Data Mining Techniques of Complaint Reports for E-government: A Systematic Literature Review

Evaristus Didik Madyatmadja
Information Systems Department
School of Information Systems
Bina Nusantara University
Jakarta, Indonesia 11480
emadyatmadja@binus.edu

Debri Pristinella
Faculty of Psychology
Atma Jaya Catholic University of Indonesia
Jakarta, Indonesia
debri.pristinella@atmajaya.ac.id

Martinus Damitutsa Kurnia Dewa
School of Mechanical and Materials Engineering,
Washington State University
Washington State, USA
martinus.dewa@wsu.edu

Hendro Nindito
Information Systems Department,
School of Information Systems
Bina Nusantara University
Jakarta, Indonesia 11480
hendro.nindito@binus.ac.id

Cristofer Wijaya
Information Systems Department,
School of Information Systems
Bina Nusantara University
Jakarta, Indonesia 11480
cristofer.wijaya@binus.ac.id

***Abstract*— E-government complaint service has been frequently used by many governments around the world, but less attention have been given to big data analysis especially for data mining. There are many techniques in data mining that is frequently used by developer for business, healthcare, technology, and also government. However, the existing research has shown that selecting suitable data mining techniques for complaint service can provide useful information to government. This literature review aims to investigate data mining technique that suitable to e-government complaint service according to desired outcome and characteristic of data obtained. There are 5 determination factor in this study to determine the suitable data mining techniques such as the desired outcome, characteristic of the data, and etc. The techniques itself have been found 6 such as classification, regression, clustering, summarization, change and deviation detection, and dependency modelling. The result will benefits to government in selecting data mining techniques for complaint service to analyse a large sum of complaint service data.**

Keywords*— *Complaint service, e-government, data mining

I. Introduction

Information and Communication Technology (ICT) have advanced in unprecedented ways to the society. Ignorance of information has automatically disappears by a strong individual initiative that is willing to discover further circumstances. Public has access to resources information around the globe. The consequence, public became critical and responsive to plenty of developing issues. The rapid florescence of Information and Communication Technology (ICT) is a certain that shall be exist and participated by modern public [1].

In the present time, Information and Communication Technology (ICT) plays an important role in existences, for instance in the government. The application of ICT in the government for one of the example is to collect society complaints through e-government service. With e-government, the public will be more easily to submit complaint to government directly. As a result, lots of data is possible to be collected as public has participated with the e-government complaint service. According to lots of complaint data, analysis efficiency can be improved by data mining with appropriate techniques depend on the outcome of needed information and the data itself.

Information system is becoming an important tools in public administration. They offer an automatic routines to citizens and government the opportunity to be interacted with a new channels such as phone or an email [14].

Data mining was defined as six main functions: classification, regression, clustering, Dependency Modelling (Association Rule Learning), Deviation Detection (Anomaly Detection), and Summarization [4]. The purpose of this literature review are to review what methods of data mining are commonly used by various sectors and the outcome of the method itself. Moreover, the second purpose is to discuss possibilities to implement methods of data mining that are reviewed before for analyse e-government service complaints.

This study intends to analyse the various scientific publication to obtain suitable data mining technique of complaint reports for e-government. The purpose of this literature review are to review several related scientific publication to the data mining technique and complaint

978-1-7281-7071-8/20/$31.00 ©2020 IEEE

service e-government, especially how to relate suitable data mining technique and the desired outcome to be obtained for the government.

II. Background Study

A. E-Government

Technologies (IT) in government has changing into e-Government. It provides several benefits in efficiency, availability, costs, and ROI (return on investment) [10]. Nowadays e-Government services are becoming the main elements in sharing of real time information among the government and the non government [11].

E-Government as a room for governments to use the most innovative Information Technology services, especially for web-based applications [12]. Moreover, e-Government is a main object in the modern government around the world, to towards expanding transparent, accounting, and good governance; to make the government efficient, and enabling citizens to access government services and information easy and effective [13].

The definition is wide-raging, it can be used to develop past, present, and future development to the government. David C. Wyld strongly believes that such a definition is the good way to experience e-government, as it does not explain e-government to a only technology or even philosophy [3].

B. Data Mining

Data mining is a process that is used to find useful data through a big amount of data [5]. Data mining has been many used by financial sector, marketers, retailers, even for manufacturers. In financial sector, data mining has been used for credit scoring and fraud detection. In marketing, it has been used for direct marketing, cross-selling and up-selling. In retail sector, it has been used for market segmentation and store layout. In manufacturers, it has been used for quality control and maintenance schedule [6].

Data mining is becoming more prevail everyday, because it helps companies to find profitable patterns and trends from their databases. Companies and institutions have spent much money to get megabytes and terabytes of data, but they are not getting the advantage of the more valuable information hidden deep within their database [7]. Many people define data mining as a knowledge discovery from data, while the others define data mining as a necessary step in the process of knowledge discovery [8]. There are 6 data mining methods according to [4]: classification, regression, clustering, summarization, dependency modeling, and change and deviation detection.

C. Classification

Classification is one of the most important technique in data mining. It classify the data in to predefined targets. As targets are predefined, it is a supervised learning. The function of classification is to classify based on some data with attributes to explain an objects to the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes [9]. After that, the classes is used to determine the group of the data based on the value of another attributes. To analyze the dataset, there are examples of classification techniques such as decision tree, naïve bayes, and neural networks [15].

D. Regression

Regression is one of the most usable techniques by developers. Usually it helps to predict customer satisfaction. The main function of regression is to make a statistic model that is useful to predict some variable from valuable data of one variable.

$$Y = a1b1 + a2b2 + \ldots +C$$

Y = variable or output to predict
a1, a2 = variable to predict (used to)
b1, b2 = size of effect independent variable
c = predicted Y when variables are equal to zero

E. Clustering

Clustering is also popular among developers. Clustering separate objects into some clusters that the objects are connected to the similarity of some object to according to the defined criteria [17]. Moreover, spatial clustering have a function such as to group some objects into several clusters which the cluster contains objects that is similar each other. [18].

F. Summarization

There are many important information in organization or government is stored in an un-structured data type. The information is not stored with database oriented type. These information is found in such of natural language data such as text message, or email [19]. Summarization is often defined as a technique for finding the right label for a data set. One of the example of summarization is mean tabulation. This techniques often used in data analysis, data science, visualization [20].

Many information in organization or company exists in form of unstructured text data. Sometimes, such data or information does not have label or resided properly in database with some good methods, but contained in natural language and it is contained in web pages, email, and many other documents [19]. Summarization is one of data mining technique that is involve in categorize a complex description of data set. The example of summarization techniques are tabulation of averages and standard deviations. These are often applied to data analysis, visualization and automated report.

G. Dependency Modeling

Dependency Modelling: getting to know a model that describes a large associations or dependencies among features [21]. One of the most none method on dependency modeling is association rule mining, the objects of a transaction represent items that were purchased concurrently by using a user.

Each association rule has two measures : confidence value and support value. Confidence value is the percentage of many transactions contain A in transactions contain B; Support val is the percentage of transactions contain A and B in all transactions in the input data set. Another words, the confidence proceed the sum of the connection between objects, while the support measures the significance of the connection between objects [22].

H. Change and Deviation Detection

Change and deviation detection is also called anomaly detection. This technique is defined by the process of finding anomalies in some dataset. These anomalies are called unexpected behaviors. The anomalies are not always be categorized as a bad sign or attack, but it can be surprising which is it might turn into opportunity. It can be harmful and it can be not. This technique provides a significant information in various use, for the example in credit card fraud. Behaviors are recorded when someone is using their credit card properly or purchasing the right transaction. But when the future transaction detects unusual behavior that is not appropriate, it can be a fraud [22].

III. Research Method

In this journal drafting, Systematic Literature Review is the selected method for indentifying and summarizing various technique that are commonly used by data scientists, and obtain certain appropriate technique for complaint services that are studied before. The SLR method has several steps such as planning, conducting, and reporting in order to obtain a systematic journal. In the phase of the writing, author identify, review, evaluate types of data mining method that suitable to be implemented in e-government by using google Scholar as a reference to obtain and analyze data. Google Scholar is the only online research database, because it has a bunch of scientific publication that related to Infomation and Communication Technology (ICT) especially for data mining technique and e-complaint service. Furthermore, simplicity of using google scholar is also one of the significant reason of using the selected online research database.

IV. Result

A. Study Found

For study found in the beginning of the finding was using keywords e-government AND (data mining OR text mining) then for both keywords was found 110 papers. Next the second finding was using keywords (data mining OR text mining) AND technique and was obtained 70 papers. After that the third finding, the next was using: e- government AND complaint service and was found 46 studies.

The sum of the paper is two hundred and twenty-six. Then each paper that was found was analyze to determine the relevancy to the topic.

B. Candidates Studies

In this section, 226 papers were chosen by adjusting the abstract with the research question and the result is 72 papers are selected.

C. *Selected Studies*

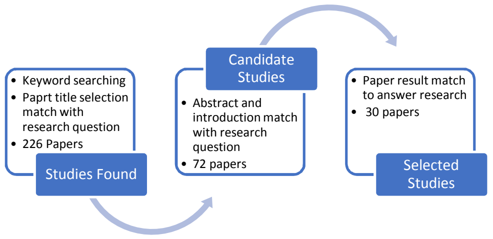
The selected articles or paper should fulfill this criteria:

a) The research must focus on e- government development by using data mining.

b) The paper contains at least 1 data mining technique.

c) The paper according to the research question.

The outcome is the 30 articles fulfill the criteria for a review which can be found in figure 1. Then the data extraction, which is how much the study found from the selected paper can be found in Table 1.



Keyword searching
Paprt title selection match with research question
226 Papers
Studies Found
Candidate Studies
Abstract and introduction match with research question
72 papers
Paper result match to answer research
30 papers
Selected Studies

Fig. 1. Finding strategy for SLR

TABLE I. NUMBER OF STUDIES IN SELECTED SOURCES

Source	Studies Found	Candidate Studies	Selected Studies
Science Direct	140	30	13
Springer	6	4	2
ACM	50	28	10
Emerald	8	2	1
Sage Journal	7	3	1
Google Scholar	15	5	3
Total	226	72	30

D. Candidate Studies

From the 30 selected studies, 45 author are participated, 30 institutions, and 20 universities. Fortunately, each author wrote one paper, and each institution also only has one paper. The location of the institution is in Canada, Thailand, Rusia, France, India, Taiwan, New Zealand, Portugal, Spain, USA, Japan, Italy, Korea, Mexico, Germany, Israel, Vietnam, Australia, Indonesia, and Singapore.

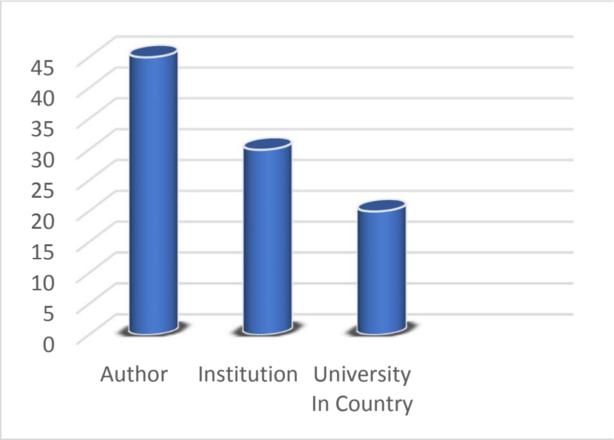


Fig. 2. Author Demography

Authors worked in 14 departments: Information Systems, Information Technology, Business Management, Business, Economics, Economics and Finance, Management, Management Information System, Public and Politics, Public Policy and Technology and Society, Public Administration, Data Science, and Data Analyze. Then all of the department above grouped into 4 groups department, namely Economy, Information System, Information Technology, Public and Politics.

Author's academic background can be found in figure 3, meanwhile the University in the country can be found in figure 4. Paper is selected by publication year between 2010 and 2019 as shown in Figure 5.

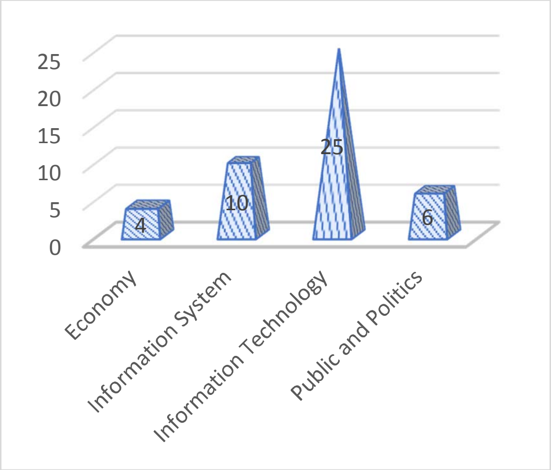


Fig. 3. Author Academic Background

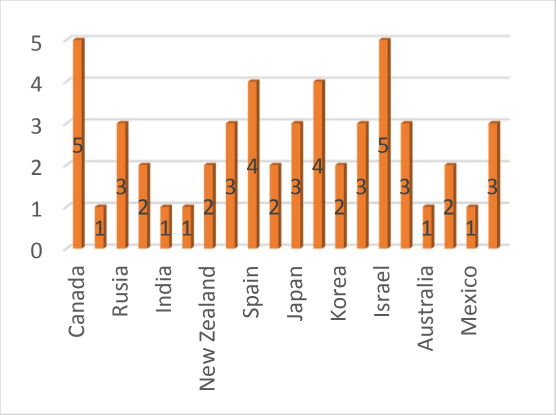


Fig. 4. University in The Country

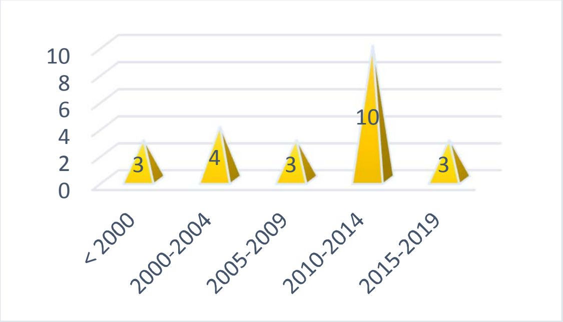


Fig. 5. Publication Year

E. *Suitable Data Mining Techniques in E-Government Complaint*

Bandung e-complain service "LAPOR" contained numbers of complain that has some kind of data such as complaint_id, date, complaint_topic, tags, complaint_title, the_complaint, delete_date, delete_category [23]. From kind of data above, some information are not necessary to be used in data mining. Location can be considered as an additional of data in complaint service. Accordingly, usable data to be used for data mining in complaint service: Complaint Title, The Complaint, Date, Category, and Location.

Detemination of suitable data mining for kind of data above:

- Classification
  Classification can be used to create a new desired category from Complaint Title, or The Complaint itself, sometimes both kind of data can be used to gain a high accuracy. Classification of category can be collaborated with another techniques such as summarization, and clustering [24].
- Regression
  Regression can be used to predict which category that is suitable to a very new complaint. This technique also records behaviour within the data. Uncategorized complaint category can be determined by using regression techniques from behavior of previous data.

- Clustering
  Sometimes, common category is not enough giving information needed for candidates. Detailed categorization can be determined using clustering. In reference [25], the author used clustering combined with other techniques to determined detailed category for the complaints.
  - Summarization
    This technique is necessary to complaint service especially for text mining. Complaint service data has many not understandable sentence. Summarization helps another techniques such as classification and clustering to understand the sentence of the complaint.
  - Dependency Modeling
    Also called association rule. Complaint service could use this technique in determining association between location and the user of complaint service or the complaint itself.
  - Change and Deviation Detection
    This technique is usually used to determine anomaly of behaviors that were captured before. This technique is not suitable for complaint service based on the data found because no anomaly can be found. Sometimes this technique was used to determine fraud of the system most of the time.

F. Determination Factors of Suitable Data Mining Techniques in E-Government Complaint Service

The desired outcome of the analysis of complaints from reports obtained is the main factor of data mining technique selection [23]. The outcome depends on the existing complication that occur in society. The author used classification, clustering, and *Support Vector Machine (SVM)* or regression to produce information that helps government to solve society complication.

TABLE II. DETERMINATION TABLE OF DATA MINING TECHNIQUES IN E-GOVERNMENT COMPLAIN SERVICE

Data mining techniques	Algorithm	Selected journals	Total
Classification	Neural Network	1	12
	Decision Tree	3	
	Naïve Bayes	2	
	K-Nearest Neighbor	1	
	SVM	5	
Regression	Linear Regression	2	2
Summarization	Text Summarization	8	8
Clustering	K-Means	5	6
	Non Hierarchical	1	
Dependency modelling	Generalized Sequence Patterns	1	2
	Apriori	1	

Factors of Suitable Data Mining Technique in E-Government Complaint :

- Desired outcome
  Complication that occur in society varies greatly. Government shall determine the desired outcome of analysis to helps developer to select the most suitable data mining technique.
- Data variant
  Some technique is require particular data variant. The author only used the complaint text for classification and clustering [23]. If the author used multiple data variant, the outcome might get more accurate.
- Size of available data
  The bigger the data, more accurate the outcome. A very small sample size may generate many false rules, thus degrade the performance [26]. A large size of available data can generate a large data sample and generate high accuracy rules.
- Data label
  To build a model that an image is contained some value or not, the first thing to do is to take picture to the right sample of picture [27]. Labeled data often determine whether *supervised* or *unsupervised* learning in data mining. For example: clustering and classification
- Language library for text mining
  The author recommend a specific *stopwords* in sentiment analysis (Indonesian language) which is not sourceful in languages except for English [28]. Availability and resourcefulness language library is very important especially for complaint service text mining to determine which technique is suitable [2].

V. CONCLUSION

There are 6 techniques found in this literature review that frequently used in data mining. According to suitable technique for e-government complaint service, 5 techniques were found. These techniques are classification, regression, clustering, summarization, and dependency modeling but there are only 3 techniques that are precision according study found and that 3 techniques are classification, clustering and summarization. Some of the technique was frequently found combined to another technique. Also there are 5 factors that determine why the found techniques are suitable for data mining in e-government complaint service. The factors are desired outcome, data variant, size of available data, data label, and language library (text mining).

Result of this study will benefit to Government to analyze complaint service using the suitable data mining technique. Furthermore, it could build more appropriate systems to society according to the outcome of the analysis. For future research, it is preferable to understand the new techniques that might be useful for complaint service, to obtain the quality of information that can be useful for Government.

REFERENCES

[1] Rosana, Anita Septiani. , "Kemajuan Teknologi Informasi dan Komunikasi dalam Industri Media di Indonesia." , pp. 144-156, 2010.

[2] Saputra, N., Teguh Bahrata Aji., & Adhistya Erna Permanasari., “Analisis Sentimen Data Presiden Jokowi Dengan PreProcessing Normalisasi & Stemming Menggunakan Metode Naïve Bayes Dan SVM” , Vol.5 No 1, 2016.

[3] Wyld, David C., “The Essential Elements of a Definition of E-Government, Vol.1 No 1, 2004.

[4] Fayyad, Usama., Gregory Piatetsky Shapiro dan Padhraic Smyth., “Knowledge Discovery and Data Mining : Toward a Unifying Framework” ,1996.

[5] Neelamegam, S., E Ramaraj. , “Classification algorithm in Data Mining : An Overview”, Vol.3, 2013.

[6] Koh, H.C. and Tan, G., “Data mining applications in healthcare,” Journal of healthcare information management, Vol 19 No 2, pp. 65, 2011.

[7] Daniel T. Larose., “Discovering Knowledge in Data: An Introduction to Data Mining. Wiley-Interscience”, USA, 2005.

[8] Jiawei Han, Micheline Kamber, and Jian Pei. ,” Data Mining: Concepts and Techniques (3rd. ed.). Morgan Kaufmann Publishers Inc”.,San Francisco, CA, USA, 2011.

[9] Gupta, Shelly & Sharma, Anand. , “Data mining classification techniques applied for breast cancer diagnosis and prognosis,” Indian Journal of Computer Science and Engineering, Vol 2, 2011,

[10] Ozkan, S., & Kanat, I. , “E. e-Government adoption model based on theory of planned behavior: Empirical validation,” Government Information Quarterly, Vol. 28, pp. 503–513, 2011.

[11] Fedotova, Olga, Leonor Teixeira, Helena Alvelos. , “E-participation in Portugal: evaluation of government electronic platforms,” Procedia Technology, Vol. 5, pp. 152–161, 2012.

[12] Alawneh, A., Al-Refai, H., & Batiha, K. , “Measuring user satisfaction from e-Government services:Lessons from Jordan,”Government Information Quarterly, Vol. 30, pp. 277–288, 2013.

[13] Lin, F., Fofanah, S. S. & Liang, D. , “Assessing citizen adoption of e- Government initiatives in Gambia: A validation of the technology acceptance model in information systems success,” Government Information Quarterly, Vol. 28, pp. 271–279, 2011.

[14] Holgersson, & Karlsson. , “Public e-service development: Understanding Citizens’ conditions for participation,” Government Information Quarterly, Vol. 31, pp. 396-410, 2014.

[15] Dangare, Chaitrali & Apte, Sulbha. , “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,” International Journal of Computer Applications, Vol 47, pp. 44-48, 2012.

[16] Awang, M. K., Rahman, M. N. A., & Ismail, M. R. , “Data Mining for Churn Prediction: Multiple Regressions Approach,” Computer Applications for Database, Education, and Ubiquitous Computing, pp. 318–324, 2012.

[17] Huang, Zhexue. , "A fast clustering algorithm to cluster very large categorical data sets in data mining," DMKD 3.8, pp. 34-39, 1997.

[18] Han, Jiawei, Micheline Kamber, and Anthony KH Tung. , "Spatial clustering methods in data mining," Geographic data mining and knowledge discovery, pp. 188-217, 2001.

[19] Chandola, V., & Kumar, V. , “Summarization – compressing data into an informative representation,” Knowledge and Information Systems, Vol 12 No 3, pp. 355–378, 2006.

[20] Crangle, C. E. , “Text summarization in data mining,” In Int. Conference on Soft Issues in the Design, Development, and Operation of Computing Systems, pp. 332-347. Springer, Berlin, Heidelberg, 2002.

[21] Giraud-Carrier, C., & Povel, O. , “Characterising Data Mining software,” Intelligent Data Analysis, Vol 7 No 3, 2003.

[22] Agrawal, S., & Agrawal, J. , “Survey on Anomaly Detection using Data Mining Techniques,” Procedia Computer Science, 60, pp. 708–713, 2015.

[23] Megawati, C. , “ANALISIS ASPIRASI DAN PENGADUAN DI SITUS LAPOR! DENGAN MENGGUNAKAN TEXT MINING, 2015.

[24] Miranda, E,. Aryuni, M., Irwansyah, E., “A Survey of Medical Image Classification Techniques”, 2016 International Conference on Information Management and Technology (ICIMTech), pp. 56-61, 2016.

[25] Aryuni, M., Madyatmadja, E, D., Miranda, E., Customer Segmentation in XYZ Bank using K-Means and K-Medoids Clustering, 2018 International Conference on Information Management and Technology (ICIMTech), 412-416, 2018.

[26] J. Zaki, S. Parthasarathy, Wei Li and M. Ogihara, "Evaluation of sampling for data mining of association rules," Proceedings Seventh International Workshop on Research Issues in Data Engineering. High Performance Database Management for Large-Scale Applications, Birmingham, , pp. 42-50, UK, 1997.

[27] Sheng, Victor S., Foster Provost, and Panagiotis G. Ipeirotis. "Get another label? improving data quality and data mining using multiple, noisy labelers." In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 614-622, 2008.

[28] Miranda, E., Aryuni, M., Hariyanto, R., Surya, E, S., “Sentiment Analysis using Sentiwordnet and Machine Learning Approach (Indonesia general election opinion from the twitter content)”. 2019 International Conference on Information Management and Technology (ICIMTech), pp.62-67, 2019.