ICIC Express Letters
Part B: Applications
Volume **9**, Number **2**, February **2018**
ICIC International ©2018 ISSN 2185-2766
pp. **155–162**

CORRUPTION SYSTEM DEVELOPMENT BASED ON INDONESIA'S CORRUPTION PERCEPTION INDEX

Noerlina[1], Lili Ayu Wulandhari[2], Yulyani Arifin[2], Sasmoko[3,4]
Andi Muhammad Muqsith Ashari[2], Mohammad Alamsyah[2]
Kelvin[5] and Hanindha Soenarto[5]

[1]Information System Department
School of Information System
[2]Computer Science Department
School of Computer Science
[3]Primary Teacher Education Department
[5]Psychology Department
Faculty of Humanities
[4]Research Interest Group in Education Technology
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah Jakarta 11480, Indonesia
Nurlina@binus.edu

Received August 2017; accepted November 2017

Abstract. *The purpose of this study is to develop the corruption system in Indonesia based on online news as it is considered to reflect Indonesia's current condition. The government has an important role in economic growth. However, there are still many civil servants who abuse their positions to do corruption. The central government should pay more attention to each region to avoid corruption cases. For developing a mapping system, Naïve Bayes classifier is needed to classify news about corruption and non-corruption. N-Gram and Hash Table were used to map corruption cases based on Indonesia's administrative territory. The experimental results showed that Naïve Bayes classification achieved 100% accuracy for training and testing. However, N-Gram and Hash Table achieved 85% accuracy for mapping the location of the article.*
Keywords: Corruption, Indonesia, System development, Corruption perception index

1. **Introduction.** Indonesia is a country with the largest economic development in Southeast Asia [1], as shown in gross domestic product data by World Bank in 2016 with 932.259 million dollars [2]. Indonesia economic is industry and agriculture-based. The government plays an important role in national economic growth. Nevertheless, there are still many occurrences of government officers abusing their power to do corruption. In this order, the central government should pay attention to every area in the nation to avoid corruption case. Meanwhile, the news media always constantly preach about corruption case, which makes the news media relevant for being one of the sources of measurement of corruption perception index [3].

Corruption is a misappropriation or abuse of public resources or public power for personal gain or others [4]. Moreover, corruption is the major problem faced by developing countries, including Indonesia, because corruption is perceived to complicate good governance. Also, corruption can scrap the ability of government because of the abandonment of procedure, the depletion of resources, and job appointment without legal process [5]. Corruption also includes an abuse of power by the officials such as embezzlement, nepotism, extortion, bribery, and fraud by irresponsible people. Acts of corruption are included in the category of serious and harmful criminal acts to the nation and state. Therefore,

law enforcement must work harder and more thoroughly in abolishing actions related to corruption.

The mass media coverage can reflect the current condition of some places. Besides, mass media can sum up ideas and thoughts of society on social reality [19]. Thus, the use of social media as a source of corruption perception index is very relevant. Moreover, corruption perception index can be a benchmark for corruption mapping levels in some regions. Corruption mapping will help the government analyze and manage corruption in each region. Therefore, this study proposed corruption system development based on Indonesia's corruption perception index. This study is the continuation of the previous research which was about the corruption cases mapping based on Indonesia's corruption perception index [3].

The problem formulation of this research is how to develop a corruption case mapping system based on the perception of Indonesia online mass media using Naïve Bayes algorithm.

- How to classify news stories about corruption or non-corruption.
- How to identify the place of the news using the text mining approach.
- How to map the location of corruption based on Indonesia's online news media.

The methodology used in this research consists of three techniques, which are web crawling, web scraping, and cron to process the data of the study. And based on the study, it can be concluded that web scraping and web crawling process are effective techniques to obtain reliable article data, and Naïve Bayes technique can classify the news about corruption and non-corruption with 100% accuracy for testing.

2. **Problem Statement and Preliminaries.** Corruption is the behavior of public officials, politicians, or civil servants who unreasonably enrich themselves or those close to them by abusing entrusted public authority [6]. In general, corruption is divided into three categories or types [7], which are as follows.

1) Grand corruption. It is regarding large amount of public resources being taken or misused by several public officials;
2) State or regulatory capture. It is related to the public institution and private institution that gain special benefit by collusion;
3) Bureaucratic or Petty Corruption. It is related to a large number of public officials that were misusing their power to get bribes.

Corruption has major impact on various aspects, particularly on the economic aspect. According to [8], there are four impacts of corruption.

1) Corruption can weaken the investment. So, it reduces country's economic growth.
2) Corruption causes the misallocation people in some position, or it is called talent miss allocated.
3) Corruption causes improper allocation of funds.
4) The quantity and quality of public goods and services are inadequate because corruption makes tax revenue reduced that affects the composition of government spending.

Data mining is an extraction of information from the data inside the database storage. Data mining is also known as knowledge discovery in database or KDD [9]. However, according to [10], data mining is a selection or process of digging knowledge from the large amounts of data. It can be concluded that data mining is the process of analyzing large amounts of data to get useful information.

According to [10], data mining can be categorized into two, which are:

1) Predictive. It is a process to find the pattern from data by using several future variables. The example is the classification data. The goal is to predict the value of certain attributes based on other attributes;

2) Descriptive. It is a process to find the important characteristic of data in the database and its goal to reduce patterns that summarize the main relation between data. In this category, post-processing techniques are usually used to validate and explain the result.

Data mining has three main functions [11] which are as follows.

1) Explanatory. It is to explain several observation activities or condition;
2) Confirmatory. It is to confirm an existing hypothesis;
3) Exploratory. It is to analyze new data on an unusual relation or anomalous relation.

3. **Research Method.** The main data in this study was news content that was taken from several online news media in Indonesia. In the data process there are three techniques used in this research, which are as follows.

1) Web crawling is a software that runs recursively crawl from one page to another page by following the link in each page [12]. In this research, crawling result will be saved in the database as a news article index for web scraping purpose;
2) Web scraping is a software technique that gathers and extracts information from a web page [12]; in this research, web scraping is used to get news content based on news article index and save the content into the database;
3) Cron scheduling is a UNIX and Linux base scheduling to run a command or a shell script periodically; in this research, cron scheduling is used to run web crawling and web scraping periodically.

There are seven processes during the extraction of information [10]:

1) Data cleaning is a process of removing noise and irrelevant data or inconsistent data;
2) Data integration is the incorporation of data from various data mining into a new data mining;
3) Data selection is the process of data selection on data mining that will be analyzed because not all the data can be analyzed;
4) Data transformation is the process of transforming or merging of data into suitable format to be processed in data mining;
5) Application of techniques data mining is a main process where a method is applied to finding valuable knowledge and hidden from the data;
6) Pattern evaluation is the process to identify patterns to be represented in knowledge base;
7) Knowledge presentation is visualization and presentation of knowledge about the technique to obtain knowledge from user.

The algorithm used in the data mining to classify the data set was Naïve Bayes classifier. This algorithm is a classification of probability and statistical model. It predicts future opportunities based on previous experiences, and this is also called Bayes' Theorem. Naïve Bayes for each decision class, calculates the probability on the condition that the decision class is true by giving the object information vector. This algorithm assumes that attribute of object is independent [14]. Moreover, this algorithm works very well compared to classifier model [15].

The next process used confusion matrix. Confusion matrix is visualization tools used for supervised learning. Each column in matrix is the example of prediction class, and each row represents an event in actual class [16].

The next stage is data mapping, where the process is:

a) Indonesia Administrative Division

The territory division in Indonesia is administered by local government in their respective territorial boundaries. It is in accordance with the 1945 Constitution of the Republic of Indonesia of the second amendment in Chapter VI on Regional Government Article 18 Paragraph 1, which states that the Unitary State of the Republic of Indonesia

shall be divided into provinces and those provinces shall be divided into regencies and municipalities, each of which shall have regional authorities which shall be regulated by law. Indonesia has 34 provinces and 514 regencies and municipalities.

b) Location Extraction with N-Gram Algorithm

N-Gram is a method to cut or separate string in sentences or words. N-Gram is applied to taking a certain number (n) of strings or characters of a word on ongoing basis from start to finish [17].

c) Adjacency List

Data of corruption case that will be mapped is hierarchical in which each province has several sub-municipalities. Adjacency list can be used to map by representing each region with an index number [18].

4. **Results.**

4.1. **Database design.** In this study, there are several data to be stored in database. Those data are as follows: (1) News site index generated at the data collection stage using web crawling techniques and stored in the form of files with JSON format. (2) News produced at data collection by using web scraping is stored in table format. This table stored the information of table id, news title (title), news content (content), news author (author), original news URL (original_url), news picture (thumbnail), news published date (published_date), date of news created in database (created_at), date of updated data in database (updated_at) and date of data deleted from database (deleted_at) which have initial value 0 and 1 if it has been deleted. Information has features that had extracted through web scraping in the form of data type, maximum data length, and index. (3) Next, data from web crawling and web scraping process will go through text processing to extract sentence features that indicate news class (corruption/non-corruption) and the location of corruption case. This table stored table id (id), information id of processed article (article_id), sentence that is identifying location on the article (location_string), information of class article (corruption_article) that the value is "YES" if included in the corruption article class and "NO" if not included in non-corruption articles class, province where the news has occurred (province_id), city or regency where the news has occurred (regency_id). The database also stored data that will be used for data training process and it is distinguished by one attribute of is_training. (4) For measuring the accuracy of the classification and location extraction of an article, the results of the classified and extracted location will be compared with the result of the manual measurement, which is stored in Figure 1.

The database design is depicted in entity relation diagram (ERD).

4.2. **Development of news classification algorithm.** In this study, classification is divided into two classes, which are corruption and non-corruption then in the equation of variable $V_j = \{\text{Corruption, Non-corruption}\}$. So the equation becomes:

$$V_{MAP} = \underset{V \in \{\text{Corruption, Non-corruption}\}}{\operatorname{argmax}} \prod_{i=1}^{n} (P(x_i|V_j)P(V_j)) \tag{1}$$

x_i: News i $(1, 2, 3, \ldots, n)$

V_j: Class j (Corruption, Non-corruption)

$P(x_i|V_j)$: Probability of x_i at class V_j, where $P(x_i|V_j) = \frac{n_k+1}{n+|\text{data}|}$

$P(V_j)$: Probability of V_j, where $P(V_j) = \frac{|\text{vocabulary}_j|}{|\text{vocabulary}|}$

V_{MAP} is the highest probability for corruption and non-corruption class, n is the news total, and x_i is news that is being classified in V_j class.

$P(V_j)$ and $P(x_i|V_j)$ are calculated during training or training data, where

$|\text{vocabulary}_j|$: the number of words in j class

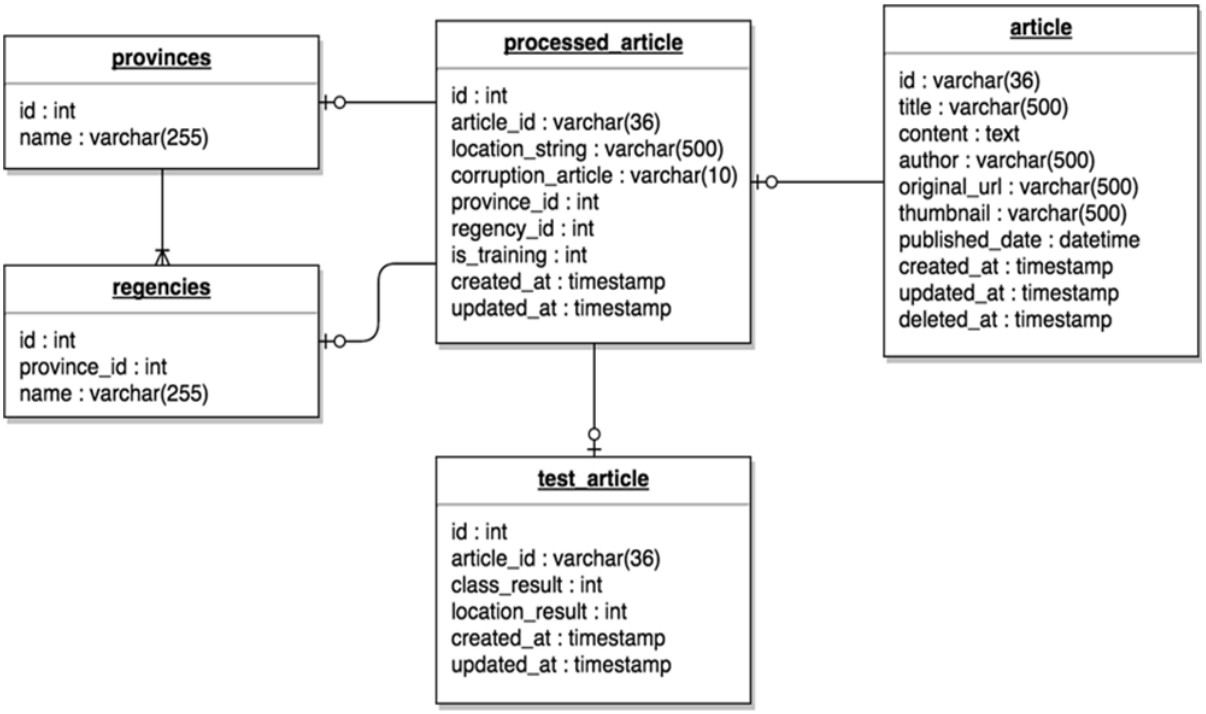

FIGURE 1. Entity relation diagram (ERD)

|vocabulary|: the number of unique words in all class
n_k: the number of emergence frequency of each word
n: the number of word emergence frequency in each class
|data|: the number of all word from all class
Classification has two main processes, which are:

a) Naïve Bayes Classifier Learning Process
For training and learning process of Naïve Bayes, the |vocabulary| must be determined in exercise article.

b) Classification Process
In classification stage, D3 article data is used for data test with the description of D3: KPK called Setya Novanto as witness.

Based on the reference of learning process, the calculation result is

$$\begin{aligned} V_{MAP} &= \underset{V\in\{\text{Corruption, Non-corruption}\}}{\text{argmax}} \prod_{i=1}^{n}(P(x_i|V_j)P(V_j)) \\ V_{MAP} &= \underset{V\in\{\text{Corruption, Non-corruption}\}}{\text{argmax}} P(V_j)P(\text{setya}|V_j)P(\text{novanto}|V_j) \\ &\quad P(\text{called}|V_j)P(KPK|V_j)P(\text{witness}|V_j) \end{aligned} \tag{2}$$

V_{MAP} for corruption class

$$V_{MAP} = \underset{V\in\{\text{Corruption}\}}{\text{argmax}} \frac{1}{2} * \frac{1}{9} * \frac{1}{9} * \frac{1}{9} * \frac{2}{9} * \frac{1}{9} = 0.0000169350878084$$

V_{MAP} for non-corruption class

$$V_{MAP} = \underset{V\in\{\text{Non-corruption}\}}{\text{argmax}} \frac{1}{2} * \frac{1}{9} * \frac{1}{9} * \frac{1}{9} * \frac{1}{9} * \frac{1}{9} = 0.0000084675439042$$

From the value of V_{MAP}, the highest value was taken into the category of article. Based on above description, it is found that V_{MAP} for the category of corruption, has the

highest value compared to other categories. Thus, the article is classified as a category of corruption.

4.3. **The mapping of news location with N-Gram.** N-Gram method is distinguished by the number of strings or character cutting that is being processed. There are many kinds of N-Gram type which are Uni-Gram for the cutting of one character string, Bi-Gram for the cutting of two character strings, Tri-Gram for the cutting of three character strings, Quad-Gram for the cutting of four character strings, and so on.

For this study, the cutting of string was based on space and the number of N = 5, because the biggest number of words in administrative area of Indonesia is five words that are "KABUPATEN PENUKAL ABAB LEMATANG ILIR". The searching of location was performed by using Hash Table algorithm for the location that had found in the database.

4.4. **System summary.** In Figure 2, system summary describes the percentage of total article classes, classroom testing result, and overall location testing result.

4.5. **Location mapping.** In this component, Figure 3 showed the result of corruption articles mapping on the division of administrative territory of Indonesia.

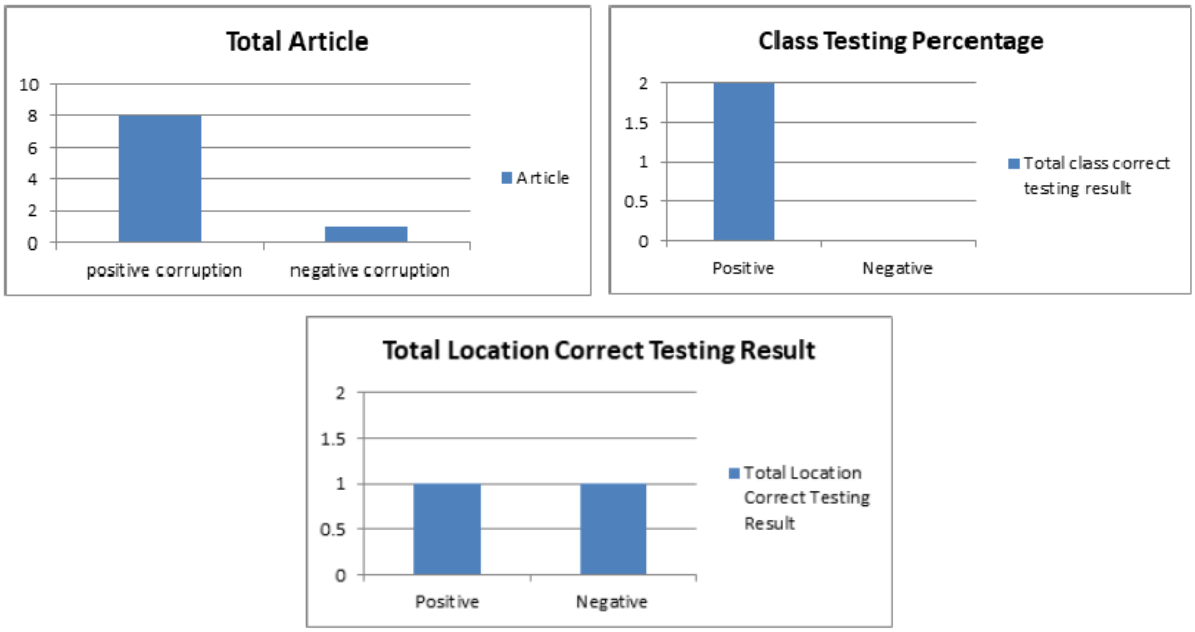

FIGURE 2. System summary

LOCATION MAPPING

Location	Article Count
SUMATERA UTARA	1
KABUPATEN TAPANULI UTARA	1
JAKARTA	4
KOTA JAKARTA SELATAN	2
KOTA JAKARTA PUSAT	1
KOTA JAKARTA UTARA	1
JAWA BARAT	1
KABUPATEN BANDUNG	1
BANTEN	1

FIGURE 3. Location mapping

5. **Validation and Evaluation.** The validation of the Indonesian corruption case mapping model was carried out in two phases: training and testing. Training was to classify the news types of corruption and non-corruption, and the selection of sentences that will be input on N-Gram to find out the location of news about corruption. In this process, the data came from the results of web crawling and web scraping process.

In the training stage of news classification, 50 positive articles and 50 negative articles about corruption were used. In the testing stage, 15 news were used to identify its news class. Based on the experiment, it obtained 100% accuracy for testing. Location mapping with N-Gram algorithm and Hash Table algorithm showed 85% accuracy for mapping areas of provincial and municipal levels, which came from 100 news which were read manually. From reading the news, their locations would be known and then would be inputted to the system on and on until all 100 news had been inputted. After that, the 100 news were tested again by inputting them to an application and were again checked one by one manually. The result is that from 100 news there are 85 news whose locations are accurate.

6. **Conclusions.** Based on this study, it can be concluded that web scraping and web crawling process are an effective technique to obtain reliable article data, and Naïve Bayes technique can classify the news about corruption and non-corruption with 100% accuracy for testing. Location mapping with N-Gram algorithm and Hash Table algorithm showed 85% accuracy for mapping areas of provincial and municipal levels. From the process of text mining, the result was the data regarding corruption mode based on the news phenomenon from online media related to corruption. It will be processed further to extract the location referred by the article. Thus, forming the data of the number of articles that have been classified by the location can help the central government to perform certain actions on a particular place based on the data result of the text mining. It is expected to reduce cases of corruption in various regions in Indonesia.

REFERENCES

[1] World Bank, *Overview*, http://www.worldbank.org/en/country/indonesia/overview, 2014.

[2] World Bank, *Gross Domestic Product 2016*, http://databank.worldbank.org/data/download/GDP.pdf, 2017.

[3] Noerlina, L. A. Wulandhari, Sasmoko, A. M. Muqsith and M. Alamsyah, Corruption cases mapping based on Indonesia's corruption perception index, *Journal of Physics: Conference Series*, vol.801, no.1, 2017.

[4] Corruption Watch, *What is Corruption?*, http://www.corruptionwatch.org.za/learn-about-corruption/what-is-corruption/our-definition-of-corruption/, 2017.

[5] M. Siddique and R. Ghosh, *Corruption, Good Governance and Economic Development: Contemporary Analysis and Case Studies*, World Scientific Publishing, Singapore, 2014.

[6] Transparency International, *Transparency International Plain Language Guide*, https://www.transparency.org/whatwedo/publication/the_anti_corruption_plain_language_guide, 2009.

[7] A. Shah and M. Schacter, Combinating corruption: Look before you leap, *Finance Development*, vol.4, pp.40-43, 2004.

[8] P. Mauro, *Why Worry about Corruption?*, International Monetary Fund, Washington, D.C., 1997.

[9] Y. G. Sucahyo, *Data Mining Menggali Informasi Yang Terpendam*, Artikel Populer IlmuKomputer.com, 2003.

[10] J. Han, J. Pei and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2011.

[11] J. A. Hoffer, M. Presco and H. Topi, *Modern Database Management*, Pearson, 2012.

[12] M. Scherenk, *Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL*, No Starch Press, Inc., San Fransisco, 2007.

[13] C. Carlos, *Effective Web Crawling*, Ph.D. Thesis, Department of Computer Science, University of Chile, 2004.

[14] D. Olson and D. Delen, *Advanced Data Mining Techniques*, Springer Berlin Heidelberg, Berlin, 2008.

[15] D. Xhemali, C. J. Hinde and R. G. Stone, Naïve Bayes vs. decision trees vs. neural networks in the classification of training web pages, *International Journal of Computer Science Issues*, vol.4, no.1, pp.16-23, 2009.
[16] F. Gorunescu, *Data Mining: Concepts, Models, and Techniques*, Springer, Berlin, 2011.
[17] W. Cavnar and J. Trenkle, N-Gram-based text categorization, *Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.
[18] T. H. Cormen, C. Leiserson, R. Rivest and C. Stein, *Introduction to Algorithms*, The MIT Press, Massachusetts, 2001.
[19] D. Malik, Dari konstruksi ke dekonstruksi: Refleksi atas pemberitaan televisi kita, *Jurnal ISKI, Pers Indonesia Era Transisi*, vol.4, 2001.