Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management
Kuala Lumpur, Malaysia, March 8-10, 2016

Application of Text Mining for Classification of Textual Reports: A Study of Indonesia's National Complaint Handling System

Isti Surjandari
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Kampus UI, Depok 16424, Indonesia
isti@ie.ui.ac.id

Chyntia Megawati
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Kampus UI, Depok 16424, Indonesia
chyntia.megawati@ui.ac.id

Arian Dhini
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Kampus UI, Depok 16424, Indonesia
arian@ie.ui.ac.id

I. B. N. Sanditya Hardaya
Department of Industrial Engineering
Faculty of Engineering, Universitas Indonesia
Kampus UI, Depok 16424, Indonesia
ida.bagus23@ui.ac.id

***Abstract*— The rapid development of Information and Communication Technology (ICT) has made ICT an important part in the daily life of society. In that connection, the Indonesian government also tried to take advantage of ICT to be able to establish two-way communication with the public or commonly known as e-Government. One way is to create a website called LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat or National Complaint Handling System). All kind of reports that conveyed by public through LAPOR! could be important inputs for the government to develop and improve public services. The high number of reports makes manual analysis becomes ineffective so that big data analysis becomes important. This study uses Text Mining methods for analyzing textual data in the form of opinions or complaints submitted by the public through LAPOR! by classifying those reports into classes. Then the data set in each class was clustered into specific topics. The results of this study show that the majority of public report is associated with poverty, particularly regarding social assistance, such as KPS (Kartu Perlindungan Sosial or Social Security Card) and BLSM (Bantuan Langsung Sementara Masyarakat or Temporary Direct Cash Assistance), which were not well distributed or not on target.**

Keywords*— *Text Mining, Classification, Clustering, Support Vector Machine, Self-Organizing Maps, Public's Reports

I. INTRODUCTION

The rapid development of Information and Communication Technology (ICT) has made ICT an important part in the daily life of today's society. Many sectors use Internet for various purposes. The Indonesian government also tried to take advantage of ICT to be able to establish a two-way communication with the public or commonly known as e-Government. One way is to create a website called LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat or Indonesia's National Complaint Handling System). LAPOR! is managed by Kantor Staf Presiden (Presidential Staff Office). It is an initiative from

the Government to provide an integrated and accessible portal for public to submit complaints and inquiries as a means of enhancing public participation in national development programs and support the creation of open government in Indonesia. It is built as social platform, which enable interaction between the people and the Government. This service can be used to convey the aspirations of the Indonesian people and their complaints through several ways, namely Short Message Service (SMS 1708), www.lapor.ukp.go.id sites, mobile apps Android and Blackberry, as well as social media Twitter and Facebook @ LAPOR1708 REPORT. The homepage of LAPOR! website is shown in Fig 1.

As of March 2015, LAPOR! has received more than 610,000 reports, with an average of more than 900 reports each day. The reports were received from various regions in Indonesia and around 80-90% were reported via mobile messaging. However, LAPOR! has not been able to explore and analyze the whole incoming reports because of the massive number of data and limited resources. From a total of 610,000 data received, only 75,870 reports (12.5%) were directly approved, 6600 reports (1%) are pending, and 527,530 reports (86%) were archived. This indicates that only about 13-14% reports were processed, while around 86% reports remain unknown subjects. Therefore, LAPOR! currently still not able to identify national problems in real time based on complaints and reports coming from the public.

To date, the analysis of public's aspiration and complaints delivered via LAPOR! was done manually. The high number of reports makes manual analysis becomes ineffective, so that big data analysis becomes important. Therefore, it is necessary for the classification and grouping of public's reports that goes to LAPOR! with the help of Text Mining to determine the patterns of problems that occur in the society, so that the Government can quickly identify and respond to these problems. Text Mining is part of Data Mining that is used to analyze textual data (i.e., semi-structured and unstructured data).

The purpose of this study is to analyze public reports received through LAPOR!, so that the Government may identify the patterns of problems that occur in the society. Specifically, this study aims to establish a classification model and perform clustering of documents that go into LAPOR! using Text Mining, that is, by using Support Vector Machine (SVM) for classification and Self Organizing Map (SOM) to map the cluster of each classification class that has been formed.

II. Literature Reviews

Text mining or text analytics is a term that describes a technology that can analyze semi-structured or unstructured text data. The data structure distinguishes it from data mining, where data mining process structured data. Basically, text mining is an interdisciplinary field that refers to information retrieval, data mining, machine learning, statistical, and computational linguistics [1]. Text mining will extract meaningful numerical index of the text, and then the information contained in the text will be accessible by using various data mining algorithms [2].

Fig. 1. LAPOR! social platform

Text mining can analyze documents, classify documents based on the words contained in it, as well as determine the similarity between documents to know how they relate with other variables [3]. The most common applications of text mining are filtering spam documents, analyzing sentiment, measuring customer preferences, summarizing documents, grouping research topics, and many others. Basili, Cammisa and Moschitti proposed kernel functions to use prior knowledge in learning algorithms for document classification [4]. Cross-validation results showed the benefit of the approach for Support Vector Machines when few training examples are available [4]. Suh, Park and Jeon applied text mining techniques to unstructured data of petitions to elicit keywords from petitions and identify groups of petitions with the elicit keywords by *k*-means clustering method after determining the size of clusters by Self-Organizing Maps [5].

To obtain the ultimate goal of text mining, several stages of process are required as shown in Fig. 2 [6]. The selected data to be analyzed will first pass through the stage of pre-processing and transformation processes, until it can attain the knowledge discovery.

The main purpose of pre-processing stage is to get the data form ready for further processing. Operation process of learning algorithm could not directly execute text document in its original form. Therefore, after the pre-processing stage, the document is converted into a more manageable representation. Typically, the document will be represented by a vector [5]. Vector model is built from the document by changing tokens (pieces words) in the document into a vector which will be operated by numerical linear algebra operation [7]. In order to build the vector model, it is necessary to carry on weighting scheme. The most widely used weighting scheme is term frequency-inverse document frequency (TF-IDF). On the assumption of TF-IDF reverse weighting; words with high TF will get large weight, unless the number of documents that contain the words is large.

After passing the TF-IDF scheme, result in the form of matrix will be obtained. The matrix is represents documents in row and the separated token in column. Although it already has a form that appropriate and able to be processed using the learning algorithm, the matrix still has very high dimension. Singular Value Decomposition (SVD) is used to reduce the dimensions of the matrix by finding hidden correlations and structure of the matrix [2].

The final stage of the extracting information process in text mining is knowledge discovery. In this study, information extraction is done by classifying and clustering documents. The classification process was done with Support Vector Machine (SVM) algorithm. SVM is a classification algorithm that has purpose to find a separator function (i.e., hyperplane) with the greatest margin, so that separate two data sets optimally [1]. SVM was originally used for the classification of numerical data, but it turns out that SVM is also very effective to resolve text data problems.

The classification model needs to be evaluated to determine how well the model performs the desired classification. The accuracy value of the model will represent how the overall document classified correctly. The higher accuracy value means the model is better and more accurate in doing classification. Furthermore, the clustering process was done with the Self Organizing Map (SOM) algorithm. Kohonen Self-Organizing Maps or Self-Organizing Maps (SOM) is a type of neural network models. SOM allows visualization and projection of high-dimensional data to a lower dimension, most often as 2-D plane, while maintaining the data topology [7].

III. Research Methodology

For the purpose of this study, the data used is all the public reports that go into LAPOR! for the period October 2014 – March 2015. The data received has undergone a screening process where a report or a message that contain threats, verbal abuse, racial, and pornography are not included or considered as a spam message. The flowchart of research methodology is shown in Fig. 3.

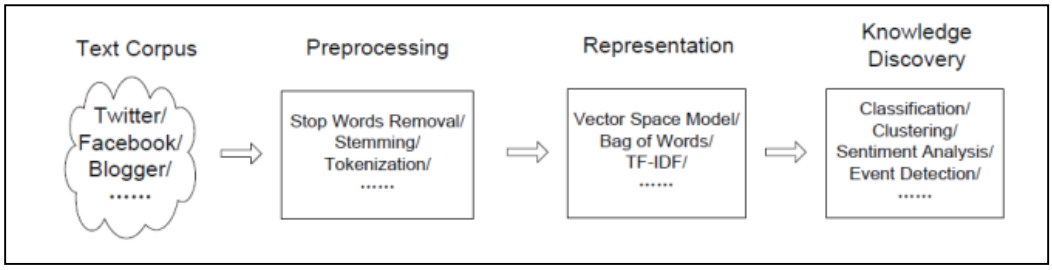

Fig. 2. Text mining analysis process scheme

Text preprocessing phase is carried out to transform the textual data that is unstructured into structured. There are several steps being taken in the text preprocessing. In general, the stages in the preprocessing text is divided into two main parts, the first is tokenization, case folding, normalization and filtering; while the second is stemming. For the purpose of this study, the data set will be processed with and without stemming proess.

- Tokenization is a process of separating documents into pieces of words known as token.
- Case Folding is a process of changing all uppercases to lowercases in a document and vice versa. This step is also to ensure that only letters a through z are contained in the document.
- Normalization is a process of substituting misspelled or abbreviated words. For example, the word "tidak" has some other forms, like "tdk", "gak", "nggak", "enggak" (which means no in English).
- Filtering is a process of eliminating of mention (@), hashtag (#), url, or punctuation marks (emoticon).
- Stemming, is a process of taking the stem. For example, the word "using", "uses", and "used" will have the same stem as "use".
- *Matrix*, which is a vector representation of word tokens based on the occurrences of words in the document [8]. In general, there will be three different matrix generated, i.e., the term frequency (tf), inverse document frequency (idf), and singular value decomposition (SVD).

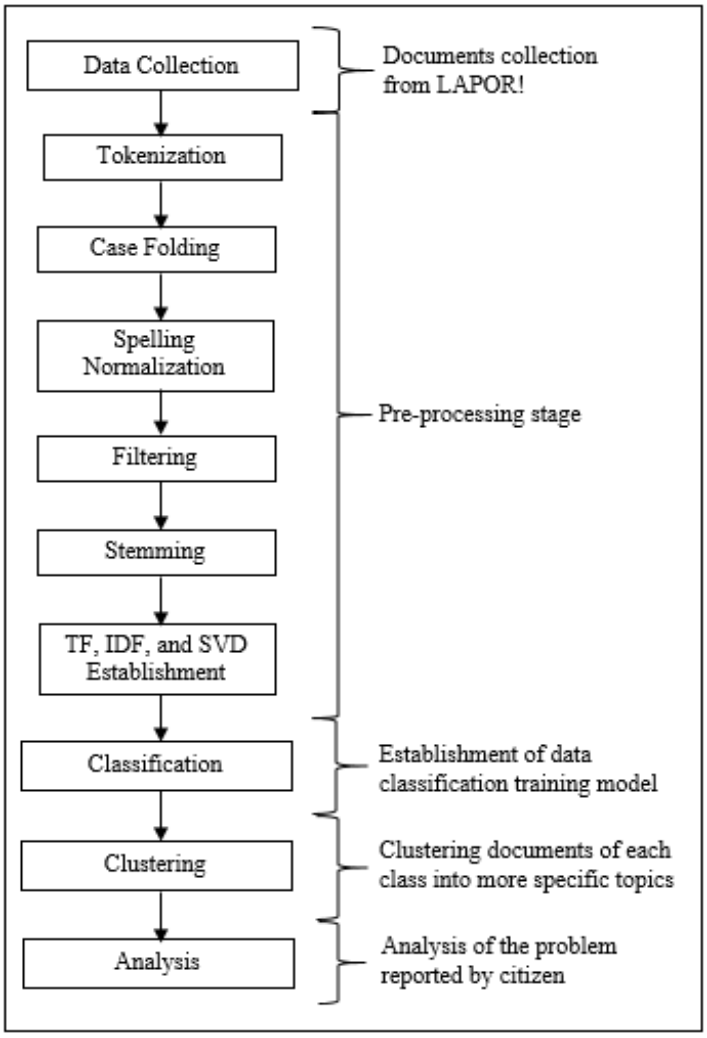

Fig. 3. Flowchart of research methodology

In the process of term frequency, token will be weighed according to the number of its emersion in a document. Then, the term frequency matrix, as shown in Fig. 4, will be obtained. After going through the process of term frequency, the procedure of inverse document frequency is done by weighing the words based on its emersion frequency in the entire documents. The inverse document frequency matrix is shown in Fig. 5. This procedure will produce a high dimension of *term frequency-inverse document frequency* (TF-IDF). To that end, *Singular Value Decomposition* (SVD) is used to reduce the number of variables by transforming correlated variables into a set of uncorrelated variables, which will reveal relationships contained in the original data. These uncorrelated variables will later be referred as *concepts* [9]. As shown in Fig. 6, the matrix is no longer contains all word entities, but it only contains some concepts that represent relationships between entities in the document.

	1 Var1	2 account	3 ada	4 adalah	5 agar	6 air	7 aju	8 akan	9 anak	10 anggota	11 antara	12 apa
1	yth erintah provinsi dk jakarta saya ingin lapor bahwa honor kami baga											1
2	rumah tangga saya hanya erima kip padahal ada anak lajar di rumah s		1						1			
3	subsidi dari pmrinth untuk siswa sd ca das rtajaya camat telagasar ka											
4	yth menteri didi dan budaya rujuk pada buku tunjuk teknis tentang bar											
5	keluarga saya belum erima kip hasan ali untuk anak anak saya yang								2			
6	kami dapat kps tetapi anak saya kelas smp tidak dapat bsm padahal						1		1			
7	saya erima kps dan saya punya anak sekolah smp namun saya tidak								1			
8	asalamualaikum yang hormat bapak ibu yang kerja di kps mohon jela								1			
9	rumah tangga di desa saya tidak erima klp untuk siswa kurang mamp											
10	kang emil kenapa bantu kenapa bantu sekolah di kota bandung di hap											
11	yth menteri didi smk muhammadiyah simo boyolalialamat jl madu ngr											
12	anak saya tidak ada yang dapat kip		1						1			
13	kepada yang hormat prov dk jakarta kami guru guru dk jakarta sampai											1
14	saya tidak erima kartu indonesia pintar kip untuk dapat bantu siswa k								1			
15	anak saya tidak dapat kip wahyud yang nama arida lutfian tidak dapat								1			

Fig. 4. Term frequency matrix

	1 Var1	2 account	3 ada	4 adalah	5 agar	6 air	7 aju	8 akan	9 anak	10 anggota	11 antara	12 apa
1	yth erintah provinsi dk jakarta saya ingin lapor bahwa honor kami bagai guru bantu di											2.50899
2	rumah tangga saya hanya erima kip padahal ada anak lajar di rumah saya mohon di u		1.43443						2.07647			
3	subsidi dari pmrinth untuk siswa sd ca das rtajaya camat telagasar kab ka rawang jat											
4	yth menteri didi dan budaya rujuk pada buku tunjuk teknis tentang bantu embang smk											
5	keluarga saya belum erima kip hasan ali untuk anak anak saya yang sekolah muham								3.51577			
6	kami dapat kps tetapi anak saya kelas smp tidak dapat bsm padahal sudah aju syara						3.73826		2.07647			
7	saya erima kps dan saya punya anak sekolah smp namun saya tidak erima kip sedar								2.07647			
8	asalamualaikum yang hormat bapak ibu yang kerja di kps mohon jelas berapa bulan s								2.07647			
9	rumah tangga di desa saya tidak erima klp untuk siswa kurang mampu erima kks kks											
10	kang emil kenapa bantu kenapa bantu sekolah di kota bandung di hapus											
11	yth menteri didi smk muhammadiyah simo boyolalialamat jl madu ngren simo boyolal											
12	anak saya tidak ada yang dapat kip		1.43443						2.07647			
13	kepada yang hormat prov dk jakarta kami guru guru dk jakarta sampai saat ini ingin ta											2.50899
14	saya tidak erima kartu indonesia pintar kip untuk dapat bantu siswa kurang mampu b								2.07647			
15	anak saya tidak dapat kip wahyud yang nama arida lutfian tidak dapat kip kartu indon								2.07647			
16	di rumah tangga saya hanya erima kip padahal ada anak yang masih seko		1.43443						2.07647			
17	maaf mau tanya kenapa bsm di sekolah kami sampai sekarang belum cair dang seko											
18	anak anak saya sekolah semua sd smp smk tetapi ko keluarga saya tidak erima kip								3.51577			
19	rumah tangga di desa saya tidak erima kartu kip dang saya punya anak yang masih s								2.07647			

Fig. 5. Inverse document frequency matrix

	1 Var1	2 Concept1	3 Concept2	4 Concept3	5 Concept4	6 Concept5	7 Concept6	8 Concept7	9 Concept8
1	yth erintah provinsi dk jakarta saya ingin lapor bahwa honor kami bagai guru bantu di smp yapi	0.01731	-0.01271	-0.00253	-0.00153	0.00134	-0.00537	0.01031	0.00006
2	rumah tangga saya hanya erima kip padahal ada anak lajar di rumah saya mohon di urus	0.00550	-0.00259	0.01393	0.01187	-0.00810	0.00167	0.01764	0.00627
3	subsidi dari pmrinth untuk siswa sd ca das rtajaya camat telagasar kab ka rawang jabar dian t	0.01171	-0.00441	0.01337	0.00354	0.00294	0.00421	0.00345	-0.00434
4	yth menteri didi dan budaya rujuk pada buku tunjuk teknis tentang bantu embang smk ruju non	0.01259	0.00127	0.00276	-0.01587	-0.00012	0.00069	-0.00228	0.01774
5	keluarga saya belum erima kip hasan ali untuk anak anak saya yang sekolah muhamad syahri	0.00453	-0.00041	0.01389	0.01217	-0.01368	0.00147	0.02171	0.00739
6	kami dapat kps tetapi anak saya kelas smp tidak dapat bsm padahal sudah aju syarat di smp	0.01195	-0.00333	0.02157	0.01262	-0.00629	-0.00229	0.01835	-0.00919
7	saya erima kps dan saya punya anak sekolah smp namun saya tidak erima kip sedang kan er	0.00749	-0.00217	0.02487	0.01891	-0.01418	0.00463	0.02639	0.00388
8	asalamualaikum yang hormat bapak ibu yang kerja di kps mohon jelas berapa bulan sekali bar	0.00822	-0.00421	0.00988	0.00692	-0.00158	0.00030	0.00932	-0.00637
9	rumah tangga di desa saya tidak erima klp untuk siswa kurang mampu erima kks kks rumah t	0.00813	-0.00516	0.02613	0.01946	0.00263	0.01243	0.00833	0.01346
10	kang emil kenapa bantu kenapa bantu sekolah di kota bandung di hapus	0.00348	-0.00185	0.00764	0.00465	-0.00013	0.00015	0.00527	-0.00252
11	yth menteri didi smk muhammadiyah simo boyolalialamat jl madu ngren simo boyolal kode pos	0.00525	0.00120	-0.00035	-0.00868	-0.00240	0.00149	0.00159	0.00635
12	anak saya tidak ada yang dapat kip	0.00282	-0.00088	0.00732	0.00686	-0.00564	0.00196	0.01084	0.00234
13	kepada yang hormat prov dk jakarta kami guru guru dk jakarta sampai saat ini ingin ta apa san	0.01508	-0.01248	-0.00961	-0.00008	-0.00753	-0.00023	0.00740	0.00189
14	saya tidak erima kartu indonesia pintar kip untuk dapat bantu siswa kurang mampu bsm saya	0.00836	-0.00321	0.02898	0.02144	-0.02253	0.00163	0.03152	0.00506
15	anak saya tidak dapat kip wahyud yang nama arida lutfian tidak dapat kip kartu indonesia pint	0.00501	-0.00028	0.01572	0.01284	-0.01684	-0.00101	0.01763	0.00511

Fig. 6. Singular value decomposition matrix

The classification process is done to classify the report data into some specified categories or classes. The purpose of this process is to build a predictive model that able to classify documents automatically to some known classes effectively and efficiently. In this study, SVM (Support Vector Machine) was used to establish a classification model. The classification process will divide the data into six classes that are: equitable access to education; public health; energy, food, and maritime; poverty alleviation; infrastructure development; and bureaucracy reformation.

SVM algorithm is a technique for regression and classification. SVM is geometrically described as a hyperplane that separates document into groups of data. Hyperplane chosen is the one with maximum distance between the hyperplane and the groups' nearest point. SVM algorithm will search for the optimum function (i.e., the hyperplane) to separate those two sets of data [6]. Hyperplane with the maximum margin of SVM is shown in Fig. 7.

The clustering process is performed on the six classes produced from the previous classification process. At each class, documents that are members of the class will be grouped into several topics. Clustering is done by using SOM (Self Organizing Map). Document members of each class will be divided into several groups or clusters. The number of cluster is determined by the initial map size of SOM. The determination of the initial map size is done by trial and error to get the size with smallest error numbers.

Self Organizing Map (SOM) is a method of clustering and visualization that has been applied in many research disciplines, ranging from engineering science, financial, social, natural sciences, and linguistics. SOM is a type of artificial neural network that is useful for performing cluster analysis visually [10]. SOM is able to classify objects with high dimension attribute to a lower-dimensional space, usually one or two dimensions [11]. Furthermore, because of its unsupervised learning algorithm, SOM is suitable at the stage of exploring the data.

IV. Results and Discussion

A. Classification Result

The accuracy model in Table I shows that the classification model with stemming has greater accuracy than that of without stemming. Hence the classification model with stemming data set was chosen to classify other data set that has not been classified. The classification result is shown in Table II.

From the classification result; it appears that majority of public reports is on poverty issues, covering conditions of poor people and various social assistance. The second largest class is infrastructure, where the reported problems is generally about road conditions, public transportation, and ongoing of government construction projects. While the least is public reports regarding energy, food and maritime issues. At this class, people generally report on fuel prices, food prices, and maritime conditions.

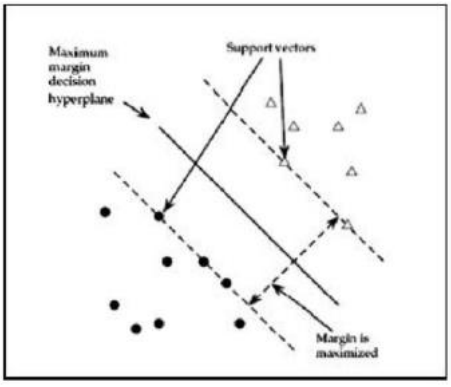

Fig. 7. SVM hyperplane

TABLE I. Accuracy of Classification Model

Data Set	Accuracy
Non-stemming	53%
Stemming	66%

TABLE II. NUMBER OF MEMBERS OF EACH CLASS

Class	Total
Equitable access to education	4195
Energy, food, and maritime	2310
Public health	3795
Infrastructure development	5077
Poverty alleviation	8347
Bureaucracy reformation	3611

B. Clustering Result

In the clustering stage, the process is performed unsupervised. Clustering is performed to obtain a specific reports topic of each class. In principle, the number of clusters obtained can be very subjective and depends on the needs of research. However, the process of searching for the optimal number of clusters is often done automatically if no restrictions or specific requirement is predetermined. For the purpose of this study, the maximum number of clusters is nine for each class. This is because LAPOR! want to be able to identify priority issues that exist in the society based on public reports. If the number of clusters formed were too large, then the number of members of each cluster would be too little, so it is difficult to provide an overview of the main problems that exist in the society. In the process of trial and error of determining the number of clusters, the results of each class show that the smallest error value is at the maximum number of clusters (i.e., 9 clusters). The clustering results are shown in Table III to VIII.

TABLE III. CLUSTER OF EDUCATION CLASS

Cluster	Cluster title	Total
Cluster 1	Reports about KJP (Kartu Jakarta Pintar)	152
Cluster 2	Complaints from public who did not get KIP (Kartu Indonesia Pintar). Majority of the reports come from outside Jakarta.	375
Cluster 3	Complaints about citizens who did not get KIP (Kartu Indonesia Pintar), but already got KIS (Kartu Indonesia Sehat) and KKS (Kartu Keluarga Sejahtera)	44
Cluster 4	Reports and questions about how to get KIP	203
Cluster 5	Report to various ministries (especially Ministry of Education and Ministry of Research and Technology) on issues about education, in terms of teachers or students	234
Cluster 6	Brief complaint from those who have not received KIP (no further details)	766
Cluster 7	Reports from KKS holders who did not received educational aids	332
Cluster 8	Complaints of less fortunate students about KIP fund distribution in school	217
Cluster 9	Complaints of household whose children have not received KIP	297

TABLE IV. CLUSTER OF ENERGY, FOOD, AND MARITIME CLASS

Cluster	Cluster title	Total
Cluster 1	Reports on the scarcity of LPG and soaring gas prices at particular area (addressed to Pertamina)	75
Cluster 2	Reports on the scarcity of fuel (addressed to Pertamina)	57
Cluster 3	Reports on the power outage and the installation of new power generation (addressed to PLN)	81
Cluster 4	Complains about the poor quality of Raskin (Beras Miskin or rice for the poor)	53
Cluster 5	Reports to the Ministry of Maritime and Fisheries Affairs, especially about illegal fishing	31
Cluster 6	Reports about PT PLN (state electricity company) services in general	168
Cluster 7	Fuel subsidy	185
Cluster 8	Reports to PDAM (regional water supply company) about water supply problems	28
Cluster 9	Protests about the soaring fuel price	52

TABLE V. CLUSTER OF HEALTH CLASS

Cluster	Cluster title	Total
Cluster 1	Reports on the difficulties of online registration for BPJS healthcare	103
Cluster 2	Complaints about payment procedures of premium/contribution/bill of BPJS healthcare	109
Cluster 3	Complaints about payment integration with online BPJS	164
Cluster 4	Public reports who have not received KIS (specifically aimed to the Ministry for Human and Cultural Development)	23
Cluster 5	Complaints about health care to the Jakarta provincial government, especially regarding the poor service at hospitals or health centers	58
Cluster 6	Problems on BPJS online registration account activation	466
Cluster 7	Reports from village household (family) who has not received KIS	148
Cluster 8	Reports about health aids in general	421
Cluster 9	Questions about BPJS healthcare	302

TABLE VI. CLUSTER OF INFRASTRUCTURE CLASS

Cluster	Cluster title	Total
Cluster 1	Reports about Transjakarta busses	104
Cluster 2	Water canal and flood	251
Cluster 3	All forms of construction projects in Jakarta	190
Cluster 4	Public works related to roads, bridges, sidewalks, etc.	101
Cluster 5	Reports related to the DKI Jakarta provincial roads and sidewalks	421
Cluster 6	Report to the provincial government of DKI Jakarta about traffic signs	125
Cluster 7	Transportation, traffics, stations, bus stops, trains, buses	660
Cluster 8	Reports to Jakarta government about illegal settlements and buildings	93
Cluster 9	Public works related to the development of underdeveloped villages	105

TABLE VII. CLUSTER OF POVERTY CLASS

Cluster	Cluster title	Total
Cluster 1	The uneven and misdirected distribution of KPS	424
Cluster 2	KPS holders who did not get educational aid	219
Cluster 3	Problems about BLSM where some villages did not get it	608
Cluster 4	Questions about cards procedures	338
Cluster 5	Problems about KPS in general	621
Cluster 6	Problems in the distribution of BLSM	800
Cluster 7	Problems concerning card recipient, where some people did not receive it	351
Cluster 8	Problems about Raskin (beras miskin or rice for the poor)	301
Cluster 9	Problems related to Government asisstance, such as KKS (kartu keluarga sejahtera), PKH (program keluarga harapan, BLT (bantuan langsung tunai), raskin, and others	1119

TABLE VIII. CLUSTER OF BUREAUCRACY CLASS

Cluster	Cluster title	Total
Cluster 1	Problems about process of making identity card	164
Cluster 2	Problems concerning the process of making land certificate	90
Cluster 3	Reports concerning the procedure and test to become civil servant	186
Cluster 4	Questions about the procedure of BLSM	85
Cluster 5	Problems concerning the procedure of making NPWP (taxpayer registration number)	151
Cluster 6	Reports related to bureaucratic problems in the provincial government of DKI Jakarta	337
Cluster 7	Questions about KPS card procedures	170
Cluster 8	Procedures to get birth certificates, driving license, building permit, STNK (vehicle document), marriage records, passports, immigration, etc.	1142
Cluster 9	Bureaucratic problems related to get card aids	82

Education class in Table III shows that the majority of the clusters discuss about KIP (Kartu Indonesia Pintar or Indonesia Smart Card), which is issued by the Government to help with the education costs for less fortunate students. Only cluster 5 that contain issues regarding teaching and learning process in the school (e.g., students, teachers, and learning problems).

Table IV shows the energy, food, and maritime class. In the energy, food, and maritime class, cluster 7 has the most number of members, where the majority of people reported problems in fuel subsidies, including complaints against the new policy on fuel subsidy and its impact on society. Cluster 6 has the second highest number, where people often complain about the service performance of PLN (Perusahaan Listrik Negara or State Company of Electricity) in general.

Results obtained from health class in Table V shows that the majority of the clusters discuss about BPJS Kesehatan (Badan Penyelenggara Jaminan Sosial Kesehatan or Healthcare Insurance Program) and KIS (Kartu Indonesia Sehat or Indonesia Health Card). BPJS healthcare and KIS is Government assistance to facilitate the poor in health services. Cluster 6 has the highest number of members which contains public complaints related to the activation process of BPJS healthcare. For example, the activation process that always fails or there is no confirmation from the BPJS.

Infrastructure class in Table VI indicates that transportation is the most widely reported topic. Cluster 7 has the highest number of members. On this cluster, people reported about the poor service quality of public transportation (e.g., trains, buses) including their supporting facility conditions (e.g., roads, bus stops, traffic signs). Cluster with the second highest number is cluster 5, which reported damaged roads and inadequate sidewalks in Jakarta.

Based on Table VII, in the poverty class, the entire cluster discusses about government assistance. Cluster 9 has the most number of members and the majority of the report is on KPS (Kartu Perlindungan Sosial or Social Security Card), KKS (Kartu Keluarga Sejahtera or Prosperous Family Card), PKH (Program Keluarga Harapan or Conditional Cash Transfer) and BLT (Bantuan langsung Tunai or Direct Cash Assistance). KPS, KKS, PKH, and BLT are government assistance to the poor, but each has a different content and rules. The majority of public complaints about the report contain: they do not receive government assistance, the procedures to become participants, aid distribution, or the occurrence of irregularities in the implementation of the assistance.

Table VIII shows the bureaucratic class. In the bureaucratic class, cluster 8 has the most number of members. Cluster 8 discusses the various complaints and questions in the application process of certificates or letters (e.g., birth certificates, driving licenses, land certificates), where public reported that the process is long, complicated, and there are many illegal levies.

V. Conclusion

All forms of inputs and aspirations from the public is essential for national development in many ways. Therefore, the Indonesian government built LAPOR! (Layanan Aspirasi dan Pengaduan Online Rakyat or Indonesia's National Complaint Handling System) as a social platform to convey any community complaints. On the other hand, the government is expected to provide rapid response and take appropriate decisions. With the large number of reports received in LAPOR!, the manual analysis would require considerable time and become unresponsive. Data analysis techniques, such as data mining and text mining are needed to automate the manual data analysis, so the long process become shorter.

This study is based on public reports through LAPOR! during the period of October 2014 to March 2015. This period was chosen because October 2014 marks the beginning of the new Presidential era. The classification model indicates the accuracy level of 66%. This accuracy level could still be improved by adding data, that is, by extending the period of data collection.

The results of this study found that poverty reduction and infrastructure development were the problems most frequently reported by the public through LAPOR! during the period of October 2014 to March 2015, which indicates that both issues should receive primary attention by the Government. Especially regarding BLSM (Bantuan Langsung Sementara Masyarakat or Temporary Direct Cash Assistance) and some other types of government assistance, where people have complained that those aids were not well distributed or not on target. While in infrastructure class, people reported about the poor service quality of public transportation (e.g., trains and buses), as well as their supporting facilities (e.g., rail stations, bus stops).

In accordance with the main purpose of LAPOR! that is to provide an integrated and accessible portal for public to submit complaints and inquiries as a means of enhancing public participation in national development programs and support the creation of open government in Indonesia, the model developed from this study can be used to assist the Government in identifying issues that occur in the society. Hence the Government may determine the priority issues that need to be resolved.

References

[1] H. Jiawei, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed, Waltham, MA: Morgan Kaufmann, 2012.

[2] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill and R. Nisbet, Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. Oxford: Elsevier. 2012.

[3] Statsoft, Text Mining Introductory Overview, retrieved April 19, 2015, from Statsoft: http://www.statsoft.com/textbook/text-mining, 2015.

[4] R. Basili, M. Cammisa and A. Moschitti, "A semantic kernel to classify texts with very few training examples", in Informatica, 2006, pp. 163-172.

[5] J. H. Suh, C. H. Park, S. H. Jeon, "Applying text and data mining techniques to forecasting the trend of petitions filed to e-People", in Expert Systems with Applications, 2010, pp. 7255-7268.

[6] C. Zhai and C. C. Aggarwal, Mining Text Data, New York: Springer, 2012.

[7] R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[8] K. Baker, Singular Value Decomposition Tutorial, *Ohio State University, Departement of Linguistics*, retrieved April 21, 2015, from Ohio State University web: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf.

[9] I. Surjandari, M. S. Naffisah, I. Prawiradinata, "Text mining of Twitter data for public sentiment analysis of staple foods price changes", Journal of Industrial and Intelligent Information, vol. 3 , no. 3, 2015, pp. 253-257.

[10] T. Kohonen, "Essentials of the self-organizing map", Neural Networks, 2013, pp. 52-65.

[11] T. Kohonen, "The self organizing map", Proceedings of the IEEE, 1990, pp. 1464 – 1480.

BIOGRAPHY

Isti Surjandari is a Professor and Head of Statistics and Quality Engineering Laboratory in the Department of Industrial Engineering, Faculty of Engineering, Universitas Indonesia. She holds a bachelor degree in Industrial Engineering from Universitas Indonesia and a Ph.D. degree from the Ohio State University. Her areas of interest are industrial management, quality engineering, data mining and applied statistics. She is a senior member of American Society for Quality (ASQ) and also ASQ country counselor for Indonesia. She has a vast experience in manufacturing and service industries.

Chyntia Megawati is a research associate in the Statistics and Quality Engineering Laboratory, Department of Industrial Engineering, Universitas Indonesia. She received her bachelor in Industrial Engineering from Universitas Indonesia.

Arian Dhini is a lecturer in the Department of Industrial Engineering, Faculty of Engineering, Universitas Indonesia. She obtained her bachelor degree in Industrial Engineering from Bandung Institute of Technology and master degree from Universitas Indonesia. Her interests include quality engineering and data mining. She is currently pursuing her doctoral degree in the field of data mining in the Department of Industrial Engineering, Universitas Indonesia.

I. B. N. Sanditya Hardaya is a research associate in the Statistics and Quality Engineering Laboratory, Department of Industrial Engineering, Universitas Indonesia. He is currently pursuing his bachelor degree in the Department of Industrial Engineering, Universitas Indonesia.