International Journal of Statistics and Probability; Vol. 9, No. 1; January 2020
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Role of Ensemble Learning in Stock Market Classification Model Accuracy Enhancement Based on Naive Bayes Classifiers

Ghaith Abdulsattar A.Jabbar Alkubaisi

Correspondence: Ghaith Abdulsattar A.Jabbar Alkubaisi, Department of Computing, Muscat College, Muscat, Sultanate of Oman. E-mail: ghaith.alkubaisi@outlook.com

Received: December 7, 2019 Accepted: December 27, 2019 Online Published: December 30, 2019

doi:10.5539/ijsp.v9n1p36 URL: https://doi.org/10.5539/ijsp.v9n1p36

Abstract

Over the last years, methods of hybrid and ensemble have attracted the attention of the data mining community. Moreover, in the computational intelligence area such as machine learning, constructing and adaptive hybrid models have become essential to achieve good performance. However, the accuracy of stock market classification models is still low, and this has negatively affected the stock market indicators. Furthermore, there are many factors that have a direct effect on the classification models' accuracies which were not addressed by previous research such as the automatic labelling technique which results in low classification accuracy due to the absence of specific lexicon, and the suitability of the classifiers to the data features and domain. In this research, a proposed model is designed to enhance the classification accuracy by the incorporation of stock market domain expert labelling technique and the construction of an ensemble Naïve Bayes classifiers to classify the stock market sentiments. The methodology for this research consists of five phases. The first phase is data collection, and the second phase is labelling, in which polarity of data is specified and negative, positive or neutral values are assigned. The third phase involves data pre-processing. The fourth phase is the classification phase in which suitable patterns of the stock market are identified by Ensemble Naïve Bayes classifiers, and the final is the performance and evaluation. The classification method has produced a significant result; it has achieved accuracy of more than 89%.

Keywords: stock market classification model, ensemble Naive Bayes classifiers, classification accuracy

1. Introduction

Accurate classification of the data sources in the stock market domain is necessary for investors to make suitable decisions, such as selling or buying stocks (Hsu, Lessmann, Sung, Ma, & Johnson, 2016; Zhong & Enke, 2017; Alkubaisi, Kamaruddin, & Husni, 2017). The current stock market classification models that utilize sentiment analysis on consumers reaction suffer from low accuracy in classification after being implemented on a dataset with different sources (Zhang, 2013; Navale et al., 2016; Arvanitis & Bassiliades, 2017). The poor accuracy of classification has a direct impact on the reliability of stock market indicators (Bollen et al., 2011; Ludwig et al., 2013; Li et al., 2016). Many factors have a direct effect on the accuracy of classification models, such as sample size, labelling technique and the classification method (Jiang et al., 2007; Sathyadevan et al., 2014). This research focuses on issues regarding consumers reaction labelling, and the classification method.

The automatic technique used in the labelling phase affects the accuracy of the classification model in the absence of a specific lexicon. Automatic labelling recognizes sentiments expressed in given consumers reaction based on existing general lexicons not specifically concerning the research domain (He & Zhou, 2011; Makrehchi, Shah, & Liao, 2013). The weakness here is related to the automatic assigning of polarity (positive, negative or neutral) that affects classification accuracy because it does not carry the real weight of the sentiment for each reaction by consumers.

Regarding the classification method, the supervised learning approach is recommended for building a classification model (Kuhn & Johnson, 2013; Ali et al., 2017). Support Vector Machine (SVM), Decision Trees, Naïve Bayes Classifiers (NBCs), K-Nearest Neighbours (KNNs) and Neural Networks (NNs) are widely used classifiers in this domain with various features (Marsland, 2015; Raschka & Mirjalili, 2017). In this research, NBCs have been selected over the other classifiers as their characteristics are more suitable for the requirements of a stock market classification model using sentiment analysis of consumers reaction, as follows. Firstly, NBCs are highly scalable, requiring a number of parameters matching to the number of variables in a learning problem (Liu et al., 2013; Shoeb & Ahmed, 2017). This research is based on tweets that reflect consumer reactions, high scalable because it includes short keywords, slang, etc. Secondly, this research focuses on fast training and fast results. This can be achieved by using maximum-likelihood

(Gong & Yu, 2010; Bollen et al., 2011; Anjaria & Guddeti, 2014). Lastly, Naive Bayes (NB) embraces four models, Gaussian Naïve Bayes (GNB), Multinomial Naïve Bayes (MNB), Bernoulli Naive Bayes (BNB), and Semi-Supervised Naïve Bayes (SSNB). In fact, the availability of different models in one model supports the validity of the proposed model's results through comparing the results of each model using the same dataset at the same time (Sarkar & Sana, 2009; Jain & Mandowara, 2016; Catal & Nangir, 2017).

2. Literature Review

2.1 Data Source

Data has turned into the cash of this time as constantly expanding in size and worth. The accessible data online is multiplying in size like clockwork (Gantz & Reinsel, 2011). In this aspect, online social networking has gotten to be a data source and key players in dispersing data to influenced entities in crisis (Smith, 2010). Currently, Twitter is the most famous micro-blog device among other existing reciprocals and has been included broadly in people in general media; for instance, it has been utilized by business communications, product information, political campaign, and news organizations. It is the only micro-blogging service that has turned into the main quickest developing patterns on the Internet, with an exponentially- increasing users base exceeding 190 million users in July 2010. Twitter is a widespread microblogging facility where followers and users make status messages (called "tweets") (Arceneaux & Schmitz Weiss, 2010). These tweets here and there express sentiments about different themes. According to Go, Bhayani, and Huang (2009) and Rout et al. (2018) classified Twitter Streaming Application Programming Interface (API) and Twitter Representational State Transfer (REST) API as two sorts of API utilized to assemble tweets. Twitter's data is displayed using APIs (Shi, & Brigadir, 2014; Trupthi, Pabboju, & Narasimha, 2017).

2.2 Labelling Techniques

In the domain of classification models, supervised machine learning classifiers need to label the dataset to assist the learning algorithm in the classification process. The first way to label the collected tweets is done by manual labelling (Zhang, 2013), but most of the manual labelled dataset has done without referring to the experts in the research domain such as the stock market (Makrehchi et al., 2013). At the same time, most of the researchers are looking to the labelled dataset has labelled before without looking to the way that has utilized to label the dataset because at the end they need to a big labelled dataset. The most popular online source for the labelled dataset is kaggle.com. This website contains millions of labelled tweets by different users with different knowledge. Unfortunately, when we are looking for the specialization in labelling for the classification model purpose, there are very little dataset has been labelled by the experts in the research domain (Bollen et al., 2011).

2.3 Data Pre-processing

Data pre-processing can be described as a motivated theory using for computational techniques representing the human language and automatic analysis (Nadkarni, Ohno-Machado, & Chapman, 2011; Hardeniya, Perkins, Chopra, Joshi, & Mathur, 2016). This indicates that data pre-processing is a form of technology resolving most of the ubiquitous products: human dialect, as it shows up in messages, website links, tweets, social media, daily paper stories, and scientific articles, in a great many dialects and assortments. In the previous decade, effective normal dialect handling applications have turned out because is the most regular of our experience. For example, machine interpretation can be processed through the linguistic use of revision and spelling on the web, through the electronic mail programmed from spam and also from distinguishing individuals' conclusions about services or products to extricating arrangements from email.

2.4 Classification as a Supervised Machine Learning

Classification is used as a supervised data mining method that includes assigning a label to an arrangement of unlabelled information objects (Dougherty, 2012; Conneau et al., 2017). It is a data mining function whose aim is to allocate items to target classes (Bird, Klein, & Loper, 2009). In general, the aim of classification is to precisely predict the target class for each case in the data. For example, a classification model could be used to identify stock market behaviour as high, low or stable.

In this research, the selected Machine Learning (ML) classifier is NB. NBCs are a family of simple "probabilistic classifiers "based on applying Bayes' theorem with strong (Naive) independence assumptions between the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including SVM (Rennie et al., 2003). NBCs are an intuitive method that uses the probabilities of each proposed feature belonging to each class to make a classification (Mahajan Shubhrata et al., 2016; Shmueli et al., 2017). NBCs are supervised ML classifiers which represent models of simple probabilistic classifiers based on applying Bayes theorem with the hypothesis of independence between features (Wang et al., 2010).

In Computer Science, NBCs are defined as a family of algorithms apply Bayes theorem and adopt that each feature value is independent of other features value which is mean no relation between features values (Vinodhini &

Chandrasekaran, 2012). There are three occasion models for NB beside the baseline model; Multinomial MNB (Witten et al., 2016), Multivariate BNB (Raschka, 2014), and SSNB (Bhattu & Somayajulu, 2012). With MNB event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a multinomial (p1, …, p n) where p I is the probability that event i occurs (or K such multinomial in the multiclass case). A feature vector x = (x1, …, x n) is then a histogram, with x i counting the number of times event i was observed in a particular instance (Tan & Zhang, 2008; Witten et al., 2016). Di Nunzio and Sordoni (2012) explained that the multivariate BNB, features are independent Booleans (binary variables) describing inputs. This model is popular for document classification tasks, where binary term occurrence features are used rather than term frequencies. If x i is a Boolean expressing the occurrence or absence of the (i'th) term from the vocabulary, then the likelihood of a document given a class C k. This event model is especially popular for classifying short texts. It has the benefit of explicitly modelling the absence of terms. According to Bhattu and Somayajulu (2012), and Zhou (2012) the SSNB gave a way to train a naive Bayes classifier from labelled data, it's possible to construct a semi-supervised training algorithm that can learn from a combination of labelled and unlabeled data by running the supervised learning algorithm in a loop. This training algorithm is an instance of the more general Expectation Maximization algorithm (EM): the prediction step inside the loop is the E-step of EM, while the re-training of naive Bayes is the M-step. The algorithm is formally justified by the assumption that the data are generated by a mixture model, and the components of this mixture model are exactly the classes of the classification problem. In this research, MNB and BNB have been utilized to build the required classification model.

2.5 Ensemble Learning

According to Kazienko, Lughofer, and Trawiński (2013), hybridization has become essential for the data science's research due to that the hybridization can help the proposed model to achieve best results most of the times. This reasonable or high performance for the proposed learning model based on the fact behind the hybridization that is leading to combine the characteristics for different methods in one model at the same time.

As well, Castillo, Melin, and Pedrycz (2007), and Melin, Castillo, Ram ́ırez, and Pedrycz (2007) illustrated that the hybrid or ensemble classifiers are a model for combining similar or different ML classifiers for a classification mission through majority or plurality voting. The generated model based on combining similar or different ML classifiers implements hard computing (voting) or soft computing (voting). In hard voting, the constructed model classifies the final label that has been classified most frequently by the constructed classification model. On another hand, in the soft voting, the constructed model classifies the class label via averaging the class probabilities. The hard voting has a simple process when it compared with the soft voting because it is classifying based on the majority. For example, when X model has 3 classifiers (c1, c2, and c3), c1 and c2 classify a random tweet as a positive tweet, but c3 classify the same tweet as a negative, so the hard voting will take the label here based on the majority that means the label will be (positive). The soft voting uses the weight of each label besides the probabilities for the labels then take the average based on each classifier result (label). Figure 1 shows the structure of ensemble ML models.

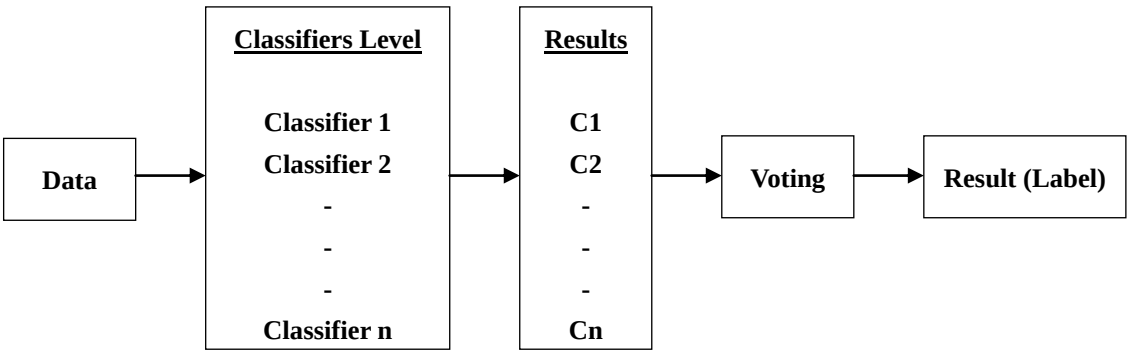

Figure 1. Structure of Ensemble ML Models

In Computer Science, there are many facts encourage the researchers in the domain of stock market classification model. One of these facts that multiple probabilistic classifiers could achieve high classification accuracy (ensemble learning-stacking), and NB consists of four classifiers (GNB, MNB, BNB, and SSNB). In conclusion, the main advantages for an ensemble ML model are a low error, less overfitting, and most of the times come with good results (Ara ́ujo & New, 2007; Hung & Chen, 2009; Chen & Chen, 2013).

2.6 Related Works

This research focuses on the classification models that utilized Arabic tweets. So, this section reviews the literature related to the stock market classification models that utilized Arabic tweets, ML and sentiment analysis.

Hamed et al.'s (2015) study is one of the few to be conducted in Arabic sentiment analysis. It demonstrates the importance of the pre-processing step as a key factor in achieving a high level of accuracy of sentiment analysis. The research was introduced for Saudi stock market tweets. It aimed to illustrate the relationship between Saudi tweets and the Saudi market index, using different implementations of SVM, KNNs and NB algorithms. The step of weighting comes after applying the most relevant of these data mining approaches in order to build a classification model based on the sentiment polarity of the tweet. These authors have developed a small desktop application to collect the corpus of Arabic tweet data from Twitter; it was created under the C# environment supported by the official developer's APIs of Twitter. The main function of the application is to collect, label and save the related tweets, as well as removing those tweets that are not relevant. They considered their Twitter micro-blog as a platform for trading opinion mining in the Saudi stock market.

Hamed et al. (2016) also proposed a sentiment analysis model for the SA stock market using sentiment analysis on Arabic tweets. It classifies the tweets into positive, negative or neutral, focusing on the role of the neutral class. The model was built based on the hybridization of ML classifiers such as SVM and NB, and the data pre-processing. It aims to show and define the role of the polarity class (label) by classifying the collected tweets in Arabic into three polarities, positive, negative, and neutral. Again, there are four stages: data collection, pre-processing, classification, and model evaluation.

The last model in the domain of Arabic tweets was proposed by AL-Rubaiee, Qiu, Alomar, and Li (2018); they aimed to enhance the labelling process, which has a direct impact on the reliability of the classification. They proposed improvements to the expert knowledge via two approaches: first, by defining neutral to include tweets with both positive and negative polarity; and second by relabelling. 2000 tweets were collected using Twitter API (all in Arabic), and they classified the dataset using SVM. The classification phase aims to show the differences in accuracy when using the original labelling process and the improved labelling process.

In summary, the various classifiers used by researchers to perform sentiment analysis for stock market classification are SVM (Go et al., 2009; Kolchyna, Souza, Treleaven, & Aste, 2015), KNNs (Barbosa analysis & Feng, 2010), Expectation Maximization (EM) (Yengi, Karayel, & Omurca, 2015) and NB (Liu, Blasch, Chen, Shen, & Chen, 2013; Narayanan, Arora, & Bhatia, 2013; Chandrasekar & Qian, 2016). The review indicates that the process of selecting the ML classifier or classifiers was based on the availability of the classifier rather than the research and model requirements. Table 1 summarizes the facts that related to previous studies.

Table 1. The Facts Sheet

Research Title, Authors, & Publication Year	Source & Size of Dataset	ML Classifier (s)	Accuracy
1. Analysis of the Relationship Between Saudi Twitter Posts and the Saudi Stock Market. Hamed AL-Rubaiee, Renxi Qiu, and Dayou Li, (2015).	Twitter, 3335 tweets	SVM, KNNs and NB	95.71%, 95.91%, and 56.28%
2. The Importance of Neutral Class in Sentiment Analysis of Arabic Tweet. Hamed AL-Rubaiee, Renxi Qiu, and Dayou Li, (2016).	Twitter, 2051 tweets	NB and SVM	76.86% and 84.97%
3. Techniques for Improving the Labelling Process of Sentiment Analysis in the Saudi Stock Market. Hamed AL-Rubaiee, Renxi Qiu, Khalid Alomar, Dayou Li (2018).	Twitter, 2000 tweets	SVM	84.92%

3. The Proposed Method and Conceptual Framework

This section illustrates the proposed framework for this research as shown in Figure 2. The conceptual frameworks consists of five phases as follows. Data collection, expert labelling, pre-processing, classification using ensemble

learning, and performance evaluation. The pre-processing prepares the data and extracts the necessary features before classification; sentiment analysis assigns the appropriate polarity for each tweet, based on the knowledge, and classification is a data mining function that assigns clauses in a collection to objective categories or classes (Bollen et al., 2011; Meesad & Li, 2014).

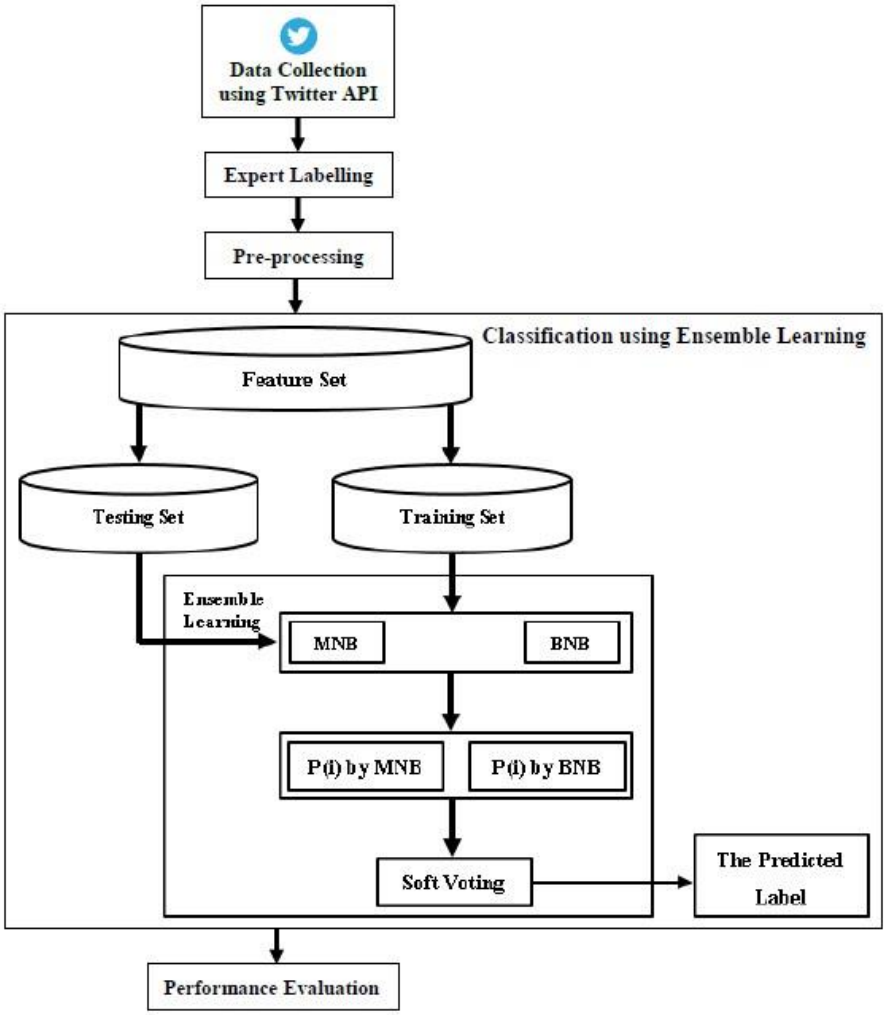

Figure 2. The Proposed Conceptual Framework for Classification using Ensemble Learning

3.1 Data Collection

The scope of this research is classification to infer tweets' polarity by implementing ensemble ML classifiers. The goal of classification is to accurately predict the target class for each clause in the dataset (Tan, Steinbach, & Kumar, 2006; Kesavaraj & Sukumaran, 2013). This research concentrates on consumer reactions and stock market behaviour for the Almarai Company and AlSafi Arabia (ASA), investigating these two companies' tweets and stock separately. Almarai and ASA are represented on the SA stock market. Tweets from all two companies have been collected from their official sites on Twitter. The size of Twitter followers for each company is another important reason for selecting these two companies. Almarai (@almarai) has more than 447,000 followers on Twitter, and ASA (@alsafiarabia) more than 40,000.

In Twitter, there are two types of APIs used to gather tweets, Twitter REST API and Twitter Streaming API (Go et al., 2009). This research utilized an internal library using REST API. Thus, it applied Open Authorization (OAuth) as required by Twitter4j library. OAuth is used in Twitter4j library Twitter and it supports access and provides authorized access to its API (Tugores and Colet, 2014). APIs streaming can provide a continuous flow of information with updates. Streaming API can access to real-time data of tweets using queries. This research adopts the REST API since it enables users to retrieve recently posted tweets by specific queries using HTTP methods (Makice, 2009). Moreover, it can filter results based on time, region, and language (Li et al., 2012). The return of queries is a list of JavaScript Object Notation

(JSON) objects containing tweets and metadata. These objects involve username, location, time, re-tweets, and the dataset saves in Excel sheet with CSV file (Aramaki et al., 2011).

3.2 Expert Labelling

In this research, this phase has done by experts in the domain of stock market based on the stock market concepts and consumer reactions analysis to give each tweet its appropriate sentiment weight such as positive, negative, and neutral. The real sentiment weight is related to the impact of a tweet on stock behaviour such as positive or negative affection. Practically, each tweet was labelled as positive: 1, negative: 2 and neutral: 0.

3.3 Pre-processing

The quality of collected data can directly tell the researcher how is the model performance, so before the classification phase, the collected tweets must be well preprocessed because of pre-processing phase is very necessary in case of classification accuracy enhancement (Hussien, Tashtoush, Al-Ayyoub, & Al-Kabi, 2016; Alkubaisi, Kamaruddin, & Husni, 2018a). There are two main steps, firstly is the transformation that consists of data cleaning, and secondly is the filtration which consists of features extraction, compute Term Frequency-Inverse Document Frequency (TF-IDF) and create TF-IDF vectorizer.

3.4 Classification Using Ensemble Learning Based on NBCs

The proposed classifiers are based on combining the MNB and BNB to be one classifier by using the soft ensemble voting as shown in Figure 2. The proposed classification method as follows:

a. Hybridizing MNB and BNB to be one classifier by using the soft voting ensemble that represents the best way to hybridize different classifiers to be one classifier in one model. The proposed hybridization based on synchronizing the implementation for both classifiers in one model.

b. Determining the result that has a maximum probability for both classifiers by using equation 1.

$$result = argmax_i(P(i)) \tag{1}$$

Determining the result that has a maximum probability, when I = #class

c. Implementing the Soft Voting by computing the average probabilities, equation 2 represents the average probabilities as follows:

$$P\left(\frac{i_n}{x}\right) \text{ when } x \text{ is the size of the hybridized classifiers. } P\left(\frac{i_n}{x}\right) = argmax_{\ i} \sum_{j=1}^{m} w_j \ P_{\ ij} \tag{2}$$

d. Selecting the high predicted class relies on the high average for the class probabilities in step 3 that calculated by using equation 2.

For example, suppose that three different classifiers there are C1, C2, and C3. Each classifier has weighted, and the probabilities as shown in Table 2 based on three kinds of classes.

Table 2. Example about the Probabilities Averaging in Soft Voting Ensemble

Classifier	Class A	Class B	Class C
Classifier 1	0.2	0.5	0.3
Classifier 2	0.6	0.3	0.1
Classifier 3	0.3	0.4	0.3
Average	≈ 0.37	0.4	0.7

The averaging for the probabilities of class A, B, and C has calculated using the averaging formula when X represents the number of the predicted classes:

$$P\left(\frac{i_n}{X}\right) = \frac{P(i_1)+P(i_2)\ldots\ldots P(i_n)}{X} \tag{3}$$

Class A: 0.2 + 0.6 + 0.3 / 3 = 0.36666666 ≈ 0.37

Class B: 0.5 + 0.3 + 0.4 = 0.4

Class C: 0.3 + 0.3 + 0.1 = 0.7

Based on the results, class C has achieved the highest average (B>A, C>A, and C>B), so the predicted class for the tested document is C.

3.5 Performance Evaluation

The purpose of the empirical evaluation is how to evaluate the performance for the proposed stock market classification

model by using the performance measurements such as classification accuracy, precision, recall, f-measure, and support (Alkubaisi, Kamaruddin, & Husni, 2018b). As well, to note the impact of the proposed methods, techniques, and features on the classification model's enhancement. This research focuses on the classification accuracy besides the precision, recall, f-measure, and support as shown in Table 3.

Table 3. Equations used for Evaluation the Classification Model

Measurements	Equations
1. Recall (Classification Completeness) = Sensitivity. They represent the total Positive Rate is a proportion of cases that were correctly identified as positive.	It is defined as: [TP / (TP + FN)] = [d / (c + d)].
2. Precision (Classification Exactness).	Its equation given as: [TP / (TP+ FP)].
3. F1-Measure is a measure that combines precision and recall (A weighted average of precision and recall).	Its equation as follows: 2 [Precision. Recall / Precision + Recall].
4. Support is the number of occurrences of each class.	Its equation is given as: Count (documents), Documents here = Tweets.
5. Accuracy is defined as the portion or part of the total number of classifications that is correct.	It is given as: [(a + d) / (a + b + c + d)] or [(TP + TN) / (TP + FP + FN + TN)].

3. Classification Results

This section presents the results of the proposed method implementation using two different datasets: Almarai and ASA Arabic tweets. This research executes Tweepy method to collect targeted tweets from specific pages as explained in section 3.1. Before implementation, the Tweepy method has modified to be compatible with the research requirements by determining the size of tweet besides the way for inserting the name of the Twitter page. These were done by using Python programming language and its libraries. Python is a very useful language due to the active developer community creates many libraries which extend the language and make it easier to be used for various services specifically data mining research domain. One of these libraries is Tweepy, an open-sourced which is hosted at (GitHub.com) that enables Python to communicate with the Twitter platform and uses its API. Table 4 shows a sample of Almarai Arabic tweets.

Table 4. Sample of Almarai Arabic Tweets

Original Tweets (In Arabic)	Tweets (After English Translation)
Tweet	
@n05450نعتز بثقتكم في منتجاتنا، ونلتزم بالجودة التي تستحق هذه الثقة	@n05450 we cherish your trust in our products and commit ourselves to the quality that deserves this trust
@mohammed205205يمكنك التواصل مع إدارة التوظيف بالمقر الرئيسي على تليفون 0114700005 لمعرفة تفاصيل أكثر عن التوظيف	@mohammed205205 you can contact the Recruitment Department at Headquarters at 0114700005 for more details on employment
@abomajed0500نأسف لا يمكننا الرد على الملاحظة دون استلامنا للعينة وتقديمها للفحص الفني	@abomajed0500 sorry, we cannot reply now. We should receive the sample first, then only it can be submitted to the technical laboratory.
صباح الخير	Good Morning.
في الوقت الحالي توزيع منتجات المراعي في دول مجلس التعاون الخليجي والأردن ومصر	At present, Almarai products are distributed in GCC, Jordan, and Egypt.
@boyonaizah6كما هو موضح على العبوة تاريخ الإنتاج 2016/12/8	@boyonaizah6 production date as shown in packaging

Table 4 shows a sample of tweets from Almarai page on Twitter (@almarai). Moreover, it shows the original tweets in Arabic besides the tweets after translation to English.

The performance and evaluation of the model has five measurements: precision, recall, F-measure, support, and accuracy, which in turn all contributing to focusing on improved classification accuracy. the proposed method was implemented with the following parameter values set: optimized = 1, fraction = 1, n-gram = 3, n-iter = 100, and alpha-val = 0.01. Eventually, all tweets were labelled by experts before the classification. The most important measurement for the implemented method is accuracy because it reflects the extent of the improvements in the proposed over other methods using the same classifiers (NBCs) and same kind of dataset. Measurement of accuracy is especially important because it reflects the percentage of correct pattern recognition and assigned polarity. Tables 5 and 6 show the implementation results for the ensemble MNB and BNB using Almarai and ASA Arabic labelled tweets. Almarai tweets were collected from 18 September 2016 to 25 May 2017, and the number of tweets for this test is 3,214. 3,168 ASA tweets were collected from 11 June 2015 to 4 January 2019.

Table 5. Ensemble MNB and BNB using Almarai Arabic Tweets (all classes: 1, 2, and 0)

Class	Precision	Recall	F-Measure	Support
0	0.82	0.93	0.88	1053
1	0.95	0.88	0.91	1942
2	0.82	0.87	0.84	219
Total	0.90	0.89	0.90	3214

- Classification Accuracy = 89.45%.
- Neutral Polarity Ratio = (1053 / 3214) = 32.77%.
- Positive Polarity Ratio = (1942 / 3214) = 60.42%.
- Negative Polarity Ratio = (219 / 3214) = 6.81%.

Table 6. Ensemble MNB and BNB using ASA Arabic Tweets (all classes: 1, 2, and 0)

Class	Precision	Recall	F-Measure	Support
0	0.81	0.99	0.89	949
1	1.00	0.87	0.93	2120
2	0.62	0.94	0.75	99
Total	0.93	0.91	0.91	3168

- Classification Accuracy = 90.8%.
- Neutral Polarity Ratio = (949 / 3168) = 29.95%.
- Positive Polarity Ratio = (2120 / 3168) =66.92 %.
- Negative Polarity Ratio = (99 / 3168) = 3.13%.

Tables 5 and 6 show that method implementation achieved a high level of accuracy, with most of the tweets carrying a positive polarity. In addition to the main performance measurements, the polarity ratios show the size of each polarity inside the classified dataset, which reflects the strength of the sentiment based on the classification and leading to construction of a stock behaviour indicator.

4. Discussion and Benchmark

The ensemble learning method performance measurements such as accuracy, precision, recall and F-score were high, reflecting the reliability of the classification. High precision and recall mean the learning algorithm generates more relevant relationships between features and labels (positive, negative, or neutral). These relationships lead to high classification accuracy and reliability at the same time. The combination was based on supervised learning using MNB and BNB. The performance and evaluation results reflect the importance of selecting appropriate classifiers to meet the research requirements.

In this research, a comparison was made between the proposed method and the baseline NBCs to see how ensemble and expert labelling improve the stock market classification accuracy. The same dataset of tweets is used, and the classifiers are Naive Bayes Classifiers. Table 7 shows the classification accuracy for the baseline NBCs and the ensemble learning method using Almarai Arabic tweets.

Table 7. The Ensemble Learning Classification Accuracy vs NBCs using Almarai Arabic tweets

ML Classifier	Accuracy using Almarai Arabic Tweets
Ensemble Learning Method (MNB&BNB)	89.45%
NB	87.9%
MNB	75.97%
BNB	75.42%

The table shows that the classification accuracy for the ensemble learning method using MNB and BNB is higher than for the baseline models despite using the same dataset.

5. Conclusion

The main goal for this research is to enhance the classification accuracy for a stock market classification model using sentiment analysis on Twitter by building an ensemble learning classifier based on NBCs, specifically MNB and BNB. The proposed conceptual framework has mainly been founded on tweets collection, sentiment analysis approach, employing expert labelling technique, tweets preprocessing, ensemble classification, and performance evaluation. This research constructed the proposed ensemble learning method using MNB and BNB as a machine learning classifier for stock market classification. Eventually, the performance and evaluation results show that the proposed classification method has achieved high classification accuracy using the different dataset in size and similar in language.

References

Ali, R., Lee, S., & Chung, T. C. (2017). Accurate multi-criteria decision making methodology for recommending machine learning algorithm. *Expert Systems with Applications, 71,* 257-278. https://doi.org/10.1016/j.eswa.2016.11.034

Alkubaisi, G. A. A., Kamaruddin, S. S., & Husni, H. (2017). A Systematic Review on The Relationship Between Stock Market Prediction Model Using Sentiment Analysis on Twitter Based on Machine Learning Method and Features Selection. *Journal of Theoretical and Applied Information Technology, 95*(24), 6924-6933.

Alkubaisi, G. A. A., Kamaruddin, S. S., & Husni, H. (2018). Conceptual Framework for Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers. *International Journal of Engineering & Technology, 7*(2.14), 57-61. https://doi.org/10.14419/ijet.v7i2.14.11156

AL-Rubaiee, H., Qiu, R., Alomar, K., & Li, D. (2018). Techniques for Improving the Labelling Process of Sentiment Analysis in the Saudi Stock Market. *International Journal of Advanced Computer Science and Applications, 9*(3), 34-43. https://doi.org/10.14569/IJACSA.2018.090307

Anjaria, M., & Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. Proceedings of the 6th IEEE International Conference on Communication Systems and Networks (COMSNETS), (pp. 1-8). https://doi.org/10.1109/COMSNETS.2014.6734907

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in ecology & evolution, 22*(1), 42-47. https://doi.org/10.1016/j.tree.2006.09.010

Arceneaux, N., & Schmitz, W. A. (2010). Seems stupid until you try it: Press coverage of Twitter, 2006-9. *New Media & Society, 12*(8), 1262-1279. https://doi.org/10.1177/1461444809360773

Arvanitis, K., & Bassiliades, N. (2017). *Real-Time Investors' Sentiment Analysis from Newspaper Articles.* In Advances in Combining Intelligent Methods, (pp. 1-23): Springer, Cham. https://doi.org/10.1007/978-3-319-46200-4_1

Barbosa, L., & Feng, J. (2010). *Robust sentiment detection on twitter from biased and noisy data.* Proceedings of the 23rd International Conference on Computational Linguistics: Posters, (pp. 36-44). Association for Computational Linguistics

Bhattu, N., & Somayajulu, D. (2012). *Semi-supervised Learning of Naive Bayes Classifier with feature constraints.* Proceedings of the 24th International Conference on Computational Linguistics, (pp. 65-78).

Bhattu, N., & Somayajulu, D. (2012). *Semi-supervised Learning of Naive Bayes Classifier with feature constraints.* Proceedings of the 24th International Conference on Computational Linguistics, (pp. 65-78).

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit: "O'Reilly Media, Inc.".

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

Castillo, O., Melin, P., & Pedrycz, W. (Eds.). (2007). *Hybrid Intelligent Systems: Analysis and Design* (Vol. 208). Springer. https://doi.org/10.1007/978-3-540-37421-3

Catal, C., & Nangir, M. (2017). A sentiment classification model based on multiple classifiers. *Applied Soft Computing, 50,* 135-141. https://doi.org/10.1016/j.asoc.2016.11.022

Chandrasekar, P., & Qian, K. (2016, June). The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 618-619). IEEE. https://doi.org/10.1109/COMPSAC.2016.205

Chen, S.-H., & Chen, M.-C. (2013). Addressing the advantages of using ensemble probabilistic models in estimation of distribution algorithms for scheduling problems. *International Journal of Production Economics, 141*(1), 24-33. https://doi.org/10.1016/j.ijpe.2012.05.010

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*. https://doi.org/10.18653/v1/D17-1070

Di Nunzio, G. M., & Sordoni, A. (2012). How well do we know Bernoulli? *IIR, 835,* 38-44.

Dougherty, G. (2012). Pattern recognition and classification: an introduction: Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-5323-9_1

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iview, 1142*, 1-12.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, *Stanford, 1*, 12.

Gong, Z., & Yu, T. (2010). Chinese Web Text Classification System Model Based on Naive Bayes. In the IEEE International Conference on E-Product E-Service and E-Entertainment (ICEEE), (pp. 1-4). https://doi.org/10.1109/ICEEE.2010.5660869

Hamed, A.-R., Qiu, R., & Li, D. (2015). *Analysis of the relationship between Saudi twitter posts and the Saudi stock market*. In the IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), (pp. 660-665).

Hamed, A.-R., Qiu, R., & Li, D. (2016). The importance of neutral class in sentiment analysis of Arabic tweets. Int. J. Comput. *Sci. Inform. Technol, 8,* 17-31. https://doi.org/10.5121/ijcsit.2016.8202

Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). Natural Language Processing: Python and NLTK. Packt Publishing Ltd.

He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing & Management, 47*(4), 606-616. https://doi.org/10.1016/j.ipm.2010.11.003

Hsu, M.-W., Lessmann, S., Sung, M.-C., Ma, T., & Johnson, J. E. (2016). Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications, 61*, 215-234. https://doi.org/10.1016/j.eswa.2016.05.033

https://doi.org/10.1109/BigData.2013.6691740

Hung, C., & Chen, J.-H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert Systems with Applications, 36*(3), 5297-5303. https://doi.org/10.1016/j.eswa.2008.06.068

Hussien, W., Tashtoush, Y. M., Al-Ayyoub, M., & Al-Kabi, M. N. (2016). Are emoticons good enough to train emotion classifiers of Arabic tweets?. *CSIT. IEEE*, 1-6. https://doi.org/10.1109/CSIT.2016.7549459

Ifrim, G., Shi, B., & Brigadir, I. (2014). Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In the Second Workshop on Social News on the Web (SNOW), (pp. 33-40). Seoul, Korea.

Jain, A., & Mandowara, J. (2016). Text classification by combining text classifiers to improve the efficiency of classification. International Journal of Computer Application (2250-1797), 6(2).

Jiang, L., Wang, D., Cai, Z., & Yan, X. (2007). Survey of improving naive Bayes for classification. In the International Conference on Advanced Data Mining and Applications, (pp. 134-145). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73871-8_14

Kazienko, P., Lughofer, E., & Trawiński, B. (2013). Hybrid and ensemble methods in machine learning J. UCS special issue. *J Univers Comput Sci, 19*(4), 457-461.

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. In the Fourth IEEE

International Conference on Computing, Communications and Networking Technologies (ICCCNT), (pp. 1-7). https://doi.org/10.1109/ICCCNT.2013.6726842

Kolchyna, O., Souza, T. T., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling, (Vol. 26). New York. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C.-C. (2012). Tedas: A twitter-based event detection and analysis system. In *the IEEE 28th International Conference on Data Engineering* (pp. 1273-1276). https://doi.org/10.1109/ICDE.2012.125

Li, X., Xie, H., Wang, R., Cai, Y., Cao, J., Wang, F., ... Deng, X. (2016). Empirical analysis: stock market prediction via extreme learning machine. *Neural Computing and Applications, 27*(1), 67-78. https://doi.org/10.1007/s00521-014-1550-z

Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013). Scalable sentiment classification for big data analysis using Naive Bayes Classifier. In *the IEEE International Conference on Big Data* (pp. 99-104). https://doi.org/10.1109/BigData.2013.6691740

Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013, October). Scalable sentiment classification for big data analysis using naive bayes classifier. In *2013 IEEE international conference on big data* (pp. 99-104). IEEE.

Ludwig, S., De Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing, 77*(1), 87-103. https://doi.org/10.1509/jm.11.0560

Mahajan, S. D., Deshmukh, K. V., Thite, P. R., Samel, B. Y., & Chate, P. (2016). Stock Market Prediction and Analysis Using Naïve Bayes. *International Journal on Recent and Innovation Trends in Computing and Communication, 4*(11), 121-124.

Makice, K. (2009). Twitter API: Up and running: Learn how to build applications with the Twitter API: "O'Reilly Media, Inc.".

Makrehchi, M., Shah, S., & Liao, W. (2013, November). Stock prediction using event-based sentiment analysis. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 01* (pp. 337-342). IEEE Computer Society.

Marsland, S. (2015). *Machine learning: an algorithmic perspective* (2nd ed.). CRC press. https://doi.org/10.1201/b17476

Meesad, P., & Li, J. (2014). Stock trend prediction relying on text mining and sentiment analysis with tweets. In *the 4th IEEE World Congress on Information and Communication Technologies (WICT)* (pp. 257-262). https://doi.org/10.1109/WICT.2014.7077275

Melin, P., Castillo, O., Ramírez, E. G., & Pedrycz, W. (Eds.). (2007). *Analysis and design of intelligent systems using soft computing techniques* (Vol. 41). Springer Science & Business Media. https://doi.org/10.1007/978-3-540-72432-2

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association, 18*(5), 544-551. https://doi.org/10.1136/amiajnl-2011-000464

Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. In *the International Conference on Intelligent Data Engineering and Automated Learning* (pp. 194-201). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41278-3_24

Navale, G., Dudhwala, N., Jadhav, K., Gabda, P., & Vihangam, B. K. (2016). Prediction of Stock Market Using Data Mining and Artificial Intelligence. *International Journal of Engineering Science, 6539.*

Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. arXiv preprint arXiv:1410.5329.

Raschka, S., & Mirjalili, V. (2017). Python Machine Learning: Packt Publishing Ltd.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the 20th international conference on machine learning (ICML-03), (pp. 616-623).

Rout, J. K., Choo, K. K. R., Dash, A. K., Bakshi, S., Jena, S. K., & Williams, K. L. (2018). A model for sentiment and emotion analysis of unstructured social media text. *Electronic Commerce Research, 18*(1), 181-199. https://doi.org/10.1007/s10660-017-9257-8

Sarkar, B. K., & Sana, S. S. (2009). A hybrid approach to design efficient learning classifiers. *Computers & Mathematics with Applications, 58*(1), 65-73. https://doi.org/10.1016/j.camwa.2009.01.038

Sathyadevan, S., Sarath, P. R., Athira, U., & Anjana, V. (2014). Improved document classification through enhanced Naive Bayes algorithm. In *the IEEE International Conference on Data Science & Engineering (ICDSE)*, (pp. 100-104). https://doi.org/10.1109/ICDSE.2014.6974619

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data mining for business analytics: concepts, techniques, and applications in R*. John Wiley & Sons.

Shoeb, M., & Ahmed, J. (2017). Sentiment Analysis and Classification of Tweets Using Data Mining. *Work, 4*(12).

Smith, A. (2010). *Government online: The internet gives citizens new paths to government services and information.* Pew Internet & American Life Project.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). Classification: basic concepts, decision trees, and model evaluation. *Introduction to data mining, 1,* 145-205.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications, 34*(4), 2622-2629. https://doi.org/10.1016/j.eswa.2007.05.028

Trupthi, M., Pabboju, S., & Narasimha, G. (2017). Sentiment analysis on twitter using streaming API. In *the 7th IEEE International Advance Computing Conference (IACC)* (pp. 915-919). https://doi.org/10.1109/IACC.2017.0186

Tugores, A., & Colet, P. (2014). Mining online social networks with Python to study urban mobility. *arXiv preprint arXiv:1404.6966*.

Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment analysis and opinion mining: a survey. *International Journal, 2*(6).

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gong, Y. (2010). Locality-constrained linear coding for image classification. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3360-3367). https://doi.org/10.1109/CVPR.2010.5540018

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yengi, Y. K., Karayel, M., & Omurca, S. İ. (2015). An Alternative Method for Sentiment Classification with Expectation Maximization and Priority Aging. *International Journal of Scientific Research in Information Systems and Engineering (IJSRISE), 1*(2), 91-96.

Zhang, L. (2013). *Sentiment analysis on Twitter with stock price and significant keyword correlation (Doctoral dissertation)*. Available from the University of Texas at Austin, Department of Computer Science.

Zhong, X., & Enke, D. (2017). Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications, 67*, 126-139. https://doi.org/10.1016/j.eswa.2016.09.027

Zhou, Z. H. (2012). *Ensemble methods: foundations and algorithms.* Chapman and Hall/CRC. https://doi.org/10.1201/b12207