Hybrid Classifier by Integrating Sentiment and Technical Indicator Classifiers

Nguyen Duc Van[1]([⊠]), Nguyen Ngoc Doanh[2,4],
Nguyen Trong Khanh[3,4], and Nguyen Thi Ngoc Anh[1,4] [iD]

[1] Hanoi University of Science and Technology,
No. 1, Dai Co Viet, Hanoi, Vietnam
`vanndkstnk57@gmail.com`, `anh.nguyenthingoc@hust.edu.vn`
[2] ThuyLoi University, No. 175 Tay Son, Dong Da, Hanoi, Vietnam
[3] Post and Telecommunications Institute Technology,
Ho Chi Minh City, Vietnam
[4] IRD, Sorbonne Universités, UPMC Univ Paris 06 Unité Mixte Internationale de,
Modélisation Mathématique et Informatiques des Systèmes Complexes
(UMMISCO), 32 Avenue Henri Varagnat, 93143 Bondy Cedex, France

Abstract. Classifiers in stock market are an interesting and challenging research topic in machine learning. A large research has been conducted for classifying in stock market by using different approaches in machine learning. This research paper presents a detail study on integrating sentiment classifier and technical indicator classifier. The research subject is investigated to classify a stock into one of three labels being top, neutral or bottom. First, using technical indicators such as relative strength index (RSI), money flow index (MFI) and relative volatility index (RVI) to classify stock, then using bagging of learning machine to classify the stock. Second, using sentiment data to classify the stock. Third, integrating technical indicator and sentiment classifiers to build hybrid classifier. In this study, hybrid machine learning by combining sentiment and technical indicator classifiers is proposed. We applied this proposal hybrid classifier for five stocks in VN30. The empirical results show hybrid classifier stock has more power than single technical indicator classifier or sentiment classifier.

Keywords: Machine learning · Stock market · Classifier · Sentiment analysis
Hybrid classifier · Technical indicator

1 Introduction

Recently, more and more researchers concentrate on analysing sentiment factors of stock market. This paper tests whether hybrid classifier integrating sentiment factors and technical indicator more power than single classifier.

In the fact that, hybrid machine learning has been studied by some other researches such as Gao and Yang [1]. They integrated sentiment factors and price volume factors. Gao and Yang [1] show that mixed-frequency stock index combining sentiment factors and price volume factors have positively predictive power statistically. Moreover,

mixed-frequency stock index futures sentiment and mixed-frequency stock index sentiment has greater positively predictive power in high sentiment period [1].

Moreover, marching learning is applied in stock market by Ballings et al. [2]. They studied the benchmark ensemble methods (Random Forest, AdaBoost and Kernel Factory) against single classifier models (Neural Networks, Logistic Regression, Support Vector Machines and K-Nearest Neighbor) [2]. They gathered data from 5767 publicly listed European companies and used the area under the receiver operating characteristic curve as a performance measure. The results indicate that Random Forest is the top algorithm followed by Support Vector Machines, Kernel Factory, AdaBoost, Neural Networks, K-Nearest Neighbors and Logistic Regression [2].

Sentiment data is important information related to news can be good, bad or neutral. The statistical analysis of relatively simple sentiment cues can provide a surprisingly meaningful sense for investors. Thus, integrating sentiment analysis and technical indicator in classifying stock is investigated in this paper. We used classifier that is called ensemble. First, we used technical indicator to classify a stock into one of three labels top, neutral and bottom. Then, sentiment data was used to classify by method that Lagarde and Arnaud was presented [4]. Last, a new classifier method to classify stock was proposed by integrating technical indicator classifier and sentiment classifier.

The research problem of this paper is applied the integrating technical indicator and sentiment classifier to label any stock. This proposal classifier predict a stock should buy or sell in the future so it could be supported investors in their decision.

After assigning the label for each stock, evaluation the results have some methods such as vote classifiers, Naïve Bayes in Ranking, min max classifiers [4]. The contribution of this paper is not only proposing new hybrid classifier but also using vote classifier and max classifier to evaluate the classifier results.

A case study in applying the proposal new hybrid is five stocks in VN30 of Vietnamese stock market. VN30 includes 30 stocks that are the most importance in Vietnamese stock market by capitalization and liquidity.

The rest of this paper is organized as follows. Section 2 – research methodology includes research methods, data collection and data analysis methods chosen. In Sect. 3, the proposal approach is applied in a stock. Discussions and ideas for further work and a short summary of the paper and the conclusions are presented in Sect. 4.

2   Research Methodology

2.1   Sentiment Definition

The market is driven by emotion of investors thus market sentiment is about feelings and emotion. Sentiment measures the positivity and negativity of references about the
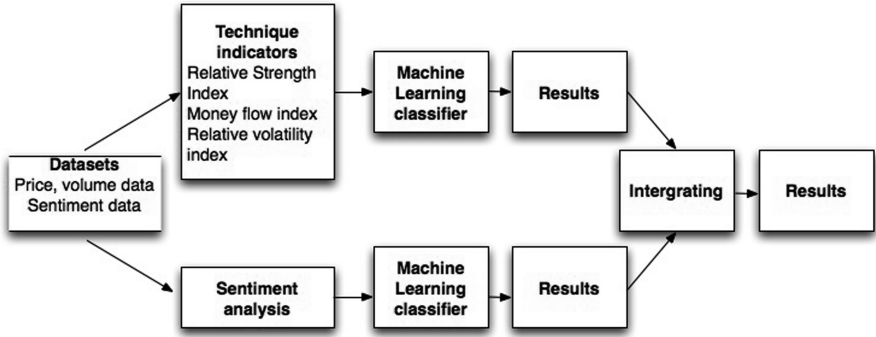
Fig. 1. Block diagram of proposal research methodology.

specific stock. The higher the measure is, the better the view of the stock is. On the other hand, market sentiment is generally described as bearish or bullish which is considered below (Fig. 1).

2.1.1 Bullish and Bearish

Returns: The close-to-close daily returns of stock i at the day t denoted $R_{t,i}$ are calculated as follows [1]:

$$R_{t,i} = 100 * \ln\left(\frac{S_{t,i}}{S_{t,i-1}}\right) \tag{1}$$

where $S_{t,i}$ is price of the stock i at the day t.

Bullish: Bullish sentiment is defined by expectations of investors who believe that stock prices will rise over time [7].

Neutral: Neutral sentiment is defined by expectations of investors who believe that stock prices will stay essentially unchanged over time [7].

Bearish: Bearish sentiment is defined by expectations of investors who believe that stock prices will fall over time [7].

2.1.2 Sentiment Ratio
Weekly, Investor's Intelligence that uses information polled directly from market professionals publish market sentiment indicator [1]. This index expresses the sentiments of investors that deal daily within the financial markets [1].

The high/low sentiment indicator compares the number of stocks making n-day highs to the number of stocks making n-day lows.

$$Ratio_{t,i} = \frac{\sum_{t=1}^{n} H(RT_{t,i})}{\sum_{t=1}^{n} L(RT_{t,i})} \tag{2}$$

where $H(RT_{t,i}) = \begin{cases} RT_{t,i}, & RT_{t,i} \geq 0 \\ 0, & RT_{t,i} < 0 \end{cases}$, $L(RT_{t,i}) = \begin{cases} 0, & RT_{t,i} \geq 0 \\ RT_{t,i}, & RT_{t,i} < 0 \end{cases}$

When stock prices are trading at their lows (highs) across the board, it means traders have a bearish (bullish) market sentiment.

2.1.3 Sentiment Trading Strategies (STS)

Sentiment trading strategies (STS): we compute the time-t returns based on the sign of the past cumulative sentiment from time $t - n - 1$ to $t - 1$. For each stock i and day t, we consider whether the past cumulative sentiment over the past n days is positive or negative. If the past cumulative sentiment over the past n days is positive, we go buy the stock. If the past cumulative sentiment over the past n days is negative, we sell the stock. We calculate a single time series of daily returns [1].

Sentiment trading strategies:

$$\sum_{t=1}^{n} RT_{t,i} \tag{3}$$

The fact that sentiment analysis is used to classify the stock. Concretely, this paper sentiment data such as bullish sentiment, neutral sentiment, bearish sentiment, sentiment ratio and sentiment trading strategies are used in classifier the stocks that will be presented in the next section.

2.1.4 Sentiment Classifier Formula

Lagarde and Arnaud proposed the effective method to classify a stock using sentiment data [4]. A strategy using sentiment data is derived. Thus, we reuse the idea in classifying sentiment that is shown by the formula as follows:

$$Target(A, t) = \ln\left[\frac{Max\left(A_{\tau \in (t \leq \tau \leq t + 182)}\right)}{A_t}\right] + \ln\left[\frac{Min\left(A_{\tau \in (t \leq \tau \leq t + 182)}\right)}{A_t}\right] \tag{4}$$

- $Target(A, t) \geq 0.0953$ will be considered as top which points to strong positive sentiment. Thus, the stock is placed into top class.
- $Target(A, t) \leq -0.223$ will be considered as bottom which points to strong negative sentiment. Thus, the stock is placed into bottom class.
- $0.0953 \geq Target(A, t) \geq -0.223$ will be considered as bottom which points to strong negative sentiment. Thus, the stock is placed into bottom class.

2.2   Technical Indicator

Technical indicators compound mathematical formulas that are used price time series data to create another time series data. These are 24 basis technical indicators are used of trading rules such as Moving average, moving average envelopes, Triple Exponential Moving Average, Average True Range (ATR), Moving Average Convergence Divergence (MACD), Bollinger Bands, bandwidth, Relative Strength Index (RSI), Relative Volatility Index (RVI), stochastic oscillator, ultimate oscillator, rate of change, relative vigor index, Ease of Movement (EMV), On Balance Volume (OBV), Accumulation Distribution Line (ADL), Chaikin Oscillator, Chaikin Money Flow (CMF) and Money Flow Index (MFI) [9]. Technical indicators can be classified into three groups that are trend, momentum and volatility based indicators. Indicators in each group are similar with each other [9] thus, this paper includes three technical indicators RVI (volatility), MFI (trend) and RSI (momentum) chosen to be on behaft of each group.

2.2.1   Relative Volatility Index (RVI)

The relative volatility index (RVI) is a volatility indicator that was developed by Donald Dorsey to indicate the direction of volatility [9]. The RVI is plotted in a range from 0 to 100 and is often used as a confirmation for other indicators such as moving average or Money Flow Index indicators.

– $S_{t,i}$ is close price of stock i at the day t, we denote

$$\text{High}_{t,i} = \begin{cases} S_{t,i}; & S_{t,t} > S_{t-1,i} \\ 0; & \text{otherwise} \end{cases}$$

$$\text{Low}_{t,i} = \begin{cases} S_{t,i}; & S_{t,t} < S_{t-1,i} \\ 0; & \text{otherwise} \end{cases}$$

$$\overline{\text{High}_{t,i}} = \frac{1}{n}\sum\nolimits_{t=1}^{n}\text{High}_{t,i}, \overline{\text{Low}_{t,i}} = \frac{1}{n}\sum\nolimits_{t=1}^{n}\text{Low}_{t,i} \tag{5}$$

$$u_{t,i} = \sqrt{\frac{1}{n-1}\sum\nolimits_{t=1}^{n}\left(\text{High}_{t,i} - \overline{\text{High}_{t,i}}\right)^{2}},$$

$$d_{t,i} = \sqrt{\frac{1}{n-1}\sum\nolimits_{t=1}^{n}\left(\text{Low}_{t,i} - \overline{\text{low}_{t,i}}\right)^{2}} \tag{6}$$

$$\text{Dsum}_{t,i} = \sum_{t=1}^{n} d_{t,i}, \text{Usum}_{t,i} = \sum_{t=1}^{n} u_{t,i}$$

The RVI formula is defined as follows

$$RVI = 100 \times \frac{Usum}{Usum + Dsum} \tag{7}$$

- RVI reading above 60 (or 70) would be considered as top which points to strong up-trend and the possibility of overbought condition. Thus, if RVI of a stock is higher than 60 (or 70), then it is placed into top class.
- RVI reading below 40 (or 30) would be considered as bottom which points to strong down-trend and the possibility of oversold condition. Thus, if RVI of a stock is lower than 40 (or 30), then it is then it is placed into bottom class.
- RVI reading between 40 to 60 (or 30 to 70) would be considered as neutral. Thus, if RVI of a stock is lower than 40 (or 30), then it is then it is placed into neutral class.

RVI is behalf volatility technical indicator group, in the next subsection we will present about volatility MFI technical indicator.

2.2.2 Money Flow Index (MFI)

MFI is an index used to measure the strength of cash flow in the market. They use both price and volume to measure sales [8]. The line is positive when the price rises, and negative when the price falls. MFI represents a momentum relative to mass, which is appropriate for determining the direction of prices and signals in the market [8]. Let $high_{t,i}, low_{t,i}, close_{t,i}, volume_{t,i}$ denote high price, low price, close price and volume of the stock i at the day t respectively. Let n be the size of time series data.

$$Typical\ price_{t,i} = \frac{high_{t,i} + low_{t,i} + close_{t,i}}{3} \tag{8}$$

$$Money\ flow_{t,i} = typical\ price_{t,i} \times volume_{t,i} \tag{9}$$

$$Positive\ money\ flow_{t,i} = \begin{cases} Money\ flow_{t,i}; & Money\ flow_{t,i} > Money\ flow_{t-1,i} \\ 0; & otherwise \end{cases} \tag{10}$$

$$Negative\ money\ flow_{t,i} = \begin{cases} Money\ flow_{t,i}; & Money\ flow_{t,i} < Money\ flow_{t-1,i} \\ 0; & otherwise \end{cases} \tag{11}$$

$$money\ ratio = \frac{\sum_{t=1}^{n} Positive\ money\ flow_{t,i}}{\sum_{t=1}^{n} Negative\ money\ flow_{t,i}} \tag{12}$$

$$MFI = 100 - \frac{100}{1 + money\ ratio} \tag{13}$$

- MFI can become overbought (>70) and prices can simply continue higher when the uptrend is strong [8]. The stock is placed into top class.

– If MFI is between 30 and 70, The stock is placed into neutral class.
– MFI can become oversold (<30) and prices can simply continue lower when the downtrend is strong [8]. The stock is placed into bottom class.

2.2.3   Relative Strength Index (RSI)

RSI is a popular momentum indicator; it represents the strength of the stock based on price fluctuations [8]. RSI is used to provide forecasts for future price fluctuations. An RSI is in the range of 0 to 100, when the RSI is less than 30, giving us a signal from the market that is concentrating on selling [8]. When the RSI is above 70 they give us a signal that the stock is being bought [8].

$$RSI = 100 - \frac{100}{1 + RS} \tag{14}$$

$$Uclose_{t,i} = \begin{cases} close_{t,i}; & close_{t,i} > close_{t-1,i} \\ 0; & otherwise \end{cases} \tag{15}$$

$$Lclose_{t,i} = \begin{cases} close_{t,i}; & close_{t,i} < close_{t-1,i} \\ 0; & otherwise \end{cases} \tag{16}$$

$$RS = \frac{\sum_{t=1}^{n} Uclose_{t,i}}{\sum_{t=1}^{n} Lclose_{t,i}} \tag{17}$$

2.3   Classification Algorithms

2.3.1   Technical Indicator Classifiers

Let $w_{x,t} = \left(w_{x,t}^1, w_{x,t}^2, w_{x,t}^3 \right)$ be the vector label of the stock x at time t that

$$w_{x,t}^1 = \begin{cases} 1; & A_{label} = \text{"top"} \\ 0; & otherwise \end{cases}$$

$$w_{x,t}^2 = \begin{cases} 1; & A_{label} = \text{"neutral"} \\ 0; & otherwise \end{cases}$$

$$w_{x,t}^3 = \begin{cases} 1; & A_{label} = \text{"bottom"} \\ 0; & otherwise \end{cases}$$

$$x \in X = \{RSI, MFI, RVI, Bullish, Bearish,$$
$$Sentiment\ ratio, Sentiment\ trading\ strategies\}$$

$$\text{Let } \mathcal{A} = \{A_1, A_2, ..A_m\} \text{ is set of stocks.}$$

Algorithm 1: Technical indicator classifiers

Input: Stock Dataset $(S_{t,A}, w_{RVI,t}, w_{MFI,t}, w_{RSI,t}, N)$

$X_1 = \{RVI, MFI, RSI\}$

Output: $A_{t,label}$

For i=1 **to** N

    Using formula (7) **return** RVI_{label}.

If (RVI_{label} ="top") **return** $w_{RVI,t} = (1,0,0)$.

Else if (RVI_{label} ="neutral") **return** $w_{RVI,t} = (0,1,0)$.

    Else return $w_{RVI,t} = (0,0,1)$.

    Using formula (13) **return** MFI_{label}

If (MFI_{label} ="top") **return** $w_{MFI,t} = (1,0,0)$.

Else if (MFI_{label} ="neutral") **return** $w_{MFI,t} = (0,1,0)$.

    Else return $w_{MFI,t} = (0,0,1)$.

    Using formula (14) **return** RSI_{label}.

If (RSI_{label} ="top") **return** $w_{RSI,t} = (1,0,0)$.

Else if (RSI_{label} ="neutral**") return** $w_{RSI,t} = (0,1,0)$.

    Else return $w_{RSI,t} = (0,0,1)$.

$$w_t = \sum_{x \in X_1} w_{x,t}$$

    If($Max_{k=1,2,3}(w_t^k) = w_t^1$) **then** $A_{t,label} = "top"$

Else If($Max_{k=1,2,3}(w_t^k) = w_t^2$) **then** $A_{t,label} = "neutral"$

Else $A_{t,label} = $ bottom.

Return $A_{t,label}$.

End.

2.3.2   Sentiment Classifiers

Let $Se_{t,A}$ denote the sentiment data value of stock A at time t.

Algorithm 2: Sentiment classifiers.

Input: Stock Dataset $(S_{et}, w_{bull,t}, w_{bear,t}, w_{ratio,t}, w_{STS,t}, N)$

$X_2 = \{Bullish, Bearish, Sentiment\ ratio, STS\}$

Output: A_{label}

For i=1 **to** N

    Using formula (2) (3) and return S_{et} sentiment data.

   Using sentiment data with formula (4) return Sentiment label.

While $(x \in X_1)$

    If $(A_{t,label} = "top")$ **return** $w_{x,t} = (1,0,0)$.

    Else if $(A_{t,label} = "neutral")$ **return** $w_{x,t} = (0,1,0)$.

      Else return $w_{x,t} = (0,0,1)$.

$$w_t = \sum_{x \in X_2} w_{x,t}.$$

If$(Max_{k=1,2,3}(w_t^k) = w_t^1)$ **then** $A_{t,label} = "top"$

Else If$(Max_{k=1,2,3}(w_t^k) = w_t^2)$ **then** $A_{t,label} = "neutral"$

Else $A_{t,label} = "bottom"$.

Return $A_{t,label}$.

End.

2.3.3   Hybrid Classifiers

Algorithm 3: Hybrid classifiers.

Input: Dataset ($w_{x,t}, N$).

Output: A_{label}

For i=1 to N

  Using Algorithm 1 return $w_{x,t}^{ind} = w_{x,t}$, $x \in X_1$, $A_{label,t}^{ind} = A_{label,t}$.

  Using Algorithm 2 return $w_{x,t}^{se} = w_{x,t}$, $x \in X_1$, $A_{label,t}^{se} = A_{label,t}$.

  If ($A_{label,t}^{ind} = A_{label,t}^{se}$) **then return** $A_{label,t}$.

  Else {

$$w_t = \sum_{x \in X_2} w_{x,t}$$

  If($Max_{k=1,2,3}(w_t^k) = w_t^1$) **then** $A_{t,label} =$ **"top".**

   Else If($Max_{k=1,2,3}(w_t^k) = w_t^2$) **then** $A_{t,label} =$ **"neutral"**

   Else $A_{t,label} =$ **"bottom".**

       }
  Return $A_{t,label}$.

  End.

---

3   Experiments and Results

3.1   Datasets

To evaluate the performance of our proposed hybrid classifier, we perform five stocks of Vietnamese stock market [6]. Vietnamese stock market is considered potential for investors. VN30 stocks being highest capitalization and liquidity in Vietnamese stock market are 80% capital of total market and 60% liquidity of total market [6]. Thus, five

stocks BID (Joint Stock Commercial Bank for Investment and Development of Vietnam), BVH (BAOVIET Holdings), CTG (Vietnam Joint Stock Commercial Bank for Industry and Trade-Main Operation Center), FPT (The Corporation for Financing and Promoting Technology) and VIC (Vingroup joint stock company), are chosen from stocks in the Top 30 stocks to apply hybrid classifier.

3.2   Experiments

The data stock prices of BID, BVH, CTG, FPT and VIC are **considered** in one year from 4$^{\text{th}}$ January 2016 to 30$^{\text{th}}$ December 2016. Thus, each stock has 252 data, the total data are 1260.

Let $S_{t,i}$ denote the price of the stock i at the day t. The price can be high, close, low, volume that depend on whatever technical indicator is used.

$$\mathcal{A} = \{VIC, FPT, CTG, BID, BVH\}.$$

N = 199.

n in the indicator classifiers is applied $n = n_{ind} = 14$.

n in the sentiment classifiers is applied $n = n_{sen} = 52$.

3.3   Results

The averages error rate and their variance for 3 algorithm are shown below (Table 1). The bold results show the best mean and variance of error rates comparing the three algorithm classifiers. Thus, the empirical results show that in hybrid classifier has not only low average error rate but also low variance of error rate which confirms its outstanding performances comparing with the left two classification methods (Fig. 2 and Tables 2, 3 and 4).

Table 1. Classification error rates of three algorithm classifiers.

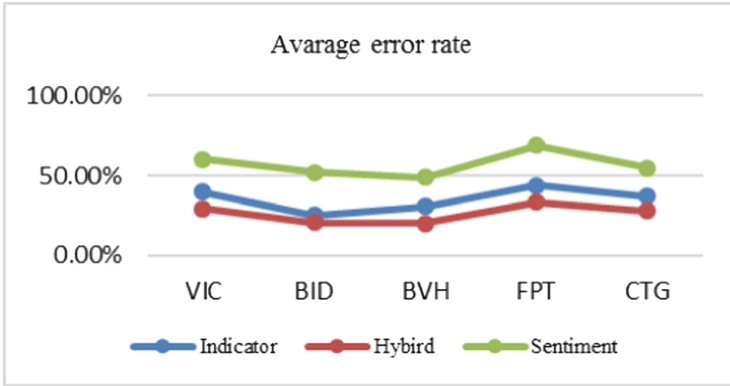|	Indicator		Sentiment		Hybrid	
	Mean	Variance	Mean	Variance	Mean	Variance
BID	0.3568	0.5699	0.3568	0.3657	**0.2060**	**0.2160**
BVH	0.3065	0.5867	0.4874	0.5347	**0.2010**	**0.2928**
CTG	0.3719	0.5212	0.5477	0.3772	**0.2814**	**0.2702**
FPT	0.4372	0.7243	0.6884	0.5040	**0.3065**	**0.2915**
VIC	0.3970	0.6747	0.6030	0.5931	**0.2915**	**0.3410**

Fig. 2. Percentage of average error rates of three algorithm classifiers.

Table 2. Correct percentage results of sentiment classifier.

Sentiment	Target				
		Top	Neutral	Bottom	
Output	Top	35	157	17	16.75%
	Neutral	109	484	182	62.45%
	Bottom	4	5	2	18.18%
	% Sentiment classifier correct	23.65%	74.92%	1.00%	

Table 3. Correct percentage results of indicator classifier.

Indicator	Target				
		Top	Neutral	Bottom	
Output	Top	34	141	34	16.27%
	Neutral	157	391	227	50.45%
	Bottom	4	4	3	27.27%
	% Indicator classifier correct	17.44%	72.95%	1.14%	

Table 4. Correct percentage results of hybrid classifier.

Hybrid	Target				
		Top	Neutral	Bottom	
Output	Top	60	133	16	28.71%
	Neutral	45	676	54	87.23%
	Bottom	4	4	3	27.27%
	% Hybrid classifier correct	55.05%	83.15%	4.11%	

4 Conclusion and Discussion

Classification in stock market is a potential and a beneficial area of research for investors. The classification of a stock into three classes that are top, neutral and bottom will be supportive for business decision makers. In this paper, historical data in Vietnamese stock market are taken over a period of one years. We attempt to classify five stocks in VN30. Different classifiers of machine learning have been used to classify future trend in these datasets. The empirical results show hybrid classifier stock have more power significantly in-sample test than single technical indicator classifier or sentiment classifier. In fact, hybrid classifier has shown good performance in terms of accuracy and low variance.

In the future work, it would be interesting to apply this classifier on a large group of stocks from different stock markets. Moreover, we consider social media data, fundamental data in addition to historical data in stock market classifier. The new data will improve our classifier results.

References

1. Gao, B., Yang, C.: Forecasting stock index futures returns with mixed-frequency sentiment. Int. Rev. Econo. Finan. **49**, 69–83 (2017). ISSN 1059-0560
2. Ballings, M., Van den Poel, D., Hespeels, N., Gryp, R.: Evaluating multiple classifiers for stock price direction prediction. Expert Syst. Appl. **42**(20), 7046–7056 (2015). ISSN 0957-4174
3. Krawczyk, B., Minku, L.L., Gama, J., Stefanowski, J., Woźniak, M.: Ensemble learning for data stream analysis: a survey. Inf. Fusion **37**, 132–156 (2017). ISSN 1566-2535
4. Lagarde, A.: Can Sentiment Indicators Signal Market Reversals? 27 May 2016. SSRN: https://ssrn.com/abstract=2785462
5. Nguyen, T.T., Nguyen, T.T.T., Pham, X.C., Liew, A.W.-C.: A novel combining classifier method based on Variational Inference. Pattern Recogn. **49**, 198–212 (2016)
6. Hoang, A.L.: Corporate governance impacts to the level of accounting information disclosure – evidence on VN30 listed companies of Vietnam. Eur. J. Account. Audit. Finan. Res. **4**(6), 1–9 (2016)
7. Mehmood, Y., Hanif, W.: Impact of bullish and bearish market on investor sentiment. Int. J. Innov. Appl. Stud. **9**(1), 142–151 (2014). ISSN 2028-9324
8. Khaidem, L., Saha, S., Dey, S.R.: Predicting the direction of stock market prices using random forest. Appl. Math. Finan. (2016)
9. Ozturk, M., Toroslu, I.H., Fidan, G.: Heuristic based trading system on Forex data using technical indicator rules. Appl. Soft Comput. **43**, 170–186 (2016). ISSN 1568-4946