BINUS UNIVERSITY GRADUATE PROGRAM Doctoral Program | Doctor of Computer Science
INSTITUT FRANÇAIS INDONESIA
IEEE The Indonesia Section Computer Society Chapter
IEEE INDONESIA SECTION
IEEE Xplore Digital Library

Mining Unstructured Data in Social Media for Natural Disaster Management in Indonesia

Rakhmat Arianto[1,2], Harco Leslie Hendric Spits Warnars[1], Ford Lumban Gaol[1], Agung Trisetyarso[1]
[1] Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University
[2] Informatics Engineering Department, Sekolah Tinggi Teknik PLN
Jakarta, Indonesia
arianto@sttpln.ac.id, spits.hendric@binus.ac.id, fgaol@binus.edu, atrisetyarso@binus.edu

***Abstract*— This paper proposed a model system for unstructured mining data in social media for natural disaster management in Indonesia. The model system of natural disaster management will be tested using real data where the application will be run from the stage crawl social media, tokenization, filtering, stemming, similarity measure, and Name Entity Recognizer so as to ascertain whether the software is built is in conformity with the rules of data collection events natural disasters that can be reliable. The proposed model system of natural disaster management can help the Indonesian government to calculate the impact of floods, landslides, and tornados that it could decide to focus fixes in the correct fields. If the government has made improvements by the mapping of disaster impact it will automatically proclamation of floods, landslides, and tornados in social media and news websites will decrease, and the value graph will change impacts linkages so that the government can focus subsequent repair.**

Keywords*— *Unstructured Data, Natural Language Processing, Natural Disaster, BNPB, Text Mining

I. INTRODUCTION

In Indonesia, the national disaster management is handled by the National Disaster Management Agency (BNPB) where BNPB composed of a head, steering element of disaster prevention, disaster management, and technical committee. BNPB has the function of coordinating the implementation of disaster management task planned, integrated, and comprehensive[12]. In carrying out its duties, BNPB assisted by the Regional Disaster Management Agency (BPBD) that serves as a connector BNPB first level local (provincial) and local levels II (district) [1,3].

In the official website BNPB, shared news of natural disasters or disasters non-natural occurring on a daily basis such as the incidence of terrorism / sabotage, flooding, flooding and landslides, tidal waves / abrasion, earthquake, earthquake and tsunami, crop pests, fires, land and forest fires, industrial accidents, transportation accidents, drought, famine, etc. [2]. From these data, the public can find out the incidence of natural disasters is based on the date of the occurrence of the accident. Moreover, province of the location of the accident a district of the territory of the scene of the disaster, as well as the victims affected by natural disasters such as fatalities, missing victims, the wounded and the victims who managed evacuated.

Moreover, Indonesia is a country located in the tropics which has two seasons, dry season and the rainy season which during the rainy season, it is often natural disasters in various regions in Indonesia such natural disasters as floods, landslides, tornados, etc. BNPB recorded from 2011 to 2016 there have been 174 natural disasters in Indonesia, where there is 81 disaster whirlwind of events, there were 56 occurrences of floods and landslides 29 events. So, we need in-depth research to take into account the impacts that occur due to natural disasters in the field of economics, infrastructure, health, education, and environment.

Law Number 24 of 2007 explains that disasters can be divided into three parts based on their causes, namely natural, non-natural, and human factors. Accidents caused by natural factors are disasters caused by events, or series of events caused by nature, including earthquakes, tsunamis, volcanic eruptions, floods, droughts, hurricanes, and landslides, disasters caused by non-natural factors are disasters caused by events or series of non-natural events which include technological failures, failed modernization, epidemics, and disease outbreaks. While social disasters (caused by human factors) are disasters caused by events or series of events caused by humans which include social conflicts between groups or between communities, and terror[22].

Meanwhile, social media as "a group of Internet-based applications that build by Web 2.0 ideology and technology, and which enable the creation and exchange of user-generated content"[18]. Social media technology takes various forms including magazines, internet forums, weblogs, social blogs, microblogging, wikis, podcasts, photos or images, videos, social ratings and bookmarks[7,8]. By implementing a set of theories in the field of media research (social presence, wealth media) and social processes (self-presentation, self-disclosure) there are six classification schemes for various types of social media, and they are:

1. Collaboration Project, where the website allows its users to be able to change, add, or remove content on this website and for example is Wikipedia.
2. Blog and Microblog, where users are more free in expressing something on this blog such as venturing or criticizing government policies and for example is Twitter.
3. Content, where users of this site share media content, such as videos, ebooks, images, and so on and for example is Youtube.

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

4. Social networking sites, where Application that allows users to be able to connect by making personal information so they can connect with others. Personal information can be like photos, and for example is Facebook.
5. The virtual game world, where it replicates a 3D environment, where users can appear in the form of avatars - desired avatars and interact with others as they should in the real world, and for example is online games.
6. The virtual social world, where users feel that they live in a virtual world, just like virtual games, interact with others. However, Virtual Social World is freer, and more towards life, for example, Second Life

Social media has the following characteristics such as:[6]

1. The message conveyed is not only for one person but many people, for example, messages via SMS or the internet.
2. The message delivered is free, without having to go through a Gatekeeper.
3. The message conveyed tends to be faster than other media.
4. Message recipients that determine interaction time.

In this paper, a model system is proposed that can address the strategic issues faced by BNPB as previously explained by using social media as an information center for natural disasters. This is because the Indonesian people are currently very quickly sharing information about natural disasters in the region through the press social rather than data sent by BPBD to BNPB.

Social media proved to be one of the most widely used communication media in the event of a disaster. This was evidenced in several previous studies such as the use of social networks to measure affected areas and centers of flood disasters in Queensland, Australia [13], analyzing geo-locations and specific keywords in a Tweet to help critical respondents get a picture of the situation most recently during disaster events [19]. A demographic analysis of online sentiments during Hurricane Irene [21,23], search for user responses to 19 crises in one place in Twitter in the period 2013- 2015 [17], extracting data on Twitter to obtain useful information in providing disaster relief during natural disasters [11].

II. Related Works

Nowadays, social media became one of the means of communication to report any incidence of natural disasters occurring around the community. Even the data of natural disasters that are happening around the community more quickly distributed so bring the idea to use social media as a comparison of the data of natural disasters that are happening.

In extracting data from Twitter related to natural disasters in Indonesia will require study of literature describing, among others, in previous studies [4], explains the technique of data collection on Twitter who have hashtags popular and monitor the news associated with the popular hashtags for additional data automatically. The proposed model analyzes traffic patterns of hashtags collected from live streaming to renew the request next collection. To evaluate the adaptive crawling proposed model, used a dataset on Twitter relating to London 2012. The analysis conducted on these studies indicate that adaptive algorithms based Refined Keyword Crawler Adaptation can gather more comprehensive dataset than the pre-defined keyword crawling. For classification datasets such as news of natural disasters of the kind of natural disasters, it is necessary WordNet Similarity method which in previous studies [5,14].

Unstructured data (or unstructured information) refers to useful information that does not have a predetermined or unorganized data model in a calculated way[9,10]. Unstructured information usually consists of many words or phrases, but may contain data such as dates, numbers, and facts as well. This causes irregularities and ambiguity which makes it difficult to understand using traditional programs compared to data stored in the form of fielded in the database or described (semantic tags) in documents.

Techniques such as data mining, natural language processing (NLP), and text analysis provide different methods for finding patterns, or interpreting, this

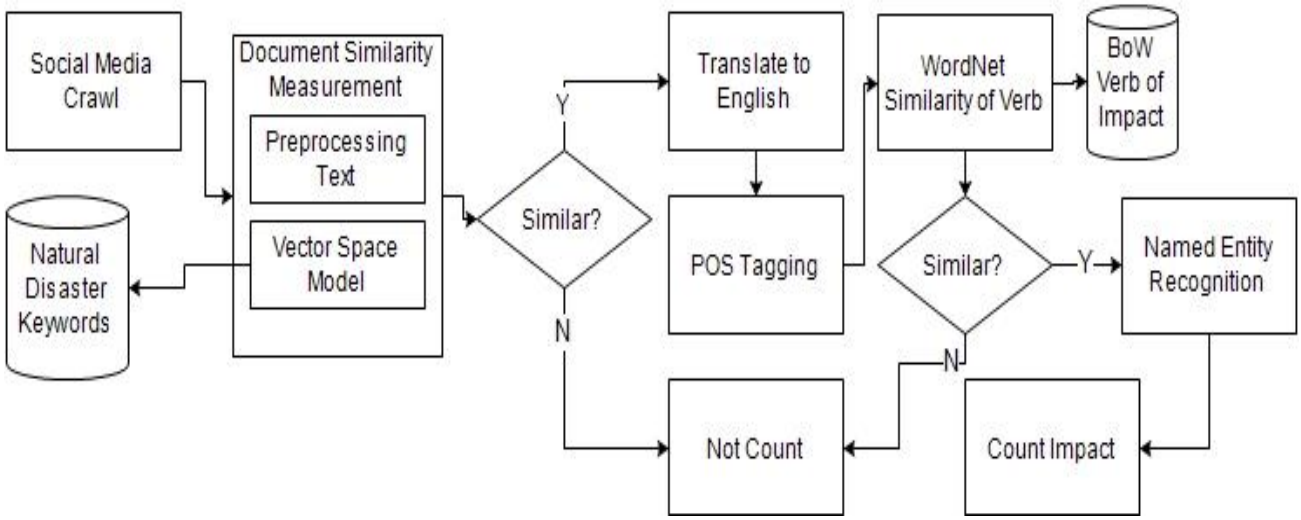

Fig. 1. Proposed Model using social media for natural disaster management in Indonesia

978-1-5386-9422-0/18/$31.00 ©2018 IEEE

information. General techniques for structuring text usually involve manual marking with metadata or part-of-speech tagging for the advanced mining-based structuring of text. Unstructured Information Management Architecture (UIMA) provides a standard general framework for processing this information to extract meaning and create structured data about that information[15]

Conditional Random Field (CRF) is a class of statistical modeling methods that are often applied in pattern recognition and machine learning, where they are used for structured predictions. Whereas ordinary classifiers predict labels for a single sample regardless of the "neighbor" sample, CRF can take the context into account; for example, a linear chain CRF (which is popular in natural language processing) predicts the order of labels for the input sample sequence.
CRF is a type of probabilistic graphical model directed discriminatively. This is used to encode the known relationship between observation and establish a consistent interpretation. This is often used for sequential data labeling or parsing, such as natural language texts or biological sequences [20] and in computer vision[16]. Specifically, CRF finds applications in shallow parsing, named entity recognition, gene discovery and critical finding of critical peptides in functional areas, CRF can be an alternative to Hidden Markov Models (HMMs). In computer vision, CRF is often used for object recognition and image segmentation.

The use of tweets that have been extracted from during the flooding that occurred in Australia in the period 2010-2011. The purpose of social network analysis is used to analyze online networks that appear at that time. The aim is to develop an understanding of online communities for Queensland, New South Wales, and the Victoria flood to identify active users and their effectiveness in disseminating relevant information. The second objective is to identify online resources that are important to be spread by the community. The primary and active users during the Queensland flood were found: local authorities (especially Queensland Police Services), government authorities (Queensland Premier, Prime Minister, Opposition Leaders, Members of Parliament), social media volunteers, traditional media journalists, and people from humanity, not-for-profit associations, and communities. Various essential resources were identified during the Queensland floods; However, this information is more general in valuable information and updates to disasters. Unlike Queensland, there is no evidence of Twitter activity from parts of local governments and governments in the floods of New South Wales and Victoria. However, the floods in New South Wales and Victoria that are active in using Twitter are local volunteers so that the local government gets accurate and reliable information.[13]

III. PROPOSED MODEL

In the proposed model, the detection of the impact of natural disasters using social media can be divided into several important stages, namely: (i) Searching for information related to natural disasters is illustrated by Document Similarity. (ii) The search for verbs that show the impact of natural disasters is described in the WordNet Similarity of Verb process, (iii) The calculation of the number of victims and the location of natural disasters can be detected using Named Entity Recognition. The whole process are illustrated in Figure 1.

Observation of the way the data were collected BNPB natural disasters where BNPB get the data from the data BPBD's natural disasters on the region level I and level II regions were obtained by an event of natural disaster. When a natural disaster has just occurred, then the data will be recorded natural disaster in natural disasters sensor which is owned by BMKG. Where BMKG will share news of natural disasters on the official twitter owned by BMKG @InfoBMKG. By relying on data obtained by the sensor which is held by the BMKG, there will be problems if the sensor is damaged or lost due to being taken by the local community as well as data obtained from BNPB only in periods of 1 year cannot be more real-time.

Data collection is done by collecting the news of natural disasters in Indonesia shared through social media such as Twitter, Facebook, Path, and Instagram in the period of date Also required data collection has been done by the National Disaster Management Authority. Moreover, in some of the period date with the data of natural disasters that occurred in the area of level I and level II. Data collection on words related to the word natural disasters that occurred in Indonesia in English that would be required at this stage of measuring the degree of similarity of the word social media to share news of natural disasters.

In conducting an examination of information related to natural disasters, it can be done by using Document Similarity where a collection of keywords related to types of natural disasters such as earthquakes, floods, landslides, tsunamis and volcanoes has been prepared. Documents obtained from social media crawl will be processed with Preprocessing Text consisting of stages of Tokenizing, Filtering, and Stemming. Tokenizing is the process of changing sentences into lowercase letters, eliminating punctuation in sentences, and separating sentences in words per word. While Filtering is the process of removing words that do not have important meaning in the sentence or contained in the Stop Word List. Stemming is the process of getting the basic words from each word resulting from Filtering.

The results of the Preprocessing Text will be processed using Term Frequency-Inverse Document Frequency (TF-IDF) which serves to give weight to each document. This process can be done using the following formula:

$$w_{t,d} = tf_{t,d} . \log \frac{|D|}{|\{d' \in D | t \in d'\}|} \quad (1)$$

Where, $tf_{t,d}$ = term frequency of term t in document d. $\log \frac{|D|}{|\{d' \in D | t \in d'\}|}$ = inverse document frequency (a global parameter). $|D|$ = the total number of documents in the document set $|\{d' \in D | t \in d'\}|$ = the number of documents containing the term t.

When we have obtained the weight of each document, the next step is to calculate the degree of similarity of the keywords of natural disasters with documents that already have weight by using the following formula:

$$\cos(d_j, q) = \frac{d_j . q}{\|d_j\| \, \|q\|} = \frac{\sum_{i=1}^{N} w_{i,j} \, w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \quad (2)$$

Where, $\cos(d_j, q)$ = similarity between document and query. d_j= document. q=query

To choose which documents have information about natural disasters, the average value of the document similarity level is taken as the lower limit. So, only documents that have similar values above the average value will be selected for processing in the next stage. By getting a similarity value above the average, it can be assumed that the document has information about natural disasters.

The next step is to measure the level of verb similarity using WordNet. Sentences or documents that have been selected, will first go through the POS Tagging process which functions to determine the position of words in sentences such as nouns, verbs, adjectives, adverbs and so forth. By understanding the position of words in sentences, a collection of verbs can be taken in the document to measure their resemblance to verbs that show the impact of disasters such as the word "died", "injured", "affected", and "evacuated" according to the column name from structured data owned by BNPB. Before being processed using WordNet, it is necessary to translate from Indonesian to English because WordNet for Indonesian is still unable to be used properly so that WordNet is used for English. But the translate process will not change data and meaning because the translated word is the basic word.

When a verb that shows the impact of a natural disaster is known, then calculating the number of impacts of natural disasters and finding the location of natural disasters using Named Entity Recognition (NER). This process will process all the nouns that are in the sentence so that it can be known that the noun that is located in sequence shows a phrase where, person, or organization. But in this case only the NER was used with results showing the place so that it could be known where the natural disaster occurred. As for calculating the impact of natural disasters, the results of POS Tagging are shown which shows the principal number and followed by nouns because if there is a noun after the principal number it can be ascertained that the noun is a unit for the principal number. If there are nouns as more than one unit, then the principal number will be added up so that the total impact of natural disasters can be known.

The proposed model system of natural disaster management will be tested using real data where the application will be run from the stage crawl social media, tokenization, filtering, stemming, similarity measure, and Name Entity Recognizer so as to ascertain whether the software is built is in conformity with the rules of data collection events natural disasters that can be reliable.

Verification and validation of the data of natural disasters will be compared between the data obtained from social media with data held by BNPB and validate back in the news on social media are included in the reporting of natural disasters manually. So it can be known whether the data collection of natural disasters to use social media can be more effective than the collection of data from BPBDs that has been summarized by the BNPB in the annual period.

If the data generated from this study has gone through several stages of validation and verification of properly and can be accounted for the results of this study can be implemented by the National Disaster Management Authority. It is because so that the data so far can only show the incidence of natural disasters in one year can be more real-time both in the first level local and regional level II.

The proposed model system of natural disaster management can calculate the impact of natural disasters, the Indonesian government can estimate on what the focus areas of improvement to reduce the effects of floods, landslides, and tornados. The implementation will have some features such as:

a. Searchable keywords related to floods, landslides, and hurricanes.
b. Search keywords related to the economy, the infrastructure, health, education, and waterspout.
c. Collecting data from social media such as news, news website (CNN, BBC, etc.).
d. Of any article or document is found, it will be grouping documents based on five categories of the impact of natural disasters, namely the economy, infrastructure, health, education, and tornado, by the way:
    - Perform Text Mining process to find an essential word in each document.
    - The searchable similarity value of each critical word of the document with the keywords of each category of impact.
    - Do counting words in documents relating to the impact categories to determine which materials are included in a report that addresses one impact category.
    - Finding the value of keyword similarity to natural disasters with disaster impact category.
    - Wanted cause-effect values for each similarity value of each keyword using the Vector Space Model.
    - Named Entity sought in the documents associated with the Google Maps API to determine the location of a discussion of the article.

IV. CONCLUSION

The proposed model system of natural disaster management can help the Indonesian government to calculate the impact of floods, landslides, and tornados that it could decide to focus fixes in the correct fields. If the government has made improvements by the mapping of disaster impact it will automatically proclamation of floods, landslides, and tornados in social media and news websites will decrease, and the value graph will change impacts linkages so that the government can focus subsequent repair.

REFERENCES

[1] P. Republik Indonesia, "Natural Disaster Management Authority." Indonesian Government, 2008.

[2] "Data Dan Informasi Bencana Indonesia." [Online]. Available: http://dibi.bnpb.go.id/data-bencana/lihat-data/per-halaman=50;halaman=2. [Accessed: 01-Jun-2016].

[3] "Gempa 6,5 pada skala Richter landa Sumatera Barat," BBC Indonesia. [Online]. Available:

http://www.bbc.com/indonesia/berita_indonesia/2016/06/160601_indonesia_gempa_sumatera_barat. [Accessed: 02-Jun-2016].

[4] X. Wang, L. Tokarchuk, F. Cuadrado, and S. Poslad, "Exploiting hashtags for adaptive microblog crawling," in Proceedings of the 2013 ieee/acm international conference on advances in social networks analysis and mining, 2013, pp. 311–315.

[5] T. Pedersen, "Duluth: Measuring degrees of relational similarity with the gloss vector measure of semantic relatedness," in Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, 2012, pp. 497–501.

[6] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005, pp. 363–370.

[7] M. T. Pilehvar, D. Jurgens, and R. Navigli, "Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity.," in ACL (1), 2013, pp. 1341–1351.

[8] Warnars, H.L.H.S. 2009. Indonesian Earthquake Decision Support System. The 5th International Conference on Information & Communication Technology and Systems (ICTS) 2009, Informatics Department, Faculty of Information Technology, Institute of Technology Sepuluh Nopember (ITS), Surabaya, Indonesia, 3-4 August 2009.

[9] Warnars, H.L.H.S.. 2009. Sistem Pengambilan Keputusan Penanganan Bencana Alam Gempa Bumi di Indonesia. Olympic Innovative Paper International Conference (Proceeding Olimpiade Karya Tulis Inovatif ,OKTI), L'association des Etudiants Indonesiens en France,Paris, pp. 89, France, 10-11 Oct 2009.

[10] Warnars, H.L.H.S. 2010. Decision Support System for Earthquake Disaster Management, Study Case in Indonesia. Journal Ilmiah Teknik Komputer, 1(2), September 2010.

[11] Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management, (May), 354–358.

[12] BNPB. (2015). Rencana Strategis Badan Nasional Penanggulangan Bencana tahun 2015-2019. BNPB. Retrieved from http://bnpb.go.id/uploads/renstra/Rancangan_Renstra_BNPB_2015-2019_26112015.pdf

[13] Cheong, F., & Cheong, C. (2011). Social media data mining: A social network analysis of tweets during the 2010-2011 Australian floods. In PACIS 2011. AIS Electronic Library (AISeL). Retrieved from https://works.bepress.com/christopher_cheong/2/

[14] Gamble, M., & Kwal, T. (2002). Communication works (7th ed.). Boston, MA: McGraw-Hill College.

[15] Grimm, S. (2008, August 1). Unstructured Data and the 80 Percent Rule. Retrieved from https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

[16] He, X., Zemel, R. S., & Carreira-Perpiñán, M. Á. (2004). Multiscale Conditional Random Fields for Image Labeling. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 695–703). Washington, DC, USA: IEEE Computer Society. Retrieved from http://dl.acm.org/citation.cfm?id=1896300.1896400

[17] Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. CoRR, abs/1605.05894. Retrieved from http://arxiv.org/abs/1605.05894

[18] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business Horizons, 53(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

[19] Kumar, S., Barbier, G., Abbasi, M. A., & Liu, H. (2011). TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence (www.aaai.org). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2736/3201-accessdate=1

[20] Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (pp. 282–289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=645530.655813

[21] Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., & Rodrigue, J. (2012). A demographic analysis of online sentiment during hurricane irene. In Proceedings of the Second Workshop on Language in Social Media (pp. 27–36). Montreal, Canada: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2390378

[22] Pemerintah Indonesia. (2010). Rencana Nasional Penanggulangan Bencana 2010-2014. Badan Nasional Penanggulangan Bencana. Retrieved from http://bnpb.go.id//uploads/renas/1/BUKU%20RENAS%20PB.pdf

[23] Zielinski, a, & Middleton, S. (2013). Social Media Text Mining and Network Analysis for Decision Support in Natural Crisis Management. Proceedings of the 10th International ISCRAM Conference, (May 2013), 840–845.

[24] Warnars, H.L.H.S. 2008. Rancangan Infrastruktur E-Bisnis Business Intelligence Pada Perguruan Tinggi. Journal Telkomnika, 6(2), 115-124, August 2008.

[25] Warnars, H.L.H.S. 2008. Analisa Dampak Investasi Teknologi Informasi Proyek *Data Warehouse* Pada Perguruan Tinggi Swasta Dengan Metode *Simple Roi*. Journal Informatika, 9(2), 101-108, November 2008.

[26] Warnars, H.L.H.S. 2015. Lecturer Decision Support System(DSS) based on Indonesian Lecturer Academic Position Rank. The International Conference on Human-Computer Interaction and User Experience (CHIuXiD 2015), Bandung, Indonesia, pp. 7-10, 8-10 April 2015.

[27] Warnars, H.L.H.S., Sasmoko and Susianna, N. 2014. Introduction investigation: Executive Information System for university. 9th ASEAN Conference on Science and Technology week 2014 (COSAT), Bogor, Indonesia, pp. 353-363, 18-20 August 2014.

[28] Christy,J., Hintarsyah, A.P. and Warnars, H.L.H.S. 2018. Forecasting Sebagai Decision Support Systems Aplikasi dan Penerapannya Untuk Mendukung Proses Pengambilan Keputusan. Jurnal Sistem Komputer, 8(1), 19-27, May 2018

[29] Muyeba, M. K., Khan, M.S., Warnars, H.L.H.S. and Keane J.A. 2011. A Framework to Mine High-Level Emerging Patterns by Attribute-Oriented Induction. The 12th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL), Universiti of East Anglia, Norwich, United Kingdom, pp. 170-177, 7-9 September 2011.