UNIVERSITI TEKNOLOGI MARA
Cawangan Melaka

ICECoT 2021

International Conference on Emerging Computational Technologies

"Future Trends of Emerging Computational Technologies in the 21st Century"
"Strategizing IR4.0 in Conquering Covid-19 and Beyond"

In Partnership With

Vol. 1, 2021
International Conference on Emerging Computational Technologies

UNIVERSITI TEKNOLOGI MARA | Cawangan Melaka

INTERNATIONAL CONFERENCE ON EMERGING COMPUTATIONAL TECHNOLOGIES (ICECoT 2021)

24 - 25 August 2021

978-967-15337©ICECoT2021

First Edition 2021

Published by:

Universiti Teknologi MARA Melaka Kampus Jasin

978-967-15337©ICECoT2021

Preface

This e-book describes the research papers presented at the International Conference on Emerging Computational Technologies (ICECoT 2021), organised by Faculty of Computer and Mathematical Sciences (FSKM), UiTM Cawangan Melaka. The main discussions of the conference is on the technological advances that help shape the skills that are required to cope with the Fourth Industrial Revolution (IR 4.0). Considering that this is our first attempt at organising a conference, we are therefore greatly honoured that the Universitas Negeri Semarang (UNNES), Indonesia, Mahasarakham University (MSU), Thailand and University of Hail (UoH), Saudi Arabia have all agreed to become our partners by contributing several reseach papers as well as providing reviewers to assess the quality of the papers.

Out of the numerous research works that had been submitted and reviewed, the Editorial Board have selected 22 papers to be published in the e-book. The discussions of these papers pertain to the use of technologies within the broad spectrum of Computer Science, Computer Networking, Multimedia, Information Systems Engineering, Mathematical Sciences and Educational Technology. It is hoped that the research findings that are shared in this e-book can benefit those who are interested in the various areas of computational technologies; such as graduate students, researchers, academicians and the industrial players, to name a few.

As the Project Manager, I would like to thank all of the committee members from the bottom of my heart for their tireless efforts in ensuring the success of ICECoT 2021. Without their continual support and excellent teamwork, this conference would not have come to fruition. In fact, holding this major event has been a good learning experience for us all, and I sincerely believe that our future conferences will become more outstanding if the same spirit is maintained.

Dr. Noor Aishikin Adam
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA Cawangan Melaka Kampus Jasin

978-967-15337©ICECoT2021

INTERNATIONAL CONFERENCE ON EMERGING COMPUTATIONAL TECHNOLOGIES (ICECoT 2021)

COMMITTEE

GENERAL CHAIR

NOR FADILAH TAHAR @ YUSOFF

PROJECT MANAGER

DR. NOOR AISHIKIN ADAM

DEPUTY PROJECT MANAGER

ZAINAL FIKRI ZAMZURI

SECRETARY

FADHLINA IZZAH SAMAN

DR. RAIHAH AMINUDDIN

TREASURER

UMMU MARDHIAH ABDUL JALIL

ANIS AFIQAH SHARIP

MOHD AZRIL SHAH ABDULLAH

NURUL AIMI MD NOR

INTERNATIONAL RELATION

DR. ELIN ELIANA ABDUL RAHIM *(Leader)*

NORBAHIYAH AWANG

NURULHUDA ZAINUDDIN

ZUHRI ARAFAH ZULKIFLI

SITI FATIMAH MOHD RUM

NOOR WAHIDA JAMIL

MOHAMMAD BAKRI CHE HARON

MUHAMMAD HAMIZ MOHD RADZI

DESIGN /MULTIMEDIA/MONTAGE/E-CERTIFICATE

DR. AHMAD FIRDAUS AHMAD FADZIL *(Leader)*

NORSHAHIDATUL HASANA ISHAK

MOHD NABIL ZULHEMAY

NUR FARAHIN MOHD JOHARI

NURAZIAN MIOR DAHALAN
NOR INTAN SHAFINI NASIRUDDIN
HAZRATI ZAINI
ZAINAB OTHMAN
FAIQAH HAFIDZAH HALIM

SPONSORSHIP

DR. NURUL HUDA NIK ZULKIPLI *(Leader)*
Ts. DR. SITI RAHAYU ABDUL AZIZ
MOHD RAHMAT MOHD NORDIN
NORZATUL BAZAMAH AZMAN SHAH
NOR AIMUNI MD RASHID

SECRETARIAT & REGISTRATION

DR AZLAN ABDUL AZIZ *(Leader)*
DR. SITI FEIRUSZ AHMAD FESOL
FADZLIN AHMADON
MARIATHY KARIM
NUR ASYIRA NAZIRON
HAJAR IZZATI MOHD GHAZALLI
MASWATI SUFFIAN
MASTURA MANSOR
NURULHUDA GHAZALI
WANNORAINI ABDUL LATIF

REVIEWER

Ts. DR. SHAFAF IBRAHIM *(Leader)*
DR. NOOR HASIMAH IBRAHIM TEO
Ts. DR. ALYA GEOGIANA BUJA
NUR NABILAH ABU MANGSHOR
Ts. DR. CHEW CHIOU SHENG
HAMIDAH MUHD IRPAN
BUSHRA ABDUL HALIM
NUR SYUHADA MUHAMMAT PAZIL
YUZAIMI YUNUS
NOOR SURIANA ABU BAKAR
NOORAZILAH IBRAHIM
ANIS AMILAH SHARI

978-967-15337©ICECoT2021

NORDIANAH JUSOH @ HUSSAIN
SALEHAH HAMZAH
DR SURYAEFIZA KARJANTO
ASSOC. PROF. DR. PHANG YOOK NGOR
AZLIN DAHLAN

EDITORIAL

Ts. DR. KHYRINA AIRIN FARIZA ABU SAMAH *(Leader)*
Ts. DR. NURUL HIDAYAH MAT ZAIN
Ts. DR. EDZREENA EDZA ODZALY
ASSOC. PROF. AISHAH AHMAD @ ABDUL MUTALIB
SITI NURAMALINA JOHARI
NOOR AFNI DERAMAN
DR. NOR AIZA MOKETAR
ANITA MOHD YASIN
FARAH NADZIRAH JAMRUS
NURUL ZAHIRAH ABD RAHIM
NOR AZYLIA AHMAD AZAM
NOR ADORA ENDUT

PROTOCOL & INVITATION

RAIHANA MD SAIDI *(Leader)*
Ts. DR. JAMALUDDIN JASMIS
IZNI SYAMSINA SAARI
AINON SYAZANA AB HAMID
ANWAR FARHAN ZOLKEPLAY
ANIS SHOBIRIN ABDULLAH SANI
MOHD HAFIFI MOHD SUPIR
ADI HAKIM TALIB
ZURAH ABU
NOR ADILA KEDIN
SHAHITUL BADARIAH SULAIMAN
YUSARIMA MUHAMAD

TECHNICAL / WEBSITE / VIDEO SUBMISSION / Q&A SESSION

SHAHADAN SAAD *(Leader)*
DR. AZMI AMINORDIN
MOHD ALI MOHD ISA

ALBIN LEMUEL KUSHAN
MOHAMMAD HAFIDZ RAHMAT
SULAIMAN MAHZAN
MOHD TAUFIK MISHAN
Ts. NOR AZLIN ROSLI
SITI FAIRUS FUZI
SITI RAMIZAH JAMA
SITI NURSYAHIRA ZAINUDIN
MOHAMAD HAFIZ KHAIRUDDIN
NORAINI HASAN
NURUL EMYZA ZAHIDI
FATIMAH HASHIM
ZAMLINA ABDULLAH
ALIAS DERAMAN

PROMOTION & PUBLICITY

SYAFNIDAR ABDUL HALIM *(Leader)*
NOR AZIDA MOHAMED NOH
FARIDAH SAPPAR
FADILAH EZLINA SHAHBUDIN
Ts. NURUL NAJWA ABDUL RAHID @ ABDUL RASHID
MOHAMAD ASROL ARSHAD

LANGUAGE EDITOR

MOHD AMIRUL ATAN
NUR AQILAH NORWAHI

978-967-15337©ICECoT2021

COMPUTER SCIENCE

978-967-15337©ICECoT2021

NETWORKING / NETCENTRIC

INFORMATION SYSTEM ENGINEERING

EDUCATIONAL TECHNOLOGY

978-967-15337©ICECoT2021

MATHEMATICAL SCIENCES

MULTIMEDIA

978-967-15337©ICECoT2021

News Sentiment and Actual Price of Stock Data: Using News Classification Technique

Anupong Sukprasert
Mahasarakham Business School
Mahasarakham University
Thailand
anupong.s @acc.msu.ac.th

Weerasak Sawangloke
Mahasarakham Business School
Mahasarakham University
Thailand
weerasak.s@acc.msu.ac.th

Benchamaphorn Sombatthira
Mahasarakham Business School
Mahasarakham University
Thailand
benchamaphorn.s@acc.msu.ac.th

***Abstract*— The objective of this study is to examine the relationship between news sentiment and actual price of stock data by using news classification technique. The effects of online news towards stock market turning points. This investigation studies the methods of news sentiment analysis. There were seventeen companies' data used to analyze the data. News classification techniques was used to sort out key features for further classification. News classification into factors affecting stock market price was done using Naïve Bayes, Deep Learning, Generalized Linear Model (GLM) and Support Vector Machine (SVM). The news classification and news sentiment were used to predict the stock market turning points. Results show that best news classification approach is based on Deep Learning techniques that provide the most accurate classification. The study suggests that the accurate and time saving decision for stock investors.**

Keywords—deep learning, news classification, news sentiment

I. INTRODUCTION

As the news has significant effects on the prediction of the stock market [1], there is a field of study that investigates how computing can be utilised to operate and interpret natural language in the text or speech that appears in the public media. This is called Natural Language Processing (NLP), and it directs how to study how humans use and comprehend language with computer systems [2]. In other words, it is "the intersection of artificial intelligence and linguistics"[3]. Normally, news is not kept as structured data, and analysing news sentiments is one of the best ways to help rearrange data in order, i.e., structured data. Once news sentiments become structured data, this structured data becomes useful for creating rules for turning points in stock market prediction.

The day-to-day news on social media is applied to predict the stock market in the incoming days. The actual stock market price is used to compare with the data from the finance news [4]. For example, on 1/1/2013, there are approximately 10 news stories about Company A. The news sentiment analysis is implemented in order to understand that the 10 news stories are classified into groups of six factors and to specify the polarity of the factors, whether it is positive or negative. After that, summaries of the news on 1/1/2013 are produced. The sentiment of factors related to Company A are represented, with whether the factor is positive or negative being incorporated into rules for predicting the stock market turning points. The results are used to predict the stock market turning points on the next day, on 2/1/2013. So, the prediction framework accomplishes exact forecasts, and the re-enactment on stocks exchanged appears to be a major benefit. Even though the predictions can only be made with a certain degree of accuracy, the ultimate aim is to minimise the risk associated with the predictions [5].

As mentioned previously, these days there are more advanced algorithm systems, such as Artificial Intelligence and learning machines, that aid in processing data. In research by Osanga [6], even though AI was successful in recording information retrieval, data mining and Natural Language Processing (NLP), there is the requirement for opinion mining to automatically deal with a large number of sentimental opinions on the internet. The growth of news distribution on websites and social necessity drives research attempts at sentiment analysis to enhance text mining that can extract the essence of words about the possibility of stock market trends.

Thus, the objective of this study is to examine the relationship between news sentiment and actual price of stock data by using news classification technique. The study problem is "Can sentiment help create rules to approach towards turning points stock market prediction?"

II. LITERATURE REVIEW

A. News Classification

News information is now easily accessible through online content contributors such as news on websites and social media [7]. However, a huge amount of news information is available in textual written forms from various online sources, which could be advantageous in many areas of research but which is difficultto use. As the number of news sources has grown considerably, it is much harder for internet users to approach news of interest which allows users to categorize news accurately in order for the news to be accessed in real time. A news classification technique is one of the vital applications of text mining because it aids in classifying textual news based on words, phrases, and combinations of words, etc., to pre-set class labels. Classification is found to be a challenging task in text mining [8] because it needs pre-processing steps to turn unstructured data into structured data.

A news classification technique performs the first part with a dataset where the class assignments are known. Classification is distinct, and it does not imply order. Its purpose is to precisely forecast the target class for each case in the data [9]. The decision strategy is probably drawn up by scrutinizing the fluctuation of the stock price on the basis of news contributed and sector preference. They performed an analysis of relevant news classification and impact calculation strategies and selected the method of stock prediction to be applied in the model by performing a

978-967-15337©ICECoT2021

comparison of prediction techniques as follows: regression, classification, and statistical techniques.

As a vast number of news articles is published on the internet, it takes a long time to select the most likely sources. Hence, a method of news classification is critical in order to acquire compatible information expeditiously [10]. As a huge amount of current news is published on diverse websites, it is necessary to transform the textual data (unstructured data) into structured data forms [11]. It is practical to mine sentiments of news of interest over a period of time which can be applied to measure the long-term influence on stock investment [12]. The process of news classification is associated with the following main steps: data collection, pre-processing, feature selection, classification techniques application, and evaluating performance measures [13].

B. Techniques of News Classification

There is a variety of news classification techniques which are used in the field of stock market prediction. However, in this study, Deep Learning, Naïve Bayes, Support Vector Machines (SVM), and Generalized Linear Models (GLM) are discussed because they are relevant to the study area. Different techniques have different advantages, for example Deep Learning is related to the way the neutral networks of the human brain help evaluate huge amounts of complex data [14]. Naïve Bayes is more likely to be related to the probability in that it has a distinctive way of explaining and classifying a set of data in order to rearrange different news into the appropriate groups [15]. Support Vector Machine (SVM) is used in a situation in which the data is non-linear to classify it into the proper groups [16]. Lastly, the General Linear Model (GLM) is related to a way to classify data which is linear, and it helps to rearrange data into suitable groups.

Basically, raw data or news (unstructured data), which consists of a huge amount of data, is a mixture of linear and non-linear sets of data which have not been classified as structured data. Hence, the different techniques that were mentioned previously are useful in classifying all kinds of data, both linear and non-linear kinds, and myriad amounts of data. Whether or not data is linear is referred to as the key difference between types of data structure. In linear data structures, the organisation of data foundations is successive, while in nonlinear data structures, the organisation of data foundations is not successive. A data structure is a method for organising and storing data, which then allows for efficient data retrieval and usage. This is necessary because a large volume of data that covers both structured and unstructured types that can overwhelm a business in its routine activities [17]. To sum up, all the techniques that have been mentioned above are suitable for solving the problem of classifying data, in particular for the text form or news.

1) Deep Learning: Deep Learning is an algorithm in the field of machine learning that allows the machine to process a huge amount of data by simulating neural networks in the same way as neutral networks in the human brain. In this section, the topic of deep learning in Pattern Recognition and Machine Learning [18] will be reviewed. At the beginning of research in the field of artificial neural networks, the most used structure was machine learning with python (MLP), in which there is only one layer. Many theories have suggested that if the number of nodes are sufficient, a one-layer structure of MLP can be used to gauge function values.

In 2006, research in this field by Hinton et al. at the University of Toronto, introduced the idea of teaching hidden layers one by one, separately, with unsupervised learning. Then, the separated layers are brought together and finely adjusted again in the form of supervised learning. The study by Hinton et al. is based on artificial neural networks called the Restricted Boltzmann Machine (RBM) that was not previously widely applied. The benefit of such networks is the connection with a probability model that is used to describe why the deeper structures should be created,− namely, there are multi-layer. As a result, this research led to a new field, which is overall called “deep learning”. The idea of teaching hidden layers in succession separately with unsupervised learning brings about another artificial neural network. That is, the structure consists of autoencoder neural networks with multi-layer. Each layer of autoencoder neural networks is an artificial neural network which is used in converting data into a new space with some structures that can be converted back completely. Joining these structures is like attempting to choose suitable data representation.

Modern Deep Learning is not like previous neural networks because it contributes training stability, generalization, and scalability with big data. It also works well for a variety of problems, so Deep Learning is rapidly growing as an alternative algorithm for the highest accuracy in prediction. Many frameworks for Deep Learning have been widely developed. In the Fig. 1, the neuron is the basic unit in the model. Naturally, the output signals of the neurons' diverse strengths move along the synaptic junctions, and they are accumulated as input for the activation of a linked neuron. In the model, the weighted combination:

$$\alpha = \sum_{i=1}^{n} w_i x_i + b$$

of input signals is accumulated. After that, an output signal f(α) is transferred by the linked neuron. The function f serves as the nonlinear activation function implemented via the network, and the bias b serves as the activation threshold of the neuron.

A multi-layer, feedforward neural network is composed of several layers of interlinked neuron units, as presented in the following image, beginning with an input layer to complement the feature space, followed by multiple layers of nonlinearity, and ending with a linear classification layer to complement the output space. The inputs and outputs of the model's units follow the fundamental logic of the one neuron represented in Fig. 1.

Fig. 2, bias units are encompassed in every non-output layer of the network. The weights connecting neurons and biases with other neurons completely verify the output of the whole network. Learning happens when these weights are adjusted to reduce the error on the labelled training data. Specifically, for each training example j, the intention is to reduce a loss function,

$$L(W, B \mid j)$$

where W is the collection $\{W_i\}_{1:N-1}$, where W_i designates the weight matrix linking layers i and $i + 1$ for a network of N layers. Likewise, B is the collection $\{b_i\}_{1:N-1}$, where b_i indicates the column vector of biases for layer $i + 1$. This primary framework of multi-layer neural networks is able to achieve Deep Learning tasks. Deep Learning are thus models of feature extraction, generally relating to multiple levels of nonlinearity. Deep Learning algorithms can learn valuable

978-967-15337©ICECoT2021

representations of raw data and have presented high performance processing complicated data such as images, speech, and text [19].

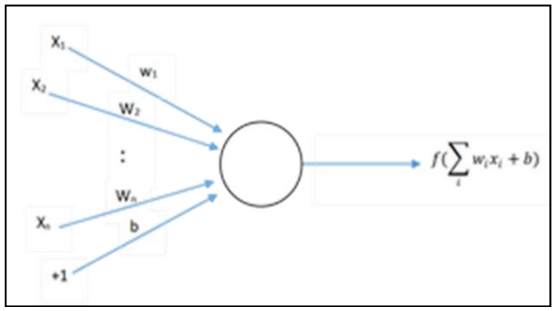

Fig. 1. A single neuron

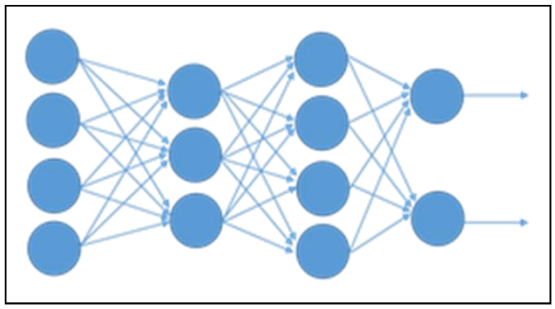

Fig. 2. Multi-Layer Perceptron

2) Naïve Bayes: Naïve Bayes is a popular classification technique which uses a statistical principle to predict conditional probabilities on the basis of Bayes' theorem, with a strong independence hypothesis. The model is titled naive because it assumes that the features are conditionally independent of each feature. This provides the ability to calculate the probabilities of Bayes' formula from a very tiny training set. The algorithm operates on the basis of conditional probabilities. The closing classification rests on the product of two probabilities, which produces the posterior probability. The first probability is termed prior probability, and the second one is termed likelihood probability. The prior probability is unconditional probability. That is to say, it is the state of knowledge before the data is detected. It is computed for each class, and the highest value class is the class of the new object.

In a study by Osanga in 2014, the probability of a tweet belonging to one of the classes (positive or negative) was computed based on the prior probability and the likelihood probability. The class that has the highest probability was chosen as the polarity of the tweet. His approach is based on the naïve hypothesis that the set of features are conditionally independent given a class. Due to this hypothesis, Naïve Bayes gives a high degree of accuracy for text classification purposes with a performance compared to other supervised methods.

Additionally, Naïve Bayes is a method that is popular and widely used in the news classification of documents due to its simplicity and great efficiency in classification. As Naïve Bayes is based on Bayes' Theorem dependent on the probability of result prediction by analysing the relationship between variables to employ in establishing the conditional probability for each relationship.

3) Support Vector Machines (SVM): Support Vector Machines is a supervised learning method that has been widely used to perform classification problems and nonlinear regressions. It is an algorithm that can help in solving classification problems, and it is used to analyse and classify data based on the principle of coefficient of equation to form the linear dividing the data put into the teaching process of the learning system. The idea of Support Vector Machines is predicated on the basis of putting data in feature space, and then the lines separating two data points will be found by forming a hyper-plane which is linear. This is to find out which linear dividing the two data points is the optimal separating hyper-plane. The line separating the data is called the maximum hyper-plane.

The basis of the Support Vector Machine is applied to linear data and can help in such problems with the use of the Kernel Function. Classification of data on multi-dimensional planes will apply the use of the most proper selection, which is called feature selection. The structure of selection is retrieved from the data teaching learning system. The number of structure sets used in describing a case is called the vector. Thus, the model of Support Vector Machines aims at dividing groups of vectors, with one group of variables on one side of the plane and other groups on different planes. All the vectors on the sides of the multi-dimensional planes are called support vectors.

In separating the data, a special rule is constructed, which is called the linear classifier, in which its function is stated as Equation (1).

$$f\ (x;\ w,\ b)\ =\ \langle w,\ x\rangle + b \qquad (1)$$

where w and b are the function parameters perpendicular to the boundary and the "<", ">" sign are the inner product of the two vectors (positive and negative in this case).

The hyperplane ought to be as far away as possible from the two datasets it divides. Fig. 3 shows that the classifier's main purpose is to search out the most effective hyperplane which separates the negative class data points from the positive class data point with the maximum possible margin for every set of points from the hyperplane [20]. The data point on the margins are known as "support vectors". The foremost necessary property of the training data in SVM is to be linearly distinct, suggesting that the two hyperplanes of both margins can be selected in the sense that there are no data points between them.

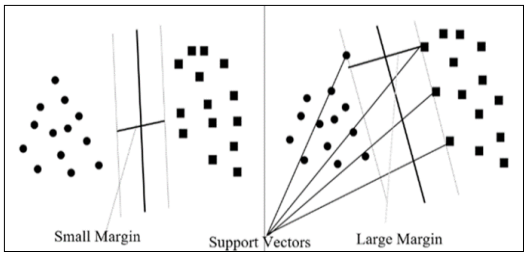

Fig. 3. SVM models depicting marginalisation

4) Generalized Linear Models (GLM): Generalized Linear Models is a group of models including linear models and classical linear models. It is intended to discover the relationship between explanatory variables and response variables on the basis of the idea that Y can be written in the

978-967-15337©ICECoT2021

sum of mean (μ) with deviation (ε). The hypothesis of GLM is detailed as follows:

a) *Random Component:* Y is independent of each other, and it contains a distribution in the exponential family.

b) *Random Component*: Y is independent of each other, and it contains a distribution in the exponential family.

c) *Systematic component:* Explanatory can be written in the form of η linear approximation that $\underline{\eta} = X\underline{\beta}$

d) *Link Function:* Relationship between random component and systematic system is determined by link function (g). Link function can find differentiable, and it is monotonic where

$$E[\underline{Y}] = \underline{\mu} = g^{-1}\underline{\eta} \quad (2)$$

Distribution in exponential family has two features as follows:

- Distribution can be written in the form of mean and variance.
- Variance is a function of mean. The formula can be

$$Var(Y_i) = \frac{\phi V(\mu_i)}{\omega_i} \quad (3)$$

where ϕ is scaled parameter and ω_i is constant determining the weight for each prior weight. Distribution in exponential family and variance of distributions can be concluded as presented in Table I.

TABLE I. VARIANCE OF DISTRIBUTIONS

Distribution	Variance
Normal	1
Poisson	x
Gamma	x^2
Binomial	$x(1-x)$ when the number of tests = 1
Inverse Gaussian	x^3

III. METHODOLOGY

As news is collected in the first step, all the news documents are obtained. The news is derived from the websites and stored in a database, with a total number of 54,849 documents that the annotator is annotating. When the annotation scheme is initially progressing, the documents are examined and discussed until consensus annotations are obtained that could be utilised as a gold standard. After each training document has been annotated, the annotators measure their annotations to construct the gold standard for the document [21].

A. Data Pre-processing

After the data is collected, the news data in datasets is processed via the use of the Natural Language Processing (NLP) technique and transformed into the form of a document matrix. Then it is used to create a model by applying machine learning techniques. For more details, the data processing consists of two sub-pre-processing steps, as shown in Fig. 4.

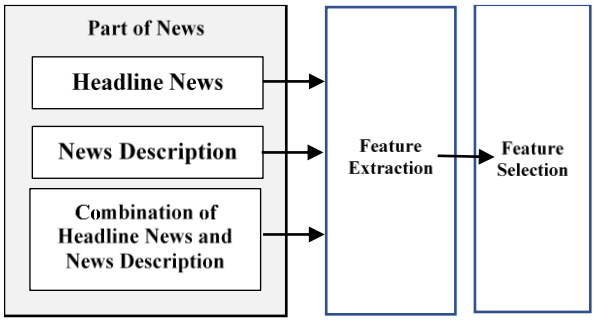

Fig. 4. Procedures in preparing data

Data Query from Database: the news is divided into three parts, and the parts of the news are identified from Headline News: headlines of news, News Description: contents, and Combination of Headline News and News Description. These three parts of the news are processed by the NLP technique to select the fit features before the data are analysed and used in creating the model to classify the category of news.

- Remove punctuation, special characters, numbers and some specific words in order to obtain the data in the form of plain text.
- Convert plain text to lowercases because the study intends to make the word appear exactly the same every time it appears. Thus, the word is changed to lowercase.
- Strip unnecessary whitespace from the document because the above processing will leave the documents with a lot of "whitespace". Whitespace is the result of all the left-over spaces that were not removed along with the words detected, and whitespaces can or should be removed.
- Stemming words is utilised to find the original form of those words or their root words without prefixes and suffixes. Most of them are nouns and verbs, for example, the original word of "walked" and "walking" is "walk". Although "ed" and "ing" is removed, the meaning of this word still remains. Finding the word roots helps in determining the word to be the same word in order to reduce the number of words (or feature redundancy) in the document. In this study, the method of finding word roots by Porter is applied; it is suitable for finding word roots in English.

After data preparation, feature extraction and feature selection re carried out.

B. News Sentiment Analysis

After news data had been collected from websites, news data is prepared by the method of text pre-processing, as shown in Fig. 5. Then news data in textual form and unstructured data is transformed into structured data; data pre-processing is undertaken, and the data is transformed into a document term matrix in order for feature extraction to then find features used for news classification. The sentiment analysis is regularly applied in financial issues, especially the stock market prediction. Feature selection is then carried out in order to select features that influence news classification for analysis. The data used in establishing the model is news about the stocks of 17 companies running businesses related to technologies that is gathered from the Yahoo! Finance website (http://finance.yahoo.com). There are a total of

978-967-15337©ICECoT2021

54,849 news documents. When obtaining data that is well-prepared for analysis, sentiment words are then calculated by a sentiment detection algorithm. After determining the result of sentiment words, the test to find the relationship between news data and stock price is undertaken to confirm that the news has effects on stock fluctuation.

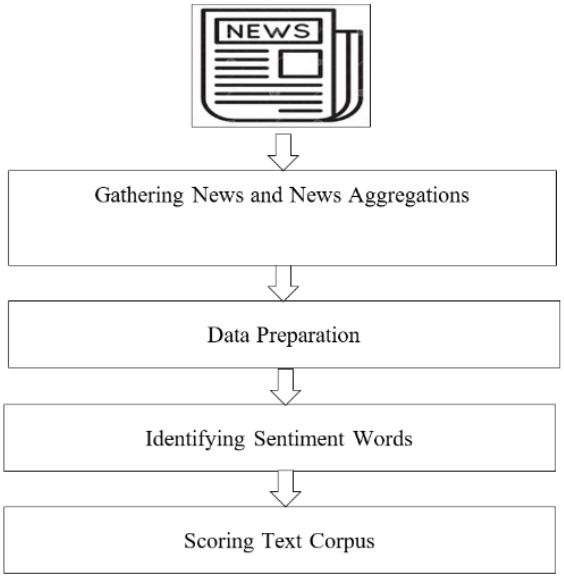

Fig. 5. Text-preprocessing stages

To find the relationship between news sentiment and actual price of stock data, it can analyse in the Fig. 6. Sentiment words as presented are using a simple algorithm that computes the number of incidents of "positive" and "negative" words in the headline news. Based on the calculation, scores are allocated to the related sentences. The polarity of the sentences is then computed in order to discover the sentiment score [22].

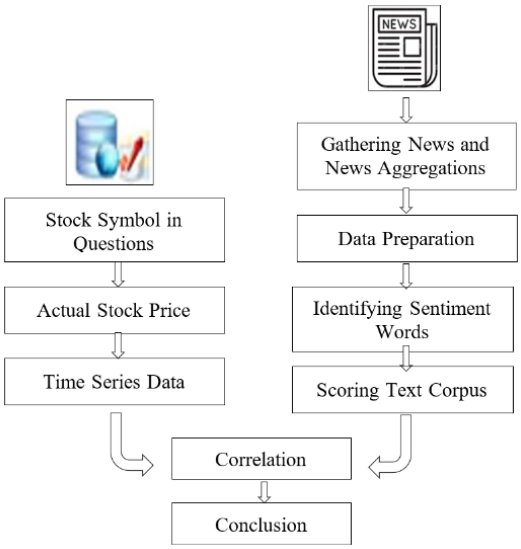

Fig. 6. Relationship between news sentiment and actual price of stock data

C. News Classification Technique

The news classification technique is illustrated in Fig. 7. The training set and testing set are separated with a k-Fold Cross Validation. That is, the data is divided into k sets equally, and the number of k rounds is calculated and examined. In each round, the data k will be selected to be test set data, and the rest of the data, which contains k-1sets, will be used as training sets. 10-Fold Validation will be employed in this test to create the model of news classification, with the use of different machine learning techniques: Naïve Bayes (NB), Support Vector Machines (SVM), Generalized Linear Mode (GLM) and Deep Learning (DL) being implemented to establish the rules of news classification with the use of RapidMiner Studio to aid in analysing. News classification performance is tested by classifying the news with the use of Accuracy, Precision, and Recall, and the number of models obtained from the research in total is 1,296.

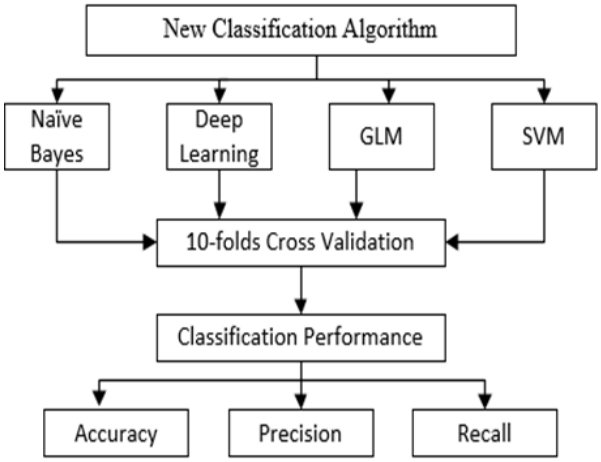

Fig. 7. News classification techniques

After the data is classified and classification performance is calculated, the hypothesis for the part of news, vector representation of text, feature extraction, feature selection and classification techniques will be examined by dividing the news items into six parts to be used in describing the results, namely a classification performance test obtained from the three parts of news, a classification performance test obtained from term weighting, a classification performance test obtained from source of feature from six sources, a classification performance test from three types of trimming the features, a classification performance test obtained from three feature selection techniques, and a classification performance measure test obtained from four classification techniques, in order to compare the efficiency of accuracy.

IV. Research Result

From the study objective and hypothesis is the test of feature selection techniques from the three techniques in the second state in order to find what feature selection techniques approach results in the most accurate feature selection techniques from the three methods used in this study, i.e., Information Gain, Gini Index and Chi-square. This is conducted to discover whether or not each feature trimming method gives different results for accuracy. The outcome of the data analysis is presented in Tables II and Table III.

TABLE II. Test of Homogeneity of Variances

Levene Statistic	df_1	df_2	Sig.
0.066	2	1617	0.936

Based on the results in Table II, it can be concluded that the variance of each group is no different in means.

978-967-15337©ICECoT2021

Therefore, the ANOVA is able to be used in testing this hypothesis.

Hypothesis Testing:

H_0: There is no difference between the means of accuracy of news classification technique (Information Gain, Chi-square and Gini Index)

Based on the analysis outcome in Table III, $p \geq 0.05$. It can be concluded that the result of the accuracy means obtained from news selection technique is statistically not significantly different when the level of significance is 0.05.

TABLE III. ANALYSIS OF VARIANCE (ANOVA)

Source of Variation	Sum of Squares	df	Mean Square	F	Sig.
Treatment	57.566	2	28.783	0.043	0.958
Error	1074072.237	1617	664.238		
Total	1074129.803	1619			

Precision and recall are standard measures that have long been in use for prediction models. The F-measure is the harmonic mean of the precision and recall. Accuracy measures the percentage of samples that are correctly classified. In this research, how to evaluate the effectiveness of the proposed approaches are defined below in terms of a 3-by-3 confusion matrix, as shown in Table IV.

Table IV shows a number of evaluation measures can be defined using these predicted upward, downward and unchanged values. The four metrics that are used in this study**.** After completing the stage of the rule-based optimisation for stock market turning points prediction illustrated, the effectiveness of the stock market turning points prediction in terms of the prediction for the six time periods is shown in Table V.

TABLE IV. TREE-CLASS CONFUSION MATRIX

		Prediction		
		up	down	unchanged
Actual	up	(up, up)	(down, up)	(unchanged, up)
	down	(up, down)	(down, down)	(unchanged, down)
	unchanged	(up, unchanged)	(down, unchanged)	(unchanged, unchanged)
	[Prediction, Actual]			

Table V shows the results of the prediction performance for the prediction of stock market turning points for the yearly period by target class, which is a stock market price that is going up. The results show that the mean of the accuracy is equivalent to 0.5294, while the mean of recall, the mean of precision and the mean of F-Measure are all equivalent to #N/A. *This the result cannot be calculated because the data used for the calculation is missing values (from four companies).

TABLE V. TESTING THE EFFECTIVENESS OF YEARLY PREDICTION BY USING NEWS INFORMATION COMPARED WITH ACTUAL STOCK PRICE DATA

No.	Stock Symbol	Performance			
		Accuracy	Precision: Up	Recall: Up	F Measure : Up
1	AAPL	0.6667	0.6667	1.0000	0.8000
2	AMZN	0.6667	0.6667	1.0000	0.8000
3	BRK-A	0.5000	0.5000	1.0000	0.6667
4	CSCO	0.5000	0.5000	1.0000	0.6667
5	CVX	0.5000	0.0000	#N/A	#N/A
6	EBAY	1.0000	1.0000	1.0000	1.0000
7	FB	1.0000	1.0000	1.0000	1.0000
8	GE	0.3333	0.5000	0.5000	0.5000
9	GOOG	0.3333	0.5000	0.5000	0.5000
10	GOOGL	0.0000	#N/A	0.0000	#N/A
11	IBM	0.0000	0.0000	#N/A	#N/A
12	INTC	0.5000	0.5000	1.0000	0.6667
13	JNJ	0.5000	0.5000	1.0000	0.6667
14	MSFT	1.0000	1.0000	1.0000	1.0000
15	ORCL	0.5000	0.5000	1.0000	0.6667
16	WMT	0.5000	0.5000	1.0000	0.6667
17	XOM	0.5000	#N/A	#N/A	#N/A
Average		0.5294	#N/A	#N/A	#N/A
Standard deviation		0.2814	#N/A	#N/A	#N/A

V. CONCLUSION

In conclusion, the results of the prediction performance for the prediction of stock market turning points in the yearly period, the mean of accuracy for each time period prediction are used to test the hypotheses in order to study the prediction of the different periods of time in order to determine whether or not they are statistically different, which is explained earlier.

More specifically, in this study, the concepts applied included sentiment analysis and rule-based optimisation. It would be helpful for future study to replicate the entire process of investigating the research to acquire a deeper understanding regarding the prediction of stock market price by adding concepts such as big data analytics. It would be useful to provide a very large number of datasets, and this would help to reduce the length of time used at the data analysis stage.

Finally, regarding the unit of study, the stock market was selected as the unit of study; hence, for future study, it would be useful to extend the findings of this study to apply news on social media to predict the prices of crude oil, gold, and the exchange rate. In summary, the sentiment can help create rules, and these rules are practical in optimising the rule-based approach towards turning points stock market prediction. The objective was answered from the study results.

ACKNOWLEDGMENT

This research was supported by Mahasarakham Business School, Mahasarakham University, Thailand.

REFERENCES

[1] R. P. Schumaker and H. Chen, “Textual analysis of stock market prediction using breaking financial news: The AZF in text system”, ACM Transactions on Information Systems (TOIS), 27(2), 1-19, 2009.

[2] 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(1), 306-324, 2015.

[3] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman,, “Natural language processing: An introduction”, Journal of the American Medical Informatics Association, 18(5), 544-551, 2011.

[4] B. Weng, M. A. Ahmed, and F. M. Megahed, “Stock market one-day ahead movement prediction using disparate data sources”, Expert Systems with Applications, 79, 153-163, 2017.

[5] I. S. Osanga, Twitter opinion mining about leaders (Master' s thesis), Faculty of Computing, Universiti Teknologi Malaysia (Doctoral

dissertation), Faculty of Computing, Universiti Teknologi Malaysia, 2014.

[6] R. Bose, A. Das, J. Poray, and S. Bhattacharya, "Risk analysis for long-term stock market trend prediction", International Conference on Advances in Computing and Data Sciences, Springer, Singapore, pp. 381-391, April 2019.

[7] S. Kaur, and N. K. Khiva, "Online news classification using deep learning technique", International Research Journal of Engineering and Technology (IRJET), 3(10), 558-563, 2016.

[8] D. R. Kawade, and K. S. Oza, "News classification: A data mining approach", Indian Journal of Science and Technology, 9, 46, 2016.

[9] N. Sharma, and P. Kaur, "Categorize Online news Using Various Classification Techniques", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 4(2), 2015.

[10] S. Jianfeng, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, "Exploiting topic based twitter sentiment for stock prediction", Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Vol. 2: Short Papers, pp. 24-29, August 2013.

[11] G. V. Attigeri, M. M. Pai, R. M. Pai and A. Nayak, "Stock market prediction: A big data approach," TENCON 2015 - 2015 IEEE Region 10 Conference, 2015, pp. 1-5, doi: 10.1109/TENCON.2015.7373006.

[12] F. Ming, F. Wong, Z. Liu, and M. Chiang, "Stock market prediction from WSJ: text mining via sparse matrix factorization", IEEE International Conference on Data Mining, pp. 430-439, 2014.

[13] A. E. Khedr, and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis", International Journal of Intelligent Systems and Applications, 9(7), 22, 2017.

[14] Gao, Y., Zhou, L., Zhang, Y., Xing, C., Sun, Y., and Zhu, X., "Sentiment classification for stock news", In 5th International Conference on Pervasive Computing and Applications, IEEE, pp. 99-104, December 2010.

[15] R. D. Edwards, J. Magee, and W. H. C. Bassetti, Technical Analysis of Stock Trends 9th Edition: Boca Raton, 2007, CRC Press.

[16] D. G. McMillan, "Nonlinear predictability of stock market returns: Evidence from nonparametric and threshold models", International Review of Economics & Finance, 10(4), 353-368, 2001.

[17] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources", Expert Systems with Applications, 79, 153-163, 2017.

[18] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion", Journal of Machine Learning Research, 11(Dec), 3371-3408, 2010.

[19] D. Fradkin, and I. Muchnik, "Support vector machines for classification", DIMACS series in discrete mathematics and theoretical computer science, 70, 13-20, 2006.

[20] H. C. Wu, C. M. Tseng, P. C. Chan, S. F. Huang, W. W. Chu, and Y. F. Chen, "Evaluation of stock trading performance of students using a web-based virtual stock trading system", Computers & Mathematics with Applications, 64(5), 1495-1505, 2012.

[21] Z. Hu, W. Liu, J. Bian, X. Liu, and T. Y. Liu, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction", Proceedings of the eleventh ACM international conference on web search and data mining, pp. 261-269, 2018.

[22] S. Birjepatil, and K. Dalal, "Behavioural investment and artificial intelligence", Chronicle of The Neville Wadia Institute Of Management Studied and Research, 113–120, 2016.

978-967-15337©ICECoT2021