices for a short time-line prediction.

From investigating the methods in the literature review and building the same techniques on our dataset, we can easily justify and benchmark our approach. The classification accuracy is used to predict the performance of the stock price feedback using Naïve Bayes [[14] , [41]] and SVM [[15] , [28] , [34]]. Therefore, we can testify that our results improvements are feasible based on the stock market feedback.

In our experiment, we measured the performance of the stock price using two classification methods, the NB and SVM, using three class labels PA, DA and CA, and we compared the performance of the proposed class label prediction accuracy (PA) against the closeness accuracy (CA) [[34] , [38]] and the direction accuracy (DA) [[12] , [34]]. We used the number of correctly classified instances and the accuracy for the whole test set. In addition, to evaluate the best classification method, we used the F-measure value for each direction (up, down) and the weighted accuracy for PA-SVM against PA-NB. Finally, we conducted the experiment using different feature sizes. We calculated the average and standard deviation values for PA-SVM to identify the best feature size and the best statistical metrics.

Table 3 Classification accuracy for NB using chi-sqr

No of features	CHI – NB								
	PA			DA			CA		
	D.S	ACC	NUM	D.S	ACC	NUM	D.S	ACC	NUM
50	DS1	73.03	1349	DS5	53.87	995	DS101	53.87	995
100	DS6	73.03	1349	DS10	58.20	1075	DS106	53.87	995
150	DS11	73.09	1350	DS15	58.20	1075	DS111	53.87	995
200	DS16	73.09	1350	DS20	58.2	1075	DS116	53.87	995
250	DS21	73.09	1350	DS25	58.2	1075	DS121	53.81	994
300	DS26	73.09	1350	DS30	58.14	1074	DS126	53.81	994
400	DS31	73.09	1350	DS35	58.14	1074	DS131	53.81	994
500	DS36	73.09	1350	DS40	58.14	1074	DS136	53.81	994
600	DS41	61.88	1143	DS45	58.2	1075	DS141	53.7	992
700	DS46	61.28	1132	DS50	58.2	1075	DS146	53.7	992
800	DS51	73.09	1350	DS55	58.2	1075	DS151	53.7	992

Table 4 Classification accuracy for NB using DF

No of features	DF – NB PA	DA	CA
	D.S ACC NUM	D.S ACC NUM	D.S ACC NUM
50	DS2 70.381300	DS52 52.355967	DS10252.35967
100	DS7 68.701269	DS57 56.41 1042	DS10752.03961
150	DS1268.591267	DS62 55.65 1028	DS11252.35967
200	DS1766.811234	DS67 56.25 1039	DS11752.35967
250	DS2265.671213	DS72 53.7 996	DS12250.83939
300	DS2764.591193	DS77 54.19 1001	DS12750.67936
400	DS3273.091350	DS82 58.14 1074	DS13253.81994
500	DS3773.091350	DS87 58.14 1074	DS13753.81994
600	DS4273.091350	DS92 58.2 1075	DS14253.7 992
700	DS4773.091350	DS97 58.2 1075	DS14753.7 992
800	DS5273.091350	DS10258.2 1075	DS15253.7 992

Table 5 Classification accuracy for NB using TF-IDF

No of features	TF-IDF – NB PA	DA	CA
	D.S ACC NUM	D.S ACC NUM	D.S ACC NUM
50	DS3 71.081313	DS53 53 979	DS10353 979
100	DS8 71.141314	DS58 57.39041060	DS10853.05980
150	DS1370.271298	DS63 57.71 1066	DS11352.84976
200	DS1870.051294	DS68 56.95 1052	DS11852.35967
250	DS2368.971274	DS73 57.33 1059	DS12352.78975
300	DS2869.031275	DS78 57.33 1059	DS12852.57971
400	DS3368.861272	DS83 56.63 1046	DS13352.24965
500	DS3868.431264	DS88 56.9 1051	DS13853.16982
600	DS4366.051220	DS93 55.49 10.25	DS14351.1 944
700	DS4864.911199	DS98 54.52 1007	DS14850.4 931
800	DS5365.021201	DS10356.19 1038	DS15349.21909

Table 6 Classification accuracy for NB using IG

No of features	IG – NB PA	DA	CA
	D.S ACC NUM	D.S ACC NUM	D.S ACC NUM
50	DS4 70.381300	DS54 52.35967	DS10452.35967
100	DS9 68.701269	DS59 56.411042	DS10952.03961
150	DS1468.591267	DS64 55.651028	DS11452.84976
200	DS1966.811234	DS69 56.251039	DS11952.35967
250	DS2465.671213	DS74 53.7 992	DS12450.83939
300	DS2964.591193	DS79 54.191001	DS12950.67936
400	DS3462.851161	DS84 54.841013	DS13450.51933
500	DS3962.751159	DS89 53.05980	DS13950.13926
600	DS4461.881143	DS94 53.38986	DS14449.48914
700	DS4961.281132	DS99 53.38986	DS14950.4 931
800	DS5459.981108	DS10453.43987	DS15449.21909

Table 7 Classification accuracy for NB using occurrences

No of features	OCC – NB PA	DA	CA
	D.S ACC NUM	D.S ACC NUM	D.S ACC NUM
50	DS5 70.491302	DS55 52.78975	DS10552.78975
100	DS1067.191241	DS60 54.251002	DS11052.03961
150	DS1565.181204	DS65 52.95978	DS11549.43913
200	DS2063.561174	DS70 53.7 992	DS12049.43913
250	DS2563.561174	DS75 53.22983	DS12549.75919
300	DS3063.881180	DS80 53.11981	DS13050.4 931
400	DS3562.961163	DS85 51.75956	DS13549.7 918
500	DS4061.931144	DS90 52.51970	DS14048.78901
600	DS4560.851124	DS95 53.22983	DS14549.26910
700	DS5060.361115	DS10053.22983	DS15049.64917
800	DS5560.251113	DS10553.22983	DS15550.24928

Table 8 Classification accuracy for SVM using chi-sqr

No of features	CHI – SVM PA	DA	CA
	D.S ACC NUM	D.S ACC NUM	D.S ACC NUM

50	DS1	72.82	1347	DS5	57.82	1068	DS101	54.46	1006
100	DS6	72.9	1350	DS10	57.71	1066	DS106	54.3	1003
150	DS11	72.92	1337	DS15	57.71	1066	DS111	54.41	1005
200	DS16	72.92	1307	DS20	57.6	1064	DS116	54.46	1006
250	DS21	72.87	1269	DS25	57.6	1064	DS121	54.46	1006
300	DS26	72.92	1225	DS30	57.49	1062	DS126	54.53	1004
400	DS31	72.92	1347	DS35	57.69	1062	DS131	54.35	1004
500	DS36	72.92	1347	DS40	57.44	1061	DS136	54.3	1003
600	DS41	72.82	1345	DS45	57.44	1061	DS141	54.14	1000
700	DS46	72.71	1343	DS50	57.76	1067	DS146	54.35	1004
800	DS51	72.65	1342	DS55	57.44	1061	DS151	54.35	1004

Table 9 Classification accuracy for SVM using DF

No of features	DF – SVM								
	PA			DA			CA		
	D.S	ACC	NUM	D.S	ACC	NUM	D.S	ACC	NUM
50	DS2	72.92	1347	DS52	57.66	1065	DS102	54.41	1005
100	DS7	73.09	1350	DS57	57.06	1054	DS107	53.7	992
150	DS12	72.38	1337	DS62	56.68	1047	DS112	52.46	969
200	DS17	70.76	1307	DS67	57.7	1066	DS117	54.25	1002
250	DS22	68.7	1269	DS72	57.71	1066	DS122	51.54	952
300	DS27	67.94	1225	DS77	57.17	1056	DS127	52.46	969
400	DS32	72.92	1347	DS82	57.49	1062	DS132	54.35	1004
500	DS37	72.92	1347	DS87	57.44	1061	DS137	54.3	1003
600	DS42	72.82	1345	DS92	57.44	1061	DS142	54.14	1000
700	DS47	72.71	1343	DS97	57.76	1067	DS147	54.35	1004
800	DS52	72.65	1342	DS102	57.44	1061	DS152	54.35	1004

Table 10 Classification accuracy for SVM using TF-IDF

No of features	Tf-DF – SVM								
	PA			DA			CA		
	D.S	ACC	NUM	D.S	ACC	NUM	D.S	ACC	NUM
50	DS3	72.92	1347	DS53	57.66	1065	DS103	54.25	1002
100	DS8	72.6	1341	DS58	57.76	1067	DS108	54.84	1013
150	DS13	72.6	1341	DS63	57.71	1066	DS113	54.79	1012
200	DS18	72.49	1339	DS68	57.6	1064	DS118	54.08	999
250	DS23	72.33	1336	DS73	57.6	1064	DS123	54.53	1004
300	DS28	72.22	1334	DS78	58.68	1084	DS128	54.52	1007
400	DS33	72	1330	DS83	59.177	1093	DS133	55.11	1018
500	DS38	70.92	1310	DS88	59.28	1095	DS138	55.17	1019
600	DS43	70.33	1293	DS93	58.9	1088	DS143	54.73	1011
700	DS48	68.43	1264	DS98	57.66	1066	DS148	54.84	1013
800	DS53	67.19	1241	DS103	57.33	1059	DS153	53.54	989

Table 11 Classification accuracy for SVM using IG

No of features	IG – SVM								
	PA			DA			CA		
	D.S	ACC	NUM	D.S	ACC	NUM	D.S	ACC	NUM
50	DS4	72.92	1347	DS54	57.66	1065	DS104	54.41	1005
100	DS9	73.09	1350	DS59	57.06	1054	DS109	53.7	992
150	DS14	72.38	1337	DS64	56.68	1047	DS114	52.46	969
200	DS19	70.76	1307	DS69	57.71	1066	DS119	54.25	1002
250	DS24	68.81	1271	DS74	57.71	1066	DS124	51.48	951
300	DS29	68	1256	DS79	57.71	1056	DS129	52.4	968
400	DS34	67.24	1242	DS84	56.84	1050	DS134	51.922	959
500	DS39	67.08	1239	DS89	54.08	999	DS139	51.92	959
600	DS44	65.78	1215	DS94	55.92	1033	DS144	51.54	952
700	DS49	64.69	1195	DS99	54.68	1010	DS149	51.43	950
800	DS54	63.23	1168	DS104	56.03	1035	DS154	53.49	988

Table 12 Classification accuracy for SVM using OCC

No of features	OCC – SVM								
	PA			DA			CA		
	D.S	ACC	NUM	D.S	ACC	NUM	D.S	ACC	NUM
50	DS5	72.87	1346	DS55	57.66	1065	DS105	53.97	997
100	DS10	72.55	1340	DS60	57.93	1070	DS110	54.57	1008
150	DS15	71.3	1317	DS65	58.14	1074	DS115	53.438	987
200	DS20	67.89	1254	DS70	59.01	1090	DS120	54.41	1005
250	DS25	66.97	1237	DS75	59.01	1090	DS125	53.92	996

300	DS30	67.02	1238	DS80	58.41	1079	DS130	55.22	1020
400	DS35	65.29	1206	DS85	56.47	1043	DS135	53	979
500	DS40	63.61	1175	DS90	56.03	1035	DS140	53.438	987
600	DS45	64.37	1189	DS95	54.57	1008	DS145	53.49	968
700	DS50	64.42	1190	DS100	53.92	996	DS150	52.73	974
800	DS55	63.72	1177	DS105	54.52	1007	DS155	51.81	957

5.1 Evaluation results for ELR-BoW using the NB and SVM methods based on the prediction ...

In this experiment, Naïve Bayes (NB) and support vector machines (SVM) are used with different sizes of features sets, namely, 50, 100, 150, 200, 250, 300, 400, 500, 600, 700 and 800. Five feature selection metrics, namely, chi-sqr, df, tf-idf, ig and occ, were used over those different sizes of feature sets. Tables [3] –[7] show the results for the NB classifier, while Tables [8] –[12] show the results for the SVM classifier. To measure the performance of the two classification methods, we focused on the percentage of accuracy for the test set and correctly classified the instances for each news article. The performance measurement assesses the ability of the ELR-BoW algorithm using the feedback measurements, which are PA, DA, and CA, to evaluate the best accuracy between two classifiers and the best number of correctly classified instances using different sizes of feature sets.

5.1.1 Finding 1: Investigate the performance of ELR-BoW using NB against three class labels

Evaluation of the effectiveness of the ELR-BoW using naïve baye NB based on the weighted accuracy. The aim of this test is to compare the performance of proposed feedback measure PA against the state-of-the-art feedback measure (DA, CA), and the impact of different feature representations on the prediction accuracy using naïve Bayes classifier. Tables [3] –[7] tabulate the results for the NB classifier that used different statistical metrics using the PA, DA and CA class labels to measure the price fluctuations for the stock price. The implementation of PA achieves the highest accuracy in (DS11, DS32), with an accuracy of 73.09%, the number of correctly classified instances was 1314 and 1300 for chi-sqr and df respectively, while for DA, the best result that was reported for (DS61) and achieved an accuracy of 58.20% for chi-sqr. The CA scored the lowest accuracy compared to the other class labels, and the best accuracy is in (DS111), with 53.87% for chi-sqr as well.

The obtained results show that the PA achieved a higher accuracy than the DA and CA using different statistical measurement in all of the test datasets. The performance of chi-sqr achieves the best results using the three feedback measures. We also notice that the best accuracy are recorded when the number of features size is between (150–400). It is indicated that using these features number have a remarkable influence on the feedback measurements on stock price movements. From this point, we can conclude that the previous studies were focusing on introducing a stock price models, rather than investigating the performance of the feedback measurements. The strong determination of the extracted features and the stock price for short timeline based on 20 minutes using ELR-BoW yielded to significant enhancements in PA the performance.

5.1.2 Finding 2: Investigate the performance of ELR-BoW using SVM against three class labels

In order to determine the performance of ELR-BoW using SVM the prediction accuracy and the correctly classified instances. Also, in this test we compare the performance of proposed feedback measure PA against the state-of-the-art feedback measure DA and CA. In Tables [8] –[12] , the SVM classifier was implemented similarly to the same datasets. The obtained results demonstrated that the PA in (DS7 and DS9) for DF and IG respectively. The number of correctly classified instances was 1350 and scored an accuracy of 73.09% using 100 features for both. We also note that the accuracy decreased using the DA and CA in (DS83 and DS130), which had best accuracies of 59.17 and 55.22, respectively.

Based on the obtained results, we observed that the PA also achieved better results compared to DA and CA, which were due to the effectiveness of the ELR-BoW algorithm to measure the feedback and represent the features for the stock price modeling. To be more exact, when the linear regression was used, the accuracy was significantly increased for PA, as can be seen in Tables [8] –[12] . Comparing the other statistical metrics, we found that PA performs betters than the other class labels as well. The drop in the accuracy for DA and CA was caused by the inaccurate evaluation for the class label.

The achieved results indicate that PA obtained a significant performance to understand the impact of news articles on the stock price. Additionally, the implementation of ELR-BoW proved to be a successful improvement in discovering the relationships and representing the direct effect of news articles on the stock prices. In addition, ELR-BoW for short time-line intraday stock prediction had a strong impact when exploring different feature representations for the stock prices. The results might be useful for market traders, whereas the results were that it was easier to predict the stock prices efficiently. In addition, the results make logical sense for clarifying realistic stock price fluctuation behavior, to show that the prediction is close to the eventual outcomes. From this point onward, we want to shed light on the impact of the ELR-BoW implementation for feature representation. It is evident that the proposed class label PA has significant enhancements for all of the statistical metrics.

5.2 Evaluate the performance of PA using the NB and SVM classification methods.

Based on findings 1 and 2, we found that PA scored the best results in both classification methods, NB and SVM. In this section, we evaluate the classification methods using the PA feedback measurements for NB and SVM. We calculate the F-measure value for each direction (up, down) and the weighted accuracy for PA-SVM against PA-NB. Tables [12] –[16] show the results for the classification methods using different feature sizes and feature selection metrics, chi-sqr, df, tf-idf, ig and occ. In this evaluation, we will focus on the F-measure value for the up direction and the weighted accuracy.

Table 13 The classification results for the chi-sqr

No of features	**CHI – SVM**			**CHI – NB**		
	F-measure			**F-measure**		
	Down	**Up**	**W.ACC**	**Down**	**Up**	**W.ACC**
50	0.125	0.839	0.663	0.078	0.842	0.654
100	0.128	0.839	0.664	0.078	0.842	0.654
150	0.129	0.84	0.665	0.078	0.842	0.654
200	0.129	0.84	0.665	0.078	0.842	0.654
250	0.129	0.839	0.664	0.078	0.842	0.654
300	0.129	0.84	0.665	0.078	0.842	0.654
400	0.129	0.84	0.665	0.078	0.842	0.654
500	0.129	0.84	0.665	0.078	0.842	0.654

600	0.128	0.839	0.664	0.078	0.744	0.623
700	0.128	0.838	0.663	0.242	0.74	0.617
800	0.128	0.838	0.663	0.078	0.842	0.654

Table 14 The classification results for the DF

No of features	**DF – SVM**			**DF – NB**		
	F-measure			**F-measure**		
	Down	**Up**	**W.ACC**	**Down**	**Up**	**W.ACC**
50	0.129	0.84	0.665	0.122	0.822	0.649
100	0.133	0.841	0.666	0.177	0.807	0.652
150	0.124	0.836	0.661	0.194	0.805	0.655
200	0.123	0.825	0.652	0.194	0.791	0.644
250	0.147	0.808	0.646	0.227	0.779	0.643
300	0.202	0.799	0.652	0.238	0.769	0.638
400	0.129	0.84	0.665	0.078	0.842	0.654
500	0.129	0.84	0.655	0.078	0.842	0.654
600	0.128	0.839	0.664	0.078	0.842	0.654
700	0.128	0.838	0.663	0.078	0.842	0.654
800	0.128	0.838	0.663	0.78	0.842	0.654

Table 15 The classification results for the TF-IDF

No of features	**Tf-IDF – SVM**			**TF-IDF – NB**		
	F-measure			**F-measure**		
	Down	**Up**	**W.ACC**	**Down**	**Up**	**W. ACC**
50	0.122	0.839	0.663	0.13	0.827	0.655
100	0.125	0.838	0.662	0.125	0.827	0.654
150	0.122	0.838	0.661	0.13	0.821	0.651
200	0.124	0.837	0.661	0.143	0.819	0.652
250	0.12	0.836	0.66	0.146	0.81	0.647
300	0.114	0.835	0.658	0.166	0.81	0.651
400	0.107	0.834	0.655	0.165	0.809	0.65
500	0.118	0.826	0.652	0.164	0.805	0.647
600	0.138	0.821	0.653	0.176	0.786	0.636
700	0.139	0.807	0.642	0.182	0.777	0.63
800	0.144	0.797	0.636	0.195	0.777	0.633

Table 16 The classification results for the IG

No of features	**IG – SVM**			**IG – NB**		
	F-measure			**F-measure**		
	Down	**Up**	**W.ACC**	**Down**	**Up**	**W. ACC**
50	0.129	0.84	0.665	0.122	0.822	0.649
100	0.133	0.841	0.666	0.177	0.807	0.652
150	0.124	0.836	0.661	0.194	0.805	0.655
200	0.123	0.825	0.652	0.194	0.791	0.644
250	0.148	0.809	0.646	0.227	0.779	0.643
300	0.198	0.8	0.652	0.238	0.769	0.638
400	0.213	0.793	0.65	0.229	0.755	0.626
500	0.26	0.788	0.658	0.251	0.752	0.629
600	0.26	0.777	0.65	0.254	0.744	0.623
700	0.242	0.77	0.64	0.242	0.74	0.617
800	0.251	0.756	0.632	0.229	0.73	0.607

Table 17 The classification results for the OCC

No of features	**OCC – SVM**			**OCC – NB**		
	F-measure			**F-measure**		
	Down	**Up**	**W.ACC**	**Down**	**Up**	**W.ACC**
50	0.123	0.84	0.663	0.084	0.824	0.642
100	0.08	0.839	0.652	0.158	0.796	0.639
150	0.125	0.828	0.655	0.175	0.779	0.63
200	0.121	0.804	0.636	0.18	0.766	0.622
250	0.176	0.794	0.641	0.185	0.767	0.624
300	0.208	0.792	0.648	0.209	0.766	0.629
400	0.217	0.777	0.639	0.219	0.757	0.625
500	0.255	0.759	0.635	0.23	0.747	0.62
600	0.24	0.767	0.637	0.249	0.735	0.616
700	0.271	0.765	0.643	0.244	0.731	0.611
800	0.289	0.757	0.641	0.248	0.73	0.61

5.2.1 Finding 3: The effectiveness of PA on the classification accuracy

In Tables [13] –[17] , we show the results that were obtained using the PA class label for both classifiers, NB and SVM. The classification results achieved a weighted accuracy for chi-sqr that reached 0.666% for SVM and 0.654 for NB. In addition, the F-measure for the news articles in the up direction achieved a score of 0.84 and 0.842, respectively. The results show that the implementation of SVM in different features sizes is better than NB.

Figures [NaN] –[NaN] represent the weighted classification accuracy for five statistical measures CHI-SQR, DF, TF-IDF, IG and OCC respectively. The weighted accuracy measured using the naïve Bayes and SVM algorithms, for different feature sizes. According to [[14] , [16]], the NB achieved promising results, and therefore, we used NB to compare the results against SVM. The five figures shows that the SVM trend line model is better than NB across all the comparisons. In Fig. [NaN] , the highest score recoded is 0.665 and 0.654 for SVM and NB respectively, the results also shows that the SVM performance was slightly change using all the features. While the NB performance dramatically decrease when using a large number of features. This indicates that the NB performance have weak performance to classify the large number of features.

In Fig. [NaN] , the DF performance of the both classifiers have recorded a drop in accuracy when using [150–300] features, then, the performance slightly increase from features 400 to 800. The highest accuracy reported for SVM is 0.666 when using 100 features and the high accuracy for NB is 0.655 when using 150 features. The results for DF provide a strong evidence to the impact of the features representation on the features performance. Similarly, the Figs [NaN] –[NaN] prove that the SVM is better than the NB in classify the stock market data. The reported results show that there is negative relationship between the number of features and performance, the trend line decrease based increasing the number of features. The plotted figures show that using small number of features is better than high numbers.

The implementation of statistical measures assists in exploring a wide range of features that lead to discovering more relationship of the market movement, and the results indicated that the selected features utilize the characteristic of the statistical measures for feature representation. The ELR-BoW was able successfully to identify a strong features that represents the condition of stock market. We can conclude an important remark that is related to feature selection, and it is obvious with regard to the classifier performance that with feature selection, the accuracy increases due to reducing the number of irrelevant features on the training test set.

5.3 Evaluation results for the impact of different feature sizes on the classification ...

To answer evaluate the impact of different feature sizes on the classification accuracy, the proposed algorithm ELR-BoW for the feature was tested on two classifiers SVM and NB on using 11 different feature sizes. The main purpose of using different feature sizes is to identify the best number of feature size to discover a strong correlation between the extracted features. In addition, there are five statistical metrics have been used to select features in a different representation. Based on findings 1 and 2, the weighted accuracy for PA is significantly better when compared with CA and DA. Therefore, the results for CA and DA are discarded from the analysis.

To evaluate the impact of different feature sizes, we used F-measure for PA-SVM and PA-NB to compare the performance of different feature sizes and statistical metric. Tables [13] –[17] summarize the results for the SVM and NB classifier using the PA class label in a different feature. From the above tables, plotted a five Figs [NaN] –[NaN] for feature selection (CHI-SQR, DF, TF-IDF, IG and OCC) statistical measures.

In this section, the obtained results were further analyzed by implementing statistical analysis using paired sample t-test to evaluate the performance of the proposed method PA-SVM compared against PA-NB. The results are presented in Tables [18] –[20] . The mean of the best features number (M) and their standard deviation (SD) are calculated in terms of F-measure values for each classification methods are presented in Table [18] . In addition, for each feature size, Tables [19] and [20] present the correlation between the features, significant value, and the P-value.

In Table [18] , we reported the (M and SD) for SVM and NB in each feature size. The standard deviation value to evaluate the distribution of the data and to know whether a specific data point is standard and expected or unusual and unexpected. A low standard deviation tells us that the data are closely clustered around the average, while a high standard deviation indicates that the data are dispersed over a wider range of values.

Table 18 Standard deviation and mean values for PA using SVM and NB

Feature size	Dataset	PA – SVM			PA – NB		
		Mean	Std. deviation	Std. error mean	Mean	Std. deviation	Std. error mean
50	DS1–5	0.8396	0.00055	0.00024	0.8274	0.00841	0.00376
100	DS6–10	0.8396	0.00134	0.00060	0.8158	0.01843	0.00824
150	DS11–15	0.8356	0.00456	0.00204	0.6494	0.36375	0.16268
200	DS16–20	0.8262	0.01417	0.00634	0.8018	0.02927	0.01309
250	DS21–25	0.8172	0.01949	0.00871	0.7954	0.03053	0.01366
300	DS26–30	0.8132	0.02247	0.01005	0.7912	0.03374	0.01509
400	DS31–35	0.8168	0.02968	0.01327	0.8010	0.04324	0.01934
500	DS36–40	0.8106	0.03584	0.01603	0.7976	0.04647	0.02078
600	DS41–45	0.8086	0.03439	0.01538	0.7702	0.04477	0.02002
700	DS46–50	0.8036	0.03535	0.01581	0.7660	0.04603	0.02058
800	DS51–55	0.7972	0.04075	0.01822	0.7842	0.05614	0.02511

According to the results in Table [18] , we summarize that the results indicate that the ELR-BoW assist the SVM to produce better results than NB classifier. The mean value M is larger when the number of features is small, and then, the results start to decrease while the number of features increases. In contrast, the standard deviation value SD achieved 0.00055 and 0.00841 when the number of features was 50 and increase dramatically to 0.04075 and 0.05614 at 800 features for SVM and NB respectively. The best mean value (M) is 0.8396 for SVM while the best M for NB is 0.8274. it can be clearly seen in the table that the SD is in SVM is lower than SD in NB which indicated the SVM is better using all different features sizes.

The results demonstrated the effect of having a large number features on the classifier results. Using a lower number of features minimizes the number of irrelevant features in the training set and results in an increase in the performance. On the other hand, with an increased number of features, the number of irrelevant data increases, and the accuracy decreases due to the curse of dimensionality reduction.

In Table [19] , the correlation R between the features for SVM and NB are calculated. The correlation is used to measure the relationship between two variables. The correlation is detonated by R, which is commonly used to represent a linear regression line between two values. The R value can be range from -1 to 1. Where -1 means that is a negative relationship to value and +1 that shows that there is a very storing relationship. While the value 0 (zero) indicates no relationship between the two variables. The results indicate that there is a negative correlation (-0.770 and 0.569) when using a low number of features of 50 and 100 respectively. When the number of features increases there is a high effect on the correlation value, a strong linear relation (> 0.70) appears whereas a positive correlation (0.941, 0.939, 0.961, 0.952, 0.956 and 0.998) for the features (200, 250, 300, 400, 500 and 800).

Table 19 Shows the correlation and the significant results between the PA-SVM and PA-NB

	Correlation between SVM and NB		
Feature size	**Dataset**	Correlation	Sig.
50	**DS1–5**	-0.770	0.128
100	**DS6–10**	-0.469	0.425
150	**DS11–15**	0.012	0.985
200	**DS16–20**	0.941	0.017
250	**DS21–25**	0.939	0.018
300	**DS26–30**	0.961	0.009
400	**DS31–35**	0.952	0.013
500	**DS36–40**	0.956	0.011
600	**DS41–45**	0.629	0.256
700	**DS46–50**	0.612	0.273
800	**DS51–55**	0.998	0.000

The best correlation value R = 0.998 scored when using 800 features and the significant = 0.000143. Based on the obtained results, it seems very hard to assume that 800 features were the best results, the high correlation value might occur due to the increase of the number of absence features which might lead to generate a dataset contains a very large number of (0) value. In this case, the learning process in the classifier is affected and the obtained results are inaccurate. We believe that the best correlation value R = 0.961 scored when using 300 features and the significant = 0.009.

Table 20 Presents the t-test p-value for two variables SVM and NB

		Mean	**Std. deviation**	**Std. error mean**	**Confidence interval of the difference**		**t df**	**Sig. (2-tailed)**
					Lower	**Upper**		
50	DS1–5	0.01220	0.00884	0.00395	0.00122	0.02318	30.0854	0.037
100	DS6–10	0.02380	0.01910	0.00854	0.00009	0.04751	20.7874	0.049
150	DS11–15	0.18620	0.36373	0.16267	-0.26543	0.63783	10.1454	0.316
200	DS16–20	0.02440	0.01664	0.00744	0.00374	0.04506	30.2794	0.031
250	DS21–25	0.02180	0.01395	0.00624	0.00447	0.03913	30.4934	0.025
300	DS26–30	0.02200	0.01366	0.00611	0.00504	0.03896	30.6024	0.023
400	DS31–35	0.01580	0.01753	0.00784	-0.00596	0.03756	20.0164	0.114
500	DS36–40	0.01300	0.01616	0.00722	-0.00706	0.03306	10.7994	0.146
600	DS41–45	0.03840	0.03535	0.01581	-0.00550	0.08230	20.4294	0.072
700	DS46–50	0.03760	0.03711	0.01659	-0.00847	0.08367	20.2664	0.086
800	DS51–55	0.01300	0.01575	0.00704	-0.00655	0.03255	10.8464	0.139

Moreover, a paired sample t-test conducted to evaluate whether statically significant differences existed between the PA-SVM and PA-NB in different feature sizes in Table [20] . The significant level below (α= 0.05). The tabulated results are scored significant results in 7 out of 11 feature sizes, the highest significant value t(4)= 3.602, p-value is 0.023 using 300 feature. This finding indicates that PA-SVM obtain a significant differences to reject the null hypothesis that the NB is might achieved better than SVM. Also, to prove that using 300 feature for stock market prediction have the ability to discover a temporal relationship for time-series data.

Furthermore, the Wilcoxon test is used to compare between the support vector machines SVM and the naïve Bayes NB statistical measures. The analyses of Wilcoxon statistical test is based on average value of F-measure value for each feature number. The results in Table [21] below, show that SVM is significantly better when compared with the NB, whereas the significant level below (α= 0.05). The reported significant value is t(10)=-2.936, p-value is 0.003.

Table 21 The comparisons between SVM and NB using Wilcoxon test

	SVM and NB
Z	-2.936 b
Asymp. sig. (2-tailed)	0.003

a. Wilcoxon Signed Ranks test; b. Based on positive ranks.

This finding indicates that there is a significant difference between the SVM and NB. From this point, we should shed a light to strength of SVM to predict stock market movements. In order to determine differences between SVM and naïve Bayes feature selection measures, it’s highly suggested to rank the statistical measures using the Friedman’s test based on the obtained F-measure value.

The obtained results are further analyzed using the Friedman's test for PA-SVM and PA-SVM. The test is used to rank the statistical measures, the results are tabulated in Table [22] . It can be clearly seen in Table [22] , for PA-SVM the best performing feature selection measurement was DF, with rank 2.5, whereas the worst one was IG, with rank 5.318. Moreover, the results for PA-NB, shows also the best statistical measurement was DF, with rank 5.3182 and the worst was OCC, with score 9.681. In addition, the p-value was calculated using Friedman's test (0.000426), the p-value showed highly significant differences.

Table 22 Average ranking of PA-SVM and PA-NB for different feature sizes

#	Feature selection methods using PA	Ranking
1	DF-SVM	2.5
2	(Chi-sqr) SVM	3.0455
3	(TF-IDF) SVM	3.818
4	(OCC) SVM	4.227
5	(IG) SVM	5.318
6	(DF) NB	5.3182
7	(Chi-sqr) NB	5.7273
8	(IG) NB	6.5
9	(TF-IDF) NB	8.8636
10	(OCC) NB	9.681
	P-value	0.000426

Based on statistical test analysis, the best number of features for the stock price at 300 features, the experiment results recorded the best performance for the DF as the best statistical measures, the best F-measure value obtained was 0.842. Moreover, the representations of the bag-of-word features using different statistical metrics have increased the flexibility to express the extracted features based on the characteristic of each statistical metric to capture the most discriminating features in spite of the results being slightly similar

We believe that the results met our expectations, we can conclude that we have a successful implementation of the proposed method ELR-BoW and feedback measure PA-SVM is robust for building correlation features between the news articles and stock prices. The results testify that there is an improvement in predicting the stock market

6. Conclusions and future work

In summary, our research introduced ELR-BoW algorithm for feature representation for stock market prediction. The performance of the proposed method and measured the effect of financial news articles on the S&P500 stock market. The news articles were represented as features, and the feature vectors were constructed using five statistical metrics to select the best features. Then, the class label examined the close prices using linear regression to calculate three different representations, namely, PA, DA and CA. The naïve Bayes (NB) support vector machines SVM classifier was trained to evaluate the performance in terms of correctly classified instances and the accuracy of the whole test set. Additionally, the F-measure and weighted accuracy are used to indicate the changes that occur in the stock price with the two categories, up/down.

In general, the results were satisfactory because they answered the research objectives, which were to identify the best feature extraction model using five statistical metrics, chi-sqr, DF, TF-IDF, IG, and occurrence. It was found that ELR-BoW using SVM obtained better performance than ELR-BoW using NB using the three feedback measures PA, DA, CA. The DF obtained the best performance compared to other statistical metrics, and the implication of different feature representations using the ELR-BoW algorithm helped to capture the stock market's sudden movements for short-timeline prediction. In addition, the results demonstrated the remarkable improvements in the performance using the proposed PA class label to measure the feedback between the stock price and the published news articles and introduced an accurate prediction model for the S&P500 stock market using linear regression tackled the issue of stock market prediction using short timeline movement based on a 20-min timeline prediction.

Additionally, the experimental results obtained a remarkable significance while capturing the relationship between the news and the stock prices. The ELR-BoW for SVM successfully achieved high significant correlation between the features, the p-value was 0.023 at a feature set size of 300. Moreover, the best ranking was for DF with an F-measure value of 0.842. Thereby, the implementation of statistical measures assisted in exploring a wide range of features, which led to discovering more relationships of the market movement. The results indicated that the selected features help to utilize the prediction accuracy for the stock market. Therefore, this study addresses the pre-processing concerns for text mining by implementing a prediction model that integrates features from financial news and stock price value series based on a 20-minute time series, to utilize feature representations, not only to testify to the system performance but also to capture the impact of the news articles on the stock market. The main implications of this study are briefly summarized below:

This work is considered to be different from previous studies by the nature of building the dataset. The superiority lies in using a large number of statistical measures to select the features and to delve into feature representation enhancements using the linear regression method. Additionally, this work shows an emphasis on investigating the relationship of the stock price using a feature selection method that incorporates five statistical metrics for stock market prediction. That approach captures relationships that demonstrate the interactions between the news articles and the stock prices to predict the movement into two directions up or down.

Despite the significant outcomes from this study, there are still some weak points that are open for debate. From this perspective, we propose a possible direction for future research that requires further vigorous investigation. In our work, we do not include any semantic method to select the features to reflect the condition of the market and understand the vagueness. Thus, we foresee focusing on integrating some distinct features that might be considered. To focus on text mining for market prediction techniques, we have not found any method that is dedicated to context capturing or abstraction methods that entail the required information for the stock market. Because this domain is an emerging field, the necessity for such methods is strongly required. The utilization of computational processing must be investigated rigorously. Last, given the availability of a staggering amount of online data, the implication of dimensionality reduction methods is highly recommended for further enhancements in the field of market prediction.

Acknowledgments

This work is supported by Malaysia Ministry of Education Exploratory Research Grant Scheme (ERGS/1/2013/ICT07/UKM/02/4).

References

1 *W. Antweiler and M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, The Journal of Finance. 59 (2004), 1259 – 1294.*

2 *S. Armstrong, K. Church and P. Isabelle, Natural language processing using very large corpora: Springer Publishing Company, Incorporated, 2014.*

3 P. Azar and A.W. Lo, The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds, Available at SSRN 2756815, 2016.

4 M.R. Borges, Efficient market hypothesis in European stock markets, The European Journal of Finance. 16 (2010), 711 – 726.

5 F.E.T. Burton and S.N. Shah, Efficient Market Hypothesis, CMT Level I 2017: An Introduction to Technical Analysis, 2017.

6 M. Butler and V. Kešelj, Financial forecasting using character n-gram analysis and readability scores of annual reports, in: Advances in Artificial Intelligence, ed: Springer, 2009, pp. 39 – 51.

7 D. Cao, S.L. Pang and Y.H. Bai, Forecasting exchange rate using support vector machines, vol. 6, ed: IEEE, 2005, pp. 3448 – 3452 Vol. 6.

8 L.J. Cao and F.E.H. Tay, Support vector machine with adaptive parameters in financial time series forecasting, vol. 14, ed: IEEE, 2003, pp. 1506 – 1518.

9 G.P.C. Fung, J.X. Yu and W. Lam, Stock prediction: Integrating text mining approach using real-time news, in: Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on, 2003, pp. 395 – 402.

10 L.A. Gajanan, FINANCIAL FORECASTING, Citeseer, 2008.

11 T. Geva and J. Zahavi, Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news, Decision Support Systems. 57 (2014), 212 – 223.

12 G. Gidófalvi and C. Elkan, Using news articles to predict stock price movements, Department of Computer Science and Engineering, University of California, San Diego, 2001.

13 S.S. Groth and J. Muntermann, An intraday market risk management approach based on textual analysis, Decision Support Systems. 50 (2011), 680 – 691.

14 H. Gunduz and Z. Cataltepe, Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection, Expert Systems with Applications. 42 (2015), 9001 – 9011.

15 M. Hagenau, M. Liebmann, M. Hedwig and D. Neumann, Automated news reading: Stock price prediction based on financial news using context-specific features, in: System Science (HICSS), 2012 45th Hawaii International Conference on, 2012, pp. 1040 – 1049.

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

17 H. Harasty and J. Roulet, Modeling stock market returns, The Journal of Portfolio Management. 26 (2000), 33 – 46.

18 T. Joachims, Text categorization with support vector machines: Learning with many relevant features: Springer, 1998.

19 K. Kim and I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, vol. 19, ed: Elsevier, 2000, pp. 125 – 132.

20 R. Kohavi and G.H. John, Wrappers for feature subset selection, Artificial intelligence. 97 (1997), 273 – 324.

21 C.-C. Lee, J.-D. Lee and C.-C. Lee, Stock prices and the efficient market hypothesis: Evidence from a panel stationary test with structural breaks, Japan and the World Economy. 22 (2010), 49 – 58.

22 H. Liu and R. Setiono, A probabilistic approach to feature selection-a filter solution, in ICML (1996), 319 – 327.

23 E. Lupiani-Ruiz, I. GarcíA-Manotas, R. Valencia-GarcíA, F. GarcíA-SáNchez, D. Castellanos-Nieves, J.T. FernáNdez-Breis et al., Financial news semantic search engine, Expert Systems With Applications. 38 (2011), 15565 – 15572.

24 B.G. Malkiel, Efficient market hypothesis, The New Palgrave: Finance. Norton, New York, 1989, 127 – 134.

25 M.-A. Mittermayer, Forecasting intraday stock price trends with text mining techniques, in: System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on, 2004, pp. 10.

26 D.C. Montgomery, E.A. Peck and G.G. Vining, Introduction to linear regression analysis: John Wiley & Sons, 2015.

27 Y. Mukund, V. Naresh, S. Patil, K. Chandrasekaran, V.V. Kumar and R. Gnanamurthy, Influence of News on Individual Confidence Bias in Stock Markets, in: Proceedings of the The 11th International Knowledge Management in Organizations Conference on The Changing Face of Knowledge Management Impacting Society, 2016, pp. 20.

28 A.K. Nassirtoussi, S. Aghabozorgi, T.Y. Wah and D.C.L. Ngo, Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment, Expert Systems with Applications. 42 (2015), 306 – 324.

29 A. Omidi, E. Nourani and M. Jalili, Forecasting stock prices using financial data mining and Neural Network, vol. 3, ed: IEEE, 2011, pp. 242 – 246.

30 D. Peramunetilleke and R.K. Wong, Currency exchange rate forecasting from news headlines, Australian Computer Science Communications. 24 (2002), 131 – 139.

31 J.C. Reboredo, M.A. Rivera-Castro, J.G. Miranda and R. García-Rubio, How fast do stock prices adjust to market efficiency? Evidence from a detrended fluctuation analysis, Physica A: Statistical Mechanics and its Applications. 392 (2013), 1631 – 1637.

32 R.P. Schumaker and H. Chen, A quantitative stock prediction system based on financial news, Information Processing & Management. 45 (2009), 571 – 583.

33 R.P. Schumaker and H. Chen, Textual analysis of stock market prediction using breaking financial news: The AZFin text system, ACM Transactions on Information Systems (TOIS). 27 (2009), 12.

34 R.P. Schumaker, Y. Zhang, C.-N. Huang and H. Chen, Evaluating sentiment in financial news articles, Decision Support Systems. 53 (2012), 458 – 464.

35 T.O. Sprenger, A. Tumasjan, P.G. Sandner and I.M. Welpe, Tweets and trades: The information content of stock microblogs, European Financial Management. 20 (2014), 926 – 957.

36 S. Vijayarani, M.J. Ilamathi and M. Nithya, Preprocessing Techniques for Text Mining-An Overview, International Journal of Computer Science & Communication Networks. 5 (2015), 7 – 16.

37 N. Vlastakis and R.N. Markellos, Information demand and stock market volatility, Journal of Banking & Finance. 36 (2012), 1808 – 1821.

38 B. Wuthrich, V. Cho, S.-W. Leung, K. Sankaran and J. Zhang, Daily stock market forecast from textual web data, in: Systems, Man, and Cybernetics, 1998 IEEE International Conference on, 1998, pp. 2720 – 2725.

39 S.Y. Yang, Q. Song, S.Y.K. Mo, K. Datta and A. Deane, The Impact of Abnormal News Sentiment on Financial Markets, Available at SSRN 2597247, 2015.

40 A. Yoshihara, K. Seki and K. Uehara, Leveraging temporal properties of news events for stock market prediction, Artificial Intelligence Research. 5 (2015), 103.

41 Y. Yu, W. Duan and Q. Cao, The impact of social and conventional media on firm equity value: A sentiment analysis approach, Decision Support Systems. 55 (2013), 919 – 926.

42 Y. Zhai, A. Hsu and S.K. Halgamuge, Combining news and technical indicators in daily stock price trends prediction, in: Advances in Neural Networks – ISNN 2007, ed: Springer, 2007, pp. 1087-1096.

43 W. Zhang and S. Skiena, Trading Strategies to Exploit Blog and News Sentiment, in Icwsm, 2010.

44 X.J. Zhou and T.S. Dillion, A statistical-heuristic feature selection criterion for decision tree induction, IEEE Transactions on Pattern Analysis and Machine Intelligence. 13 (1991), 834 – 841.

Graph: Figure 1. General architecture of ELR-BoW.

Graph: Figure 2. Enhanced BoW (eBoW) representation algorithm.

Graph: Figure 3. Statistical measures for BoW representation.

Graph: Figure 4. Stock price representation algorithm.

Graph: Figure 5. Stock price process for mapping between news articles and stock price.

Graph: Figure 6. Close values for a stock price in time series.

Graph: Figure 7. Heuristic feature selection technique.

Graph: Figure 8. The weighted accuracy for the CHI-SQR using SVM and NB-Based PA.

Graph: Figure 9. The weighted accuracy for the DF using SVM and NB-Based PA.

Graph: Figure 10. The weighted accuracy for the TF-IDF using SVM and NB-Based PA.

Graph: Figure 11. The weighted accuracy for the IG using SVM and NB-Based PA.

Graph: Figure 12. The weighted accuracy for the OCC using SVM and NB-Based PA.

Graph: Figure 13. F-measure value for chi-sqr.

Graph: Figure 14. F-measure value for DF.

Graph: Figure 15. F-measure value for TF-IDF.

Graph: Figure 16. F-measure value for IG.

Graph: Figure 17. F-measure value for occurrence.

~~~~~~~~

By Hani A.K. Ihlayyel; Nurfadhlina Mohd Sharef; Mohd Zakree Ahmed Nazri and Azuraliza Abu bakar

---
~~~~~~~~