GENERAL & APPLIED ECONOMICS | RESEARCH ARTICLE

The development of composite sentiment index in Indonesia based on the internet-available data
A. Rizkiana¹, H. Sari¹, P. Hardjomidjojo¹ and B. Prihartono¹

Abstract: The development of internet technology raises new sentiment measures used to predict stock market return. This raises a new problem because we must choose carefully which sentiment measures to be used to predict stock market return because various correlations and limitations of these different data sources, different sentiment measures, and its general prediction applicability to different domains are unclear. Since there are no perfect and/or uncontroversial proxies for investor sentiment, we will develop a composite sentiment index based on those different sentiment measures using principal component analysis. The investor sentiment measures we use are investor sentiment measured in social media, google search volume, and news media sentiment. We find that each investor sentiment proxies are positively related to sentiment index. We also find that investor sentiment in news media has one-day lag compared to investor sentiment in social media and investor attention in google trend. Lastly, we confirm that investor sentiment cannot be used to predict stock return.

Subjects: Quantitative Finance; Multivariate Statistics; Statistics for Business, Finance & Economics; Economic Psychology; Corporate Finance

Keywords: google search; investor sentiment; news sentiment; principal component analysis; social media

ABOUT THE AUTHOR
Our team is from Industrial Management Research Group. Industrial Management Research Group has a vision to be internationally acknowledged in the areas of system engineering and management through its contribution in universal values with local wisdom and as a group that has wide contextual knowledge, by which it is able to intellectually synthesize and interpret international knowledge within regional and national socio-cultural context. The Research topics are focused on Innovation and Business Development & Acquisition, Functional & Integrated Management, and Continuous Process Improvement linked to System Integration and Big Data & Analytics components of Industry 4.0. This research is related to functional & integrated management subtopics, especially in the field of investment and financial management, by combining the development of information technology.

PUBLIC INTEREST STATEMENT
The development of internet technology raises the possibility of new indicators that can be used to predict stock market return. These new indicators, which can be accessed publicly, can help the investor to determine the right stock to invest in. Previous research found that separately, the information in social media, google search volume, and online news prove to be useful to predict the movement of stock price. This research combined these three indicators into one composite indicator to be used to predict the stock return. As expected, each indicators positively correlated to the developed composite indicators. Also, Online news media has slower effect on the composite indicators compared to the other composite indicators. However, contrary to the findings in the previous research, the composite indicators developed cannot be used to predict the stock return. It is possibly caused by the different stock market characteristic in Indonesia.
1. Introduction

Rapid information technology development affects all industry sectors, including stock investment. Investment is one of the interesting topics in Industrial Engineering, as Industrial Engineer must find the best investment that will allow them to gain maximum profit within a given risk appetite. As there are many investment alternatives, the investors must wisely choose the investment that will maximize their return. These alternatives include the investment in stock market. Stock market is becoming more attractive for the Indonesian Investor as shown by the growing numbers of Single Investor Identification (SID) (KSEI, 2016). The number of SID on June 2017 doubled the number of SID in 2015 (KSEI, 2017). These numbers show that the people in Indonesia see investing in stock market are profitable.

There are two things that need to be considered first before investing in stock market to make sure it is profitable, that is stock selection and stock market timing (Lee & Dan Jo, 1999). Therefore, investor tries to use different strategies to maximize their return. There are two strategies commonly used by investor to determine the right stock and the right time to invest, namely technical analysis and fundamental analysis. Technical analysis uses a set of rules or chart to anticipate the movement of stock price based on past information, for example, opening price, closing price, and trading volume (Nazário, E Silva, Sobreiro, & Dan Kimura, 2017). While the fundamental analysis uses company fundamental data, such as financial data, market, government, politics, environment data available to assume the movement of stock price (Nassirtoussi et al., 2014).

Although these two strategies are commonly used by rational investors, there is another strategy that use investor sentiment as its main driver to buy or sell the stocks. Investor sentiment is defined as the belief about investment cash flow and risk that is not justified by available information (Baker & Wurgler, 2007). This strategy is often used by irrational traders as they are prone to exogenous sentiment (De Long et al., 1990). Interestingly, it is found from previous research that investor sentiment has a significant impact on stock price (Chau, Deesomsak, & Dan Koutmos, 2016) and improve the predictability of stock return (Muradoglu dan Harvey, 2012). Baker (2007) also arguments that nowadays the question is no longer whether investor sentiment affects stock price, but rather how to measure investor sentiment and quantify its effect. This is especially true in the era of information technology that enables the information (sentiment) to spread faster than before. So, in this research, we will focus on developing new investor sentiment for Indonesia stock market and test its relationship with the stock prices.

There are several measures that are commonly used in the previous research to measure investor sentiment. For example, Brown dan Cliff (2005) uses survey data from Investor’s Intelligence (II) which tracks the number of market newsletters that are bullish, bearish, or neutral. Baker dan Wurgler (2006) uses an overall sentiment index from six different market data, i.e. closed-end fund discount, stock turnover, the number and average first-day returns on IPOs, the equity share in new issues, and the dividend premium. Tetlock (2007) uses news media content to predict movement in broad indicators of stock market activity. Recently, due to the rapid development of information technology, Renault (2017) uses social media messages to measure investor sentiment and find that this measure helps forecast the market return.

Those different investor sentiment measures raise a new problem. We must choose carefully which sentiment measures to be used to predict stock market return because various correlations and limitations of these different data sources, different sentiment measures, and its general prediction applicability to different domains is unclear (Mao, Counts, & Dan Bollen, 2011). Since there are no perfect and/or uncontroversial proxies for investor sentiment, as argued by Baker and Dan Wurgler (2006), we will develop a composite sentiment index based on those different sentiment measures using principal component analysis (PCA). It is similar with the approach from Baker and Wurgler (2006, 2007) to develop composite sentiment index, but we will use different investor sentiment proxies because not all of the proxies from previous research
applicable for Indonesia stock market and we will also include social media message to reflect the recent development of investor sentiment measurement in the index calculation.

Therefore, our paper has these following contributions. First, as far as we know, this is the first paper to develop investor sentiment index for Indonesia Stock Market. Second, different from previous research, we include the recent development of sentiment measures, i.e., social media message as one of the proxies in sentiment investor index. Third, we will use AutoRegression and granger causality to test lead–lag relationship between our sentiment investor index and stock return. Fourth, we will analyse the relationship between investor sentiment and stock return on individual stock level, not on stock market level.

2. Methods
The methodology for developing the investor sentiment index consisted of four steps. First, we will review recent literature about investor sentiment then choose and modify, if necessary, several investor sentiment proxies that are relevant for Indonesia Stock Market based on the availability of the data. Second, we will collect the data for investor sentiment and stock price. Third, we will use PCA to develop sentiment index proxies with approach similar to Baker and Dan Wurgler (2006) research. Fourth, we will test the lead–lag relationship between investor sentiment index we develop and stock return using VAR and granger causality test.

2.1. Investor sentiment proxies
First step for developing the investor sentiment index is to review recent literature about investor sentiment measure to collect investor sentiment proxies which are used in the previous research. The following table summarized investor sentiment proxies.

Table 1 shows that there is a shift of sentiments measure, from traditional measures, based on financial statement and market data, to a more modern measure based on data available on the internet. This is in line with the developments of information technology that influence people’s behavior in stock investment. Based on the literature review above and the availability of the data in Indonesia, we select trading volume, social media data, news media, and search engine data to

<table>
<thead>
<tr>
<th>Measure of sentiment</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey data (Investor Intelligence market newsletter)</td>
<td>Brown and Dan Cliff (2005)</td>
</tr>
<tr>
<td>Closed-end fund discount</td>
<td>Baker and Dan Wurgler (2006)</td>
</tr>
<tr>
<td>NYSE share turnover</td>
<td>Baker and Wurgler (2007)</td>
</tr>
<tr>
<td>The number and average first-day returns on IPOs</td>
<td></td>
</tr>
<tr>
<td>Equity share in new issues</td>
<td></td>
</tr>
<tr>
<td>Dividend premium</td>
<td></td>
</tr>
<tr>
<td>Consumer confidence index</td>
<td>Schmeling (2009)</td>
</tr>
<tr>
<td>Trading volume</td>
<td>Chuang, Ouyang, and Dan La (2010)</td>
</tr>
<tr>
<td>Survey data (Investor Intelligence market newsletter)</td>
<td>Mao et al. (2011)</td>
</tr>
<tr>
<td>Search engine data (google)</td>
<td></td>
</tr>
<tr>
<td>Social media data (Twitter)</td>
<td></td>
</tr>
<tr>
<td>Advanced Declining Ratios</td>
<td>Naik and Padhi (2016)</td>
</tr>
<tr>
<td>Put-Call Ratios</td>
<td></td>
</tr>
<tr>
<td>NIPO</td>
<td></td>
</tr>
<tr>
<td>PE ratios</td>
<td></td>
</tr>
<tr>
<td>Turnover rates</td>
<td></td>
</tr>
<tr>
<td>Trading volume</td>
<td></td>
</tr>
<tr>
<td>Mutual Funds Net Flow</td>
<td></td>
</tr>
<tr>
<td>Social media data (Stocktwits)</td>
<td>Renault (2017)</td>
</tr>
</tbody>
</table>
construct the composite sentiment index. We did not select the other measures because some of the measures are not available in Indonesia (i.e. survey data) and the rest are available only for aggregate stock market, not for specific individual stocks (i.e. NYSE share turnover, consumer confidence index, NIPO, equity shares in new issues, etc.).

Trading volume reflects investor’s expectations about particular stocks thus change in trading volume can reflect fluctuations in investor sentiment. When most investors think companies are good to invest, they buy stocks, and the trading volume of those stocks goes higher, and when most investors consider companies are bad, they sell or stop buying stocks, the trading volume of those stocks would be lower (Chuang et al., 2010). Social media data, which is measured by the number of positive comments compared to negative comments, can be used as investor sentiment measure because it reflects information demand and reaction of society about stock’s future performance (Bukovina, 2016). Investor sentiment in social media positively correlated with market return, i.e. bullish comments in social media correlated with positive market return, and vice versa (Mao et al., 2011).

Investor sentiment in news media is measured by the number of bad news compared to all news in a given day (Mao et al., 2011). We calculate only the number of bad news because pessimistic media content variable is more predictive on inducing downward pressure on market prices (Tetlock, 2007). And lastly, search engine data are measured by using google trend shows the search frequency volume for the specific keywords. Increasing number of search in google reflects increasing investor attention to buy a particular stock and consequently make the stock price rise (Da, Engelberg, & Dan Gao, 2011).

2.2. Investor sentiment proxies data

Different from previous research which mainly analyzes the stock on aggregate level, we analyze the individual stock. To limit our research, we choose five banking stock listed on Indonesia Stock Exchange, namely Bank Negara Indonesia (BNI), Bank Rakyat Indonesia (BRI), Bank Central Asia (BCA), Bank Mandiri (BMRI), and Bank Tabungan Negara (BTN).

Then, to develop a composite sentiment index, we collect the data from several sources based on the investor sentiment measures discussed before, for 8 months, from 1 April 2017 to 1 December 2017. We retrieved the stock trading volume for those five banking stocks from yahoo finance website (i.e. yahoo.finance.com). We collect the social media messages for five banking stocks from Stockbit (i.e. social media for stock investment in Indonesia) using webscraping tools in Python. Then, we manually classify the messages as bullish, if it contains positive comments, and bearish, if it contains negative comments. Last, we calculate investor sentiment, SENT, in social media using Equation 1 (Mao et al., 2011):

\[
SENT_t = \frac{N_{positive}}{N_{positive} + N_{negative}}
\]

where \(N_{positive}\) is the number of bullish message on day \(t\) and \(N_{negative}\) is the number of bearish message on day \(t\).

We retrieved the data for news media from Kontan website, which is a news website specifically for economic and finance news. First, we filter the news based on article tag for five banking company, i.e. BBNI, BBRI, BBTN, BMRI, and BCA. Second, we retrieved the headline of the filtered articles from 1 April 2017 to 1 December 2017. Third, we assign the value of +1 for bullish news and value of −1 for bearish news then calculate the news sentiment, NEWS, as:

\[
NEWS_t = \frac{N_{positive} - N_{negative}}{\text{total number of news}}
\]
where $N_{\text{positive}}$ is the number of bullish news headline on day $t$ and $N_{\text{negative}}$ is the number of bearish news headline on day $t$.

Finally, we retrieved the search engine data from the google trend website (trend.google.com). We use stock ticker for each stock, i.e. BBNI, BBRI, BBTN, BMRI, and BBCA as search keyword and download the daily google search frequency volume, namely Search Volume Index, from 1 April 2017 to 1 December 2017. We use stock ticker for the search keywords because it is simple and short that ensure only the relevant information related to company stock will be obtained (Vozlyublennaia, 2014). We choose not to use the keywords from Mao et al. (2011) because it reflects the investor sentiment for aggregate stock market, not for individual company. We calculate investor attention using formula (2), following Da et al. (2011) methodology.

\[
\text{Attention}_t = \log(SVI_t).
\]  

(3)

### 2.3. Investor sentiment index using PCA

There are two issues that we need to address to develop investor sentiment index. The first issue is each investor sentiment proxies chosen are likely to include two components, that is sentiment component as well as idiosyncratic components which is not related to sentiments. The second issue is determining the relative timing of the variables, that is, if they exhibit lead–lag relationship, some variable may reflect the change in sentiment earlier than others. Therefore, we follow the approach similar to Baker and Dan Wurgler (2006) and Naik and Padhi (2016) to isolate the common components and also to incorporate the fact that some variables take longer to reveal the same sentiments. To remove the business cycle variation and fundamental components, each of the sentiment proxies has been regressed on two macroeconomic fundamentals, namely exchange rate and Indonesia Composite Index (IHSG), using formula (3).

\[
Y_t = \alpha_0 + \beta_k \sum_{k=1}^K \text{Fund}a_{kt} + \epsilon_t.
\]  

(4)

where $Y_t$ represents each investor sentiment proxies, $\text{Fund}a_{kt}$ represents two macroeconomic fundamentals, i.e. exchange rate and Indonesia Composite Index, and $\epsilon_t$ is the error term. The fitted equations would provide the rational components and the residuals capture the irrational component of sentiments. The residuals from these regressions, containing only irrational (sentiment) components, will be analyzed further to construct investor sentiment index. Different from previous research, in this research only two macroeconomic fundamentals selected because it is available in daily basis. Specifically, first, we estimate the first principal components of three proxies and their lags. Because we have three sentiment proxies, we will get a first-stage index with six loadings, one for each of the lagged and current proxies. Second, we compute the correlation coefficient between the first-stage index with the lagged and current proxies. Third, we select lagged or current proxies, whichever has a higher correlation with the first-stage index, then construct final sentiment index using PCA.

### 2.4. Investor sentiment index and stock return

Finally, we test the relationship of investor sentiment index and stock return using two linear regression models. The first model uses investor sentiment as independent variable and stock return as dependent variable. The second model uses investor sentiment as dependent variable and stock return as independent variable.

### 3. Results and discussion

#### 3.1. Data collection

After collecting the comments from Stockbit for five banking companies and manually classify the comments as bullish or bearish, we get 271 bullish comments and 70 bearish comments for BBCA, 488 bullish comments and 90 bearish comments for BBNI, 232 bullish comments and 80 bearish comments...
comments for BBRI, 355 bullish comments and 141 bearish comments for BBTN, and 549 bullish comments and 118 bearish comments for BMRI. As for the news media data, we get 5 bullish news for BBNI, 7 bullish news and 2 bearish news for BBCA, 4 bullish news for BBTN, 10 bullish news and 3 bearish news for BBRI, and 11 bullish news and 1 bearish news for BMRI. Lastly, we get 165 days-search volume data from google trend for each company ticker.

3.2. Data processing

Following the methods we explained in the previous section, we (1) remove the business cycle variation and fundamental components by regressing the variables using formula (3), (2) using the residuals from these regressions to form first-stage index using principal component analysis, (3) calculate the correlation between first-stage index and the current and lagged proxies, and (4) construct final sentiment index using PCA based on the current or lagged proxies, whichever has a higher correlation with the first-stage index. The final sentiment index for each sample is as follows:

\[
\text{SENTIMENT}_{t}^{\text{BBCA}} = 0.724 \text{SEN}_{t-1}^{\text{BBCA}} + 0.251 \text{Attention}_{t-1}^{\text{BBCA}} + 0.766 \text{NEWS}_{t}^{\text{BBCA}}
\]

\[
\text{SENTIMENT}_{t}^{\text{BBNI}} = 0.639 \text{SEN}_{t-1}^{\text{BBNI}} + 0.654 \text{Attention}_{t-1}^{\text{BBNI}} + 0.631 \text{NEWS}_{t}^{\text{BBNI}}
\]

\[
\text{SENTIMENT}_{t}^{\text{BBRI}} = 0.735 \text{SEN}_{t}^{\text{BBRI}} - 0.396 \text{Attention}_{t-1}^{\text{BBRI}} + 0.758 \text{NEWS}_{t}^{\text{BBRI}}
\]

\[
\text{SENTIMENT}_{t}^{\text{BBTN}} = 0.605 \text{SEN}_{t-1}^{\text{BBTN}} + 0.677 \text{Attention}_{t-1}^{\text{BBTN}} - 0.699 \text{NEWS}_{t}^{\text{BBTN}}
\]

\[
\text{SENTIMENT}_{t}^{\text{BMRI}} = 0.541 \text{SEN}_{t}^{\text{BMRI}} + 0.660 \text{Attention}_{t-1}^{\text{BMRI}} + 0.720 \text{NEWS}_{t}^{\text{BMRI}}
\]

where each of the index components is standardized residuals. Here, the final sentiment index for BBCA explains 39.141% of the sample variance, final sentiment index for BBNI explains 41.128% of the sample variance, final sentiment index for BBRI explains 42.366% of the sample variance, final sentiment index for BBTN explains 43.745% of the sample variance, and final sentiment index for BMRI 41.57% of the sample variance. There are several things that we can conclude from the sentiment index. First, we have mixed results for each component signs. The expected signs for investor sentiment, investor attention, and news sentiments are all positive because logically when there is an increase of investor sentiment because of bullish comments, or increasing investor attention because of investor search frequency toward particular stock, or the release of bullish news should all have a positive effect on overall sentiment index. So, we have two index components that have opposite signs from the expectation, that is the signs for investor attention in BBRI’s sentiment index and news sentiment in BBTN’s sentiment index. Second, we also have mixed results for the timings for each sentiment index components. But, we can conclude the timings for each sentiment index components based on the timing pattern from five sentiment index we formed. For sentiment investor in social media, we can conclude that the investor sentiment in social media for previous day is used to form sentiment index. But this conclusion needs further test because out of five stock samples, only three samples have this relationship. Investor attention also has this same relationship, that the component that forms sentiment index is investor attention from the previous day and this result is quite robust because four out of five samples show this relationship. However, news sentiment has different relationship. The sentiment index is formed by today news sentiment, not previous days news sentiment. Furthermore, this result is quite robust because four out of five samples have this relationship. We argue the reason of this finding is that the investor sentiment in social media and google search frequency in google trend are faster to reflect the investor sentiment than the news sentiment. Therefore, previous day investor sentiment in social media and previous day investor attention in google trend have more common variance with investor sentiment in the news today. In other words, there is a one-day lag for the same investor sentiment reflected in news compared to investor sentiment in social media and investor attention in google trend.
Finally, we test the predictability of stock return using sentiment index as the predictor and vice versa using linear regression. In Table 2, we can see that both models we develop are not significant. In other words, the sentiment index we develop cannot predict the stock return, also the stock return cannot predict the sentiment index. This is different from the finding of Mao et al. (2011) and Tetlock (2007). But, the insignificant relationship from sentiment index to stock return aligns with the finding of Rizkiana, Sari, Hardjomijoyo, Prihartono, and Yudhistira (2017) and Rizkiana et al. (2018) using different methods and measure.

4. Conclusion
In this research, we develop sentiment index using investor sentiment in social media, investor attention in google trend, and investor sentiment in news media. We find that each investor sentiment proxies are positively related to sentiment index. We also find that investor sentiment in news media has one-day lag compared to investor sentiment in social media and investor attention in google trend. We also confirm our finding in the previous research that investor sentiment cannot be used to predict the stock return.

Acknowledgements
This research is supported by The Eminent Research of Higher Education Program from the Ministry of Research, Technology, and Higher Research of Indonesia.

Funding
This work was supported by the Ministry of Research, Technology, and Higher Education of Indonesia.

Author details
A. Rizkiana1
E-mail: anaikzir@gmail.com
H. Sari1
E-mail: hasrini@ti.itb.ac.id
P. Hardjomijoyo1
E-mail: hhanjaya@yahoo.com
B. Prihartono1
E-mail: budhi.prihartono@widyatama.ac.id

1 Department of Industrial Engineering and Management, Institut Teknologi Bandung, Bandung, Indonesia.

Citation information
Cite this article as: The development of composite sentiment index in Indonesia based on the internet-available data, A. Rizkiana, H. Sari, P. Hardjomijoyo & B. Prihartono, Cogent Economics & Finance (2019), 7: 1669399.

References

Table 2. Linear regression result

<table>
<thead>
<tr>
<th>No</th>
<th>Stock</th>
<th>SENTIMENT+RETURN</th>
<th>RETURN+SENTIMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BBCA</td>
<td>P = 0.091</td>
<td>P = 0.091</td>
</tr>
<tr>
<td>2</td>
<td>BBNI</td>
<td>P = 0.059</td>
<td>P = 0.059</td>
</tr>
<tr>
<td>3</td>
<td>BBRI</td>
<td>P = 0.320</td>
<td>P = 0.320</td>
</tr>
<tr>
<td>4</td>
<td>BBTN</td>
<td>P = 0.936</td>
<td>P = 0.936</td>
</tr>
<tr>
<td>5</td>
<td>BMRI</td>
<td>P = 0.467</td>
<td>P = 0.467</td>
</tr>
</tbody>
</table>

© 2019 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.
You are free to:
Share — copy and redistribute the material in any medium or format.
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.
Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions
You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.