Corruption Cases Mapping Based on Indonesia's Corruption Perception Index

Noerlina[1], L A Wulandhari[2], Sasmoko[3,4], A M Muqsith[5], M Alamsyah[5]

[1]School of Information System, Bina Nusantara University,Indonesia
[2]School of Computer Science, Bina Nusantara University, Indonesia
[3]Faculty of Humanities, Bina Nusantara University,Indonesia
[4] Research Interest Group in Education Technology, Bina Nusantara University, Indonesia
[5]School of Computer Science, Bina Nusantara University, Indonesia

Abstract. Government plays an important role in nation economic growth. Nevertheless, there are still many occurrences of government officers abusing their offices to do an act of 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, this makes the news media relevant for being one of the sources of measurement of corruption perception index (CPI). It is required to map the corruption case in Indonesia so the central government can pay attention to every region in Indonesia. To develop the mapping system, researchers use Naïve Bayes Classifier to classify which news articles talk about corruption and which news articles are not, before implementing a Naïve Bayes Classifier there are some text processing such as tokenizing, stopping, and stemming.

1. Introduction

Government plays an important role in nation economic growth. Nevertheless, there are still many occurrences of government officers abusing their offices to do an act of corruption. Corruption becomes an undeniable disease and threatens the structure of society (Rahimi & Shakeri, 2016). Generally, corruption limits the governments' institution by disobeying procedure, use of resources, and office occupation without the legal procedure. Corruption also undermines the legitimacy of government and other democratic values (Ghosh & Siddique, 2014). As a national issue, corruption must be dealt with firm and clear with all the potential involvement that exists within the public especially the government and law enforcement officials. The concept of systemic corruption as a collective action has significant implications on policy implementation (Persson, Rothstein, & Teorell, 2013). The media can show the current conditions of the region.

Additionally, the media can also summarize thoughts and ideas of the public (Malik, 2001). Therefore, using the social media as source data to calculate Corruption Perception Index (CPI) is very possible. CPI measures the level of corruption and CPI mapping will help the government to analyze and take measures to prevent corruption in every region (Melgar, Rossi, & Smith, 2010; Rohwer, 2009). CPI may be a reference to map corruption in all regions in Indonesia; CPI mapping will help the government in the analysis and prevention of corruption in each region (Uca, Ince, & Sumen, 2016). Therefore, the researchers propose system development of corruption cases mapping based on Indonesia's Corruption Perception Index. To develop the mapping system, the researchers need to solve

the problem of how to classify news articles that talk about corruption case and news articles that do not talk about corruption case.

According to previous research, several algorithms can be used to classify the document including Naïve Bayes Classifier, Decision Trees, and Support Vector Machine, which is Naïve Bayes Classifier is the most accurate algorithm to classify the document (Huang, Lu, & Ling, 2003).

2. News Classification

2.1. Data Collection

The data collected in this research is a news content taken from some news site in Indonesia, there are some considerations made in the selection of news sites for data collection, the consideration include:

2.1.1. Alexa rank.
Alexa rank is a website that calculates the rank of a website based on traffic by counting unique visitors with page views. Alexa rank produces good indicator to determine the credibility of a website (Olteanu, Peshterliev, Liu, & Aberer, 2013)

2.1.2. News content that consistently uses 5W1H format
5W1H is an abbreviation of who, what, where, when, how and why. The information provided by an article will be very informative and clear if it can answer 5W1H (Jang & Woo, 2005)

2.1.3. Site has existed since 2010
To equalize the final result of the calculation system, the site at least has published articles before or at least since 2010.

2.1.4. Have sitemap feature that will be used for web crawling
Sitemap is a single page that displays a collection of structured links that directly connects the pages in a website, sitemap. With the sitemap, web crawling will easily index all the pages in a website (Yang et al., 2009)

2.1.5. Site content can be retrieved using a scraping technique.
There are several websites that prohibit a web crawler to do indexing on the website so that the data of news item could not be retrieved.

There are seven Indonesian news sites considered for data collection. The news sites are Tempo (tempo.co), Detik (detik.com), Kompas (kompas.com), Merdeka (merdeka.com), Liputan6 (liputan6.com), Tribun News (tribunnews.com), and Sindo News (sindonews.com).

There are three techniques for data collection:

- Web Crawling is a software that runs recursively crawl from one page to another page by following the link in each page (Schrenk, 2007), in this research, crawling result will be saved in the database as a news article index for web scraping purpose.
- Web Scraping is a software technique that gathering and extracting information from a web page (Schrenk, 2007), in this research, web scraping is used to get news content based on news article index and saved the content into the database.
- 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.

By using web crawling, web scraping, and cron scheduling, 396565 news articles successfully collected, here is the detail of data collected:

Table 1. *Data Collected.*

News Site	Article Collected
Detik	66826
Kompas	49396
Liputan 6	87950

Merdeka	54475
Sindo News	41016
Tempo	60763
Tribun News	36139
Total	396565

2.2. News Classification Algorithm

There are two processes performed in the process of classifying the news content; the processes are text processing and news content classification process.

2.2.1. Text Processing

Text processing is done to represent the text into the same form (Tated & Ghonge, 2015). Text processing has three sub-processes; tokenizing, stopping and stemming:

- Tokenizing is a process that splits the document into a single word; this process split the document by space character and saved into the array of word.
- Stopping is a searching process that searches a stop word in the array of word and removes the found word from the array of word, an Indonesian stop word list is listed from Fadillah Z Tala's research in 2003 (Tala, 2003).
- Stemming is a process that convert each word in the array of word into a root word for example "mencintai" into "cinta" or "memilikinya" into "milik", the process used Nazief and Andriani's algorithm, the algorithm is:
  Step 1, search the word in a root word database, if found then the word is a root word and algorithm is finished.
  Step 2, Inflection suffixes (-lah, -kah, -ku, -mu, or -nya) deletion, if a word is a particle (-lah, -kah, -tah, or -pun), then repeat this step to delete a possessive pronouns (-ku, -mu, or -nya) if exists.
  Step 3, Derivation suffixes (-i,-an,or -kan) deletion, after deletion occurred, search the result word in a root word database, if found, then the algorithm is finished, if not then continue to Step 3.1.
  Step 3.1, if -an was deleted and the last character was -k, then remove the -k, if the result found in root word database, then the algorithm is finished, if not then continue to Step 3.2.
  Step 3.2, restore deleted suffixes (-i, -an, or -kan), and continue into Step 4.
  Step 4, Derivation prefixes (di-, ke-, se-, me-, be-, pe-, or te-) deletion, with the maximum number of iterations is three times:
  Step 4.1, the algorithm finished if:
  - Not allowed prefixes and suffixes combination occurred based on this table (Table 2):

Table 2. *Not allowed prefixes and suffixes combination.*

Prefixes	Not allowed suffixes
be-	-i
di-	-an
ke-	-i, -kan
me-	-an
se-	-i, -kan

te-	-an

- Current detected prefixes equal to last deletion prefixes.
- Three prefixes deletion.

Step 4.2, identify the prefixes type and delete, there are two types of prefixes:

- Standard prefixes (di-, ke-, or se-)
- Complex prefixes (me-, be-, pe-, or te-), this type can create a new word according to the followed root word, so, there are rule that is used to get the right hyphenation (Table 3):

Table 3. Hyphenation rule.

No	Word Format	Hyphenation
1	berV…	ber-V… | be-rV…
2	berCAP…	ber-CAP… | where C!=r and P!=r
3	berCAerV…	ber-CaerV… where C!=r
4	belajar	bel-ajar
5	beC1erC2…	be-C1erC2… where C1 != {r|l}
6	terV…	ter-V… | te-rV…
7	terCerV…	ter-CerV … where C!=r
8	terCP…	ter-CP… where C!=r
9	teC1erC2…	te-C1erC2… where C1!=r
10	me{l|r|w|y}V…	me-{l|r|w|y} V…
11	mem{b|f|v}…	mem-{b|f|v}…
12	mempe{r|l}…	mem-pe…
13	mem{rV|V}…	me-m{rV|V}… | me-p{rV|V}…
14	men{c|d|j|z}…	men-{c|d|j|z}
15	menV…	me-nV… | me-fV
16	meng{g|h|q}…	meng-{g|h|q}…
17	mengV…	meng-V… | meng-kV…
18	menyV…	meny-sV…
19	mempV…	mem-pV… where V!=e
20	pe{w|y}V…	pe-{w|y}V…
21	perV…	per-V… | pe-rV…
22	perCAP…	per-CAP… where C!=r dan P!=er
23	perCAerV…	per-CAerV… where C!=r
24	pem{b|f|V}…	pem-{b|f|V}…
25	pem{rV|V}…	pe-m{rV|V}… | pe-p{rV|V}
26	pen{c|d|j|z}…	pen-{c|d|j|z}…
27	penV…	pe-nV… | pe-tV…
28	peng{g|h|q}…	peng-{g|h|q}
29	pengV…	peng-V… | peng-kV…
30	penyV…	peny-sV…
31	pelV…	pe-lV… exept ‘pelajar’
32	peCerV…	per-erV… where C!={r|w|y|l|m|n}
33	peCP…s	per-CP… where C!={r|w|y|l|m|n} and P!=er

- Search deleted prefixes result into a root word database, if not found then repeat the Step 4, if found, then the algorithm is finished.

Step 5, if word not found in root word database after Step 4, then repeat the Step 4, in this step, the algorithm is finished and assumed the word is a root word.

2.2.2. News Content Classification

2.2.2.1. Document Training

Documents that are trained in this process is the result of text processing, the document is 30 documents of corruption case, and 30 documents of the non-corruption case, each training result will be saved into a database, which will then be used for Naïve Bayes Classifier algorithm as a reference to an existing document.

2.2.2.2. Document Classification Algorithm

In this classification step, the used algorithm is Naïve Bayes Classifier algorithm, that uses a previous document to calculate the future probability using equation (Murphy, 2006):

$$V_{MAP} = \underset{V_j e\, V}{\arg max} \prod_{i=1}^{n} (P(x_i|V_j)P(V_j))$$

$data$ Word from text processing

x_i Data i $(1,2,3,\dots,n)$

V_j Class j $(1,2,3,\dots,n)$

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

$P(V_j)$ Probability V_j, where $P(V_j) = \frac{|documents\ j|}{|documents|}$

For $P(V_j)$ and $P(x_i|V_j)$ calculated while data training is being processed, where:

$|documents\ j|$ total documents in each class j

$|documents|$ total documents in all class

n_k total frequency of occurrence of each of data

n total frequency of data appearance of each class

$|data|$ total all data from all class

3. Result and Analysis

To produce good results of Naïve Bayes classifier algorithm, there are 30 news articles that talk about corruption case and 30 news articles that do not talk about corruption case, then do the test of 10 articles that talk about corruption case and 10 articles that do not talk about corruption case, the result of the algorithm is 10 articles that talk about corruption was accurate 100%, and 10 articles that do not talk about corruption is accurate 100%.

Can be concluded, the research method that already described above can produce the accurate results. At this stage, the research can classify news articles that talk about corruption case or talk about non-corruption case.

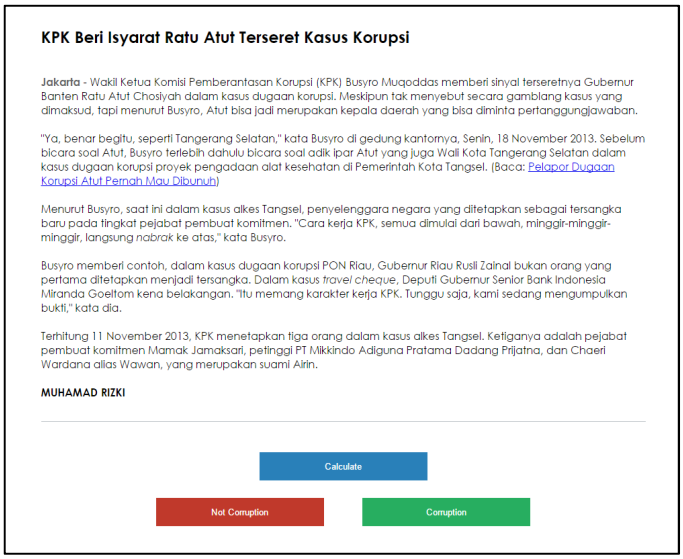

Figure 1. Training process of article that talks about corruption.

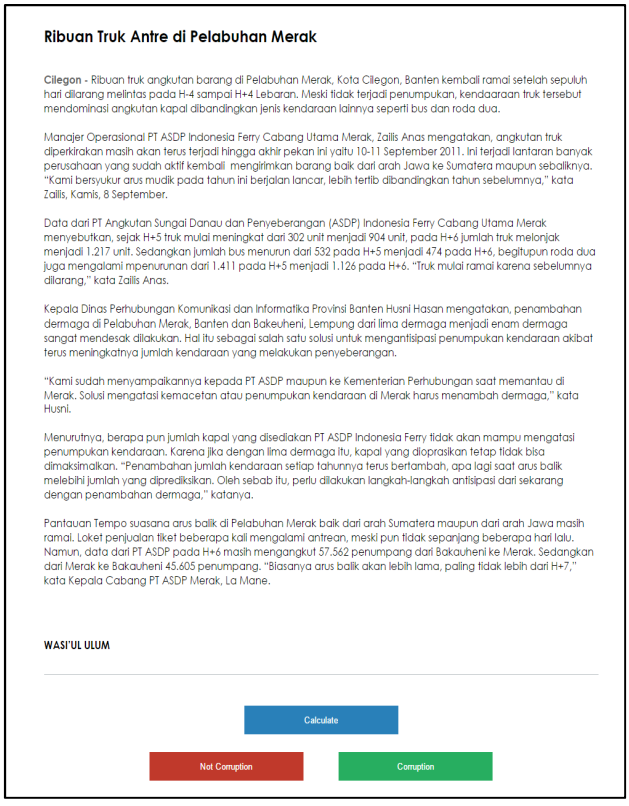

Figure 2. Training process of article that does not talk about corruption.

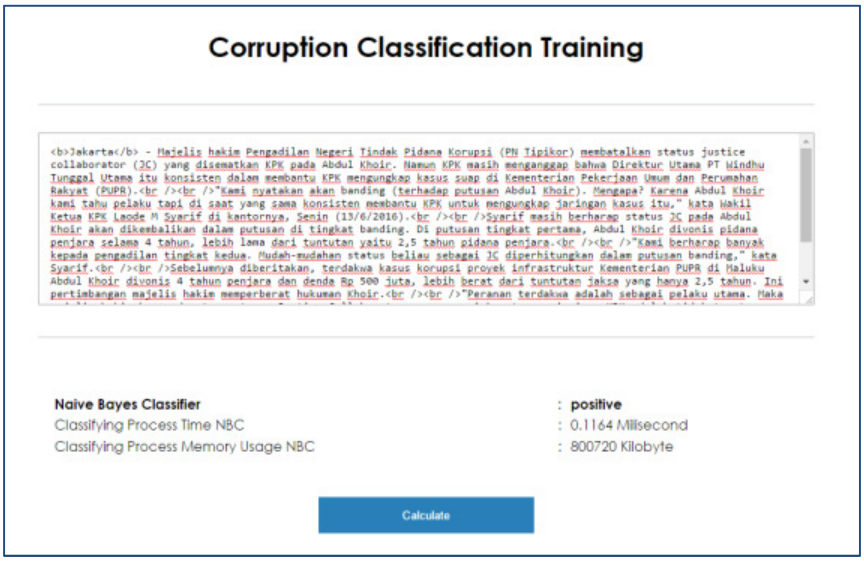

Figure 3. Testing process of article that talks about corruption.

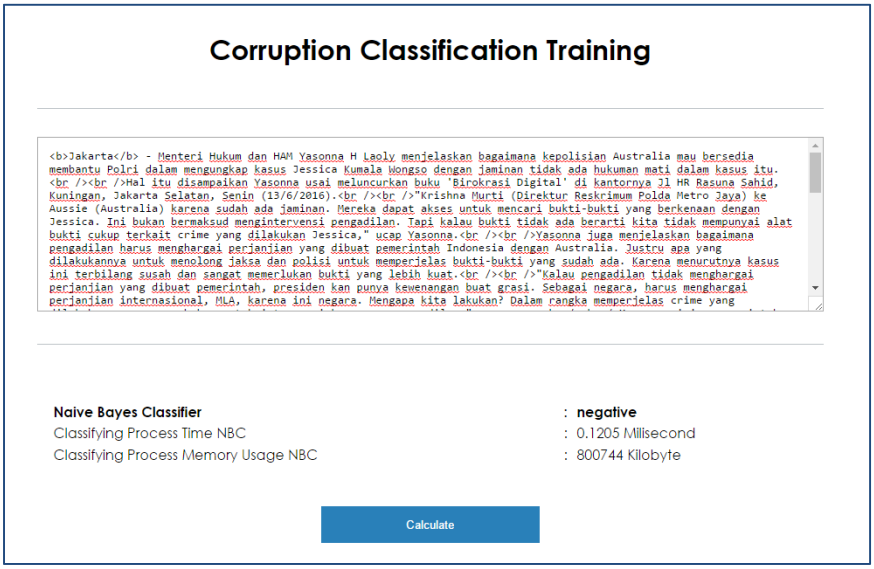

Figure 4. Testing process of article that does not talk about corruption.

From the process of text mining performed, resulting in the data on the mode of corruption based on the news phenomenon from online media related to corruption, which 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 scene, which 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] Ghosh, R., & Siddique, M. (2014). *Corruption, Good Governance and Economic Development: Contemporary Analysis and Case Studies.*

[2] Huang, J., Lu, J., & Ling, C. (2003). *Comparing naive Bayes, decision trees, and SVM with AUC and accuracy* .
[3] Jang, S., & Woo, W. (2005). 5W1H: Unified User-Centric Context. *Proceedings of the 7th International Conference on Ubiquitous Computing.*
[4] Malik, D. (2001). *Dari Konstruksi ke Dekonstruksi: Refleksi atas Pemberitaan Televisi Kita.*
[5] Melgar, N., Rossi, M., & Smith, T. W. (2010). The Perception of Corruption. *International Journal of Public Opinion Research*, *22*(1), 120–131.
[6] Murphy, K. P. (2006). *Naive Bayes Classifiers Generative Classifiers*, 4701**,** 1–8.
[7] Olteanu, A., Peshterliev, S., Liu, X., & Aberer, K. (2013). Web Credibility: Features Exploration and Credibility Prediction. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 557–568.
[8] Persson, A., Rothstein, B., & Teorell, J. (2013). Why Anti Corruption Reforms Fail — Systemic Corruption as a Collective Action Problem, *Governance*, *26*(3), 449–471.
[9] Rahimi, R., & Shakeri, H. (2016). Reflection on Judicial System's Corruption and Offering Solutions to Promote its Safety. *Journal of Politics and Law, 9*(9), 187.
[10] Rohwer, A. (2009). Measuring Corruption: A Comparison between the Transparency International's Corruption Perception Index and the World Bank's Worldwide Governance Indicators. *DICE*, 42–52.
[11] Schrenk, M. (2007). *Webbots, spiders, and screen scrapers.*
[12] Tala, F. (2003). *A study of stemming effects on information retrieval in Bahasa Indonesia*
[13] Tated, R. R., & Ghonge, M. M. (2015). A survey on text mining - techniques and application. *International Journal of Research in Advent Technology,*(1), 380–385.
[14] Uca, N., Ince, H., & Sumen, H. (2016). The Mediator Effect of Logistics Performance Index on the Relation between Corruption Perception Index and Foreign Trade Volume. *European Scientific Journal, 12*(25), 1857–7881.
[15] Yang, J. M., Cai, R., Wang, C., Huang, H., Zhang, L., & Ma, W. Y. (2009). Incorporating Site-Level Knowledge for Incremental Crawling of Web Forums: A List-Wise Strategy. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1375–1383.