THE EXTRACTION OF INFORMATION DOCUMENT SCIENTIFIC PAPERS USING MAXIMUM ENTROPY MARKOV MODEL

Dina Ilman[1], Ken Kinanti Purnamasari[2]

[1,2]Universitas Komputer Indonesia
Jl. Dipati Ukur No. 112-116, Lebakgede, Coblong, Kota Bandung, Jawa Barat 40132
E-mail: ilmandina@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

Document category detection papers haven't been able to do for various file formats. These problems can be overcome if using machine learning. Machine learning algorithm used is the Maximum Entropy Markov Model (MEMM). MEMM is one algorithm combination of Markov models with Logistic Regression. The training process of the algorithm documents papers studied first pattern of the dataset by specifying the value of the learning rate and initial weights theta. Repetition stop based on cross entropy if convergent approaches and the number of repetitions. Testing based on a model that has already trained with the viterbi algorithm to extract the document's category on scientific papers. Based on testing on 40 documents the scientific thesis of the year 2011 to 2018, average accuracy obtained with testing token-class of 77% with the learning rate is 0.001. Acquisition accuracy token-class due to the use of the less unique feature function to determine the characteristics of each category of documents, training weights will need to use an additional method of theta to be more precise and the influence of the learning rate determines how many of the training process is done.

Keywords: Information Extraction, Maximum Entropy Markov models, MEMM, document papers, thesis, Logistic Regression

1. INTRODUCTION

Information retrieval is the process of extracting information from unstructured text that generates text structured and tidy at once makes it easy to locate information from text structured [1]. Other uses of information extraction, information extraction, according to the Piskorsi aims to extract gobs of text data to get the facts relating to the incident, entity or connectedness in the form of structured information as input to the database [2]. Documents of scientific papers have diverse formats, making it difficult to detect each component in the document. Therefore, the required method for extracting information from documents of scientific papers so that information extraction rules in accordance with document format papers.

Previous research using rule-based on the scientific thesis paper documents speak in Indonesia [3]. The research on the extraction of information is done with a rule-based system to detect identity documents, including the thesis cover, abstract and abstract. Based on testing of 3 pieces of a document in the year 2017, the accuracy of which was acquired for 100% while testing against 50 diverse documents obtained on average 57% accuracy. The decline in accuracy due to the system not being able to handle various file formats. Therefore, the problem can be overcome if using machine learning.

In previous research, the use of machine learning in the field of information extraction is already done on a scientific paper [1]. In the study there is no algorithm for Maximum Entropy Markov Model (MEMM) so that the algorithm is required. The results of the previous research by Susan Mengel and Yaoquin Jing MEMM usage for the extraction of the data structure on a web page have a low error rate 0.14 (0.08 to 30 web pages) [5]. Shurthi s. did a study to named entity recognition bahasa malaysia to use the post tags TnT and method accuracy results with MEMM 82.5% [6]. A. Nedjo et al. using automatic Tag for the post MEMM on Oromo language and generate the accuracy of 99.3% on training data and test data at 93.01% [7].

Based on the exposure in the peneltian method is used to build the system of extraction MEMM information document papers speaking in Indonesia, with the limitations of the document that will be used in this research is a scientific paper documents theses Courses University Computer Engineering Informatics Indonesia.

2. CONTENT OF THE STUDY

The contents of the study describes the research methods, information extraction, document scholarly thesis,

system architecture, tokenizing, feature extraction, and the test results MEMM algorithm

2.1. Research Methods

In these studies there are five stages in the flow of research, including problem identification, literature studies, data collection, system development information extraction and withdrawal of the conclusion. The following block diagram phases flow research.

Figure 1 Research Methods

2.2. Information Extraction

Technology of the extraction of information (Information Extraction) is a technology that deals with how to make unstructured text documents with a particular domain into a structure of relevant information. In other words, the main goal of IE is looking for informations that are relevant to the domain and ignore irrelevant information [10]. In outline, the process of extraction information consists of two phases, namely identifying relevant data, then save it into shape structured for later use [11]. Study of the algorithm of extracting rules based on document text as a trainer has been given the data annotations regarding the entity information to be extracted. The human role is required to provide appropriate annotations or labels of entities that will be extracted in the document [11]:

1. Preprocessing the input data

Data input in the form of unstructured text, natural language text. Important information is obtained with the linguistic analysis, since it generates keywords and features important to identify the characteristics of the information. Linguistic analysis used include tokenizing, Division of the sentence (the sentence splitting), morpologi analysis, parsing, and named entity.

The research on using tokenisasi and extraction features including named entity.

1.1. Tokenizing

The initial state in the form of characters connected with the purpose to identify the basic parts of a natural language such as words, punctuation and separator. The result of the tokens that have meaning and connected as the basis for the process of linguistic and next text.

1.2. Extraction of features

According to Prihatini, the feature extraction is a process to find the feature values contained in documents [12]. Features can be interpreted as a sign of any data that is identified by the system so that it produces the value of the feature. Feature extraction is an important topic in the classification, since the features of good will was able to increase the level of accuracy, whereas features that aren't good tend to exacerbate the degree of accuracy [13].

There are 3 groups of features that will be used with a total of 15 local features, features, features of the layout and the named entity. Local features are characteristics that are present in the characters of each line sentence 7 features. Layout feature is the position of a sentence in the document section 3 features. The feature named entity is a feature that is extracted from certain rules based on document 3 features [14] as for the LOWERCASE and features EIGHTDIGIT is a feature that is determined by the researchers.

2.3. Documents of Scientific Papers

The document is written or printed letters used as evidence information [14]. Scientific papers is a science essay that presents the facts and written according to the methodology of writing a good and true [15]. A large Indonesian Language Dictionary explains that the scientific work in the form of a working paper containing the writings of a tree which is meant to be read in public and often arranged for publication. The sense of scientific papers are papers which contain a bouquet including tasks for learners in school education [14]. There are 2 types of scientific works, namely the research sample thesis, thesis, dissertations, books and papers and reviews or the proposed own idea/sample textbooks, diktat and modules. Theses include results of research that suggested the author based on the opinions of others. The opinions submitted must be supported by empirical data and facts-good objective, based on direct research (field observation) as well as indirect research (study library) [15]. Theses include results of research that suggested the author based on the opinions of others. The opinions submitted must be supported by empirical data and facts-good objective, based on direct research (field observation) as well as indirect research (study library) [15].

On the research of this scientific paper documents used for data input is a document categorization at the front pages or the cover art and the abstract of the study program of Informatics Engineering at random who has 17 category or class. On the cover page is composed of categories of Judul Penelitian (Sampul), Jenis Penelitian, Kalimat Pengajuan, *Other*, Penulis (Sampul), NIM (Sampul), Program Studi, Fakultas, Universitas, Kota dan Tahun. On the abstract page consists of Jenis Halaman, Judul Penelitian (Abstrak), *Other*, Penulis (Abstrak), NIM (Abstrak), Isi Abstrak dan Kata Kunci.

Table 1 Parts of the Category of scientific papers

Cover Page		
No	**Categories**	**Class**
1	Judul Penelitian (Sampul)	0
2	Jenis Penelitian	1
3	Kalimat Pengajuan	2
4	*Other*	15
5	Penulis (Sampul)	3
6	NIM (Sampul)	4
7	Prodi	5
8	Fakultas	6
9	Universitas	7
10	Kota	8
11	Tahun	9
Abstract Page		
No	**Categories**	**Class**
12	Jenis Halaman	10
13	Judul Penelitian (Abstrak)	11
14	*Other*	15
15	Penulis (Abstrak)	12
16	NIM (Abstrak)	13
17	Isi Abstrak	14
18	Kata Kunci	16
12	Jenis Halaman	10
13	Judul Penelitian (Abstrak)	11
14	*Other*	15
15	Penulis (Abstrak)	12

Some of the categories in table 1 into the classroom in the form of resistance training process MEMM. The total number of classes used in this study 17 class.

2.4. System Architecture

Construction of information extraction system there are a number of processes. The main processes in information extraction system using training data consist of MEMM and test data in the form of csv files, preprocessing data training and test data, training and testing of MEMM MEMM. For more details, can be seen on the block diagram of the architecture of the system here.

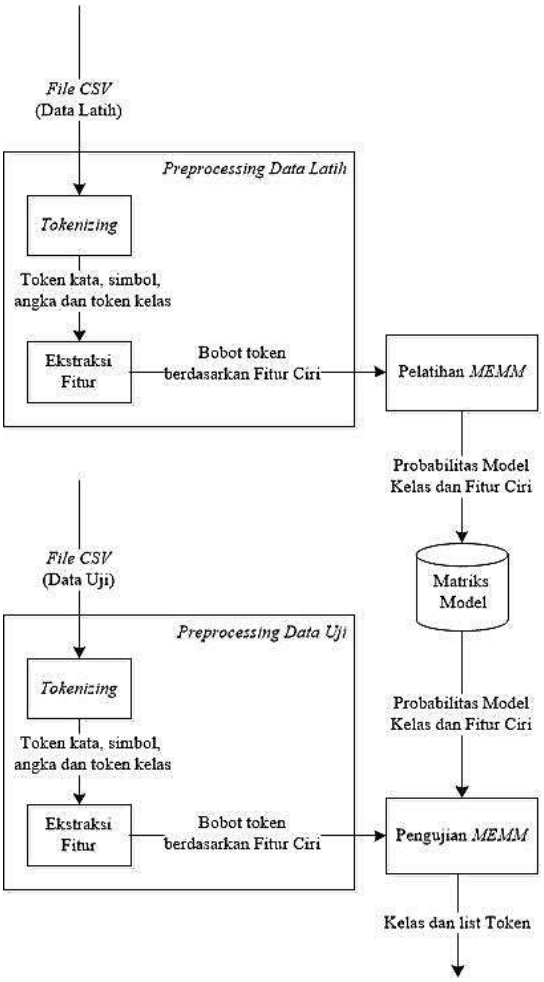

Figure 2 Block Diagram Of The System Architecture

Training data from the scanned cover page and abstract thesis of Informatics Engineering at random in the form of a .PDF and then converting into .txt manually to extract the writing process with OCR. Data preprocessing stage coach through the process of tokenizing. Feature Extraction stage is done by specifying each token features that include results where. There are three features that compared local features, features of the layout and features of the named entity. The stage of Maximum Entropy Markov Model doing training with maximum entropy then testing to determine the sequence of classes based on the data with viterbi. The results of the test data mapping based on models that have been established before.

2.5. Tokenizing

The initial state in the form of characters

punctuation and the separator. The result of the tokens that have meaning and connected as the basis for the process of linguistic and next text [11]. The following block tokenizing diagram.

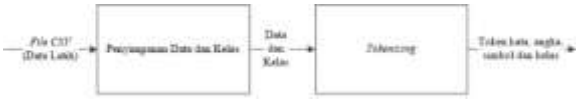

Figure 3 Block Diagram of Tokenizing

Study on the tokenisasi process does not eliminate a number of tokens or words that are not important because it is characteristic for the feature extraction process.

2.6. The Extraction of Features

According to Prihatini, the feature extraction is a process to find the feature values contained in documents [12]. Features can be interpreted as a sign of any data that is identified by the system so that it produces the value of the feature. Feature extraction is an important topic in the classification, since the features of good will was able to increase the level of accuracy, whereas features that aren't good tend to exacerbate the degree of accuracy [13].

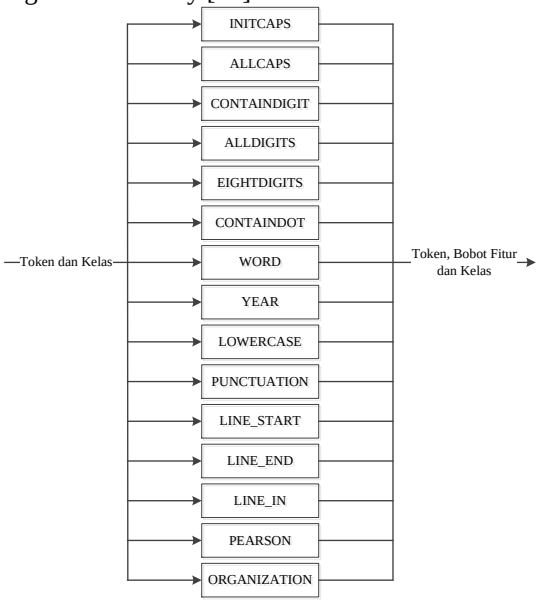

Figure 4 Block Diagram Of The Feature Extraction

There are 3 groups of features that will be used with a total of 15 local features, features, features of the layout and the named entity. Local features are characteristics that are present in the characters of each line sentence 7 features. Layout feature is the position of a sentence in the document section 3 features. The feature named entity is a feature that is extracted from certain rules based on document 3 features [14] as for the LOWERCASE and features EIGHTDIGIT is a feature that is determined by the connected with the purpose to identify the basic parts of a natural language such as words,

Table 2 Feature Extraction

Feature Name	Descriptions
Local Features	
INITCAPS	First alphabet is capital
ALLCAPS	All letters are capital
CONTAINSDIGIT	Contain digit
ALLDIGITS	All letters are digit
EIGHTDIGITS	Eight letters and number digits
CONTAINDOT	At least has one dot
WORD	Give more calculation weight, to class "KALIMAT_PENGAJUAN"
YEAR	Give assign "TAHUN"
LOWERCASE	All small letters
PUNCTUATION	Punctuation
Layout Features	
LINE_START	At starting line
LINE_END	At end line
LINE_IN	At in line
Named Entity Features	
PERSON	Person entity
ORGANIZATION	organization entity

Any tokens that have failed the terms of one of the extraction of these features will get the value weights 1 and if it does not have the value weights 0.

2.7. Maximum Entropy Markov Model

Supervised Learning is also called classification or inductive learning in machine learning. Learning is done from data that has been collected previously and describes the previous situation in the application in the real world [15]. According to Zdravko, the purpose of classification is to make a mapping (also called a model or hypothesis) between a set of documents and a class label set. The mapping results are used to determine the new document class (not yet labeled the initial class) automatically [16]. One of the classifiers that will be used in this study is the Sequence Classifier. Sequencing classifiers are models that are tasked with determining a label or class for each unit in a sequence

researcher.
so that it maps an observation sequence to a label sequence [17]. The algorithms included in sequence classifiers are the Hidden Markov Model (HMM) and the Maximum Entropy Markov Model (MEMM).

Markov models are also referred to as Markov Chain or Markov Process is a part of a stochastic process that has the Markov property, so when given the current state of the input data, the State would come in predictable and he escape from the circumstances of the past. In other words, the intended future conditions by using probabilities. The model part of a Finite state Automaton or Finite.

Finite Automaton is the set state transition between his state is done based on observation. On the Markov Chain, each bow between state contains the probability of the possibility of those lines will be drawn. The number of probabilities all bows out of a node is the one. Markov chain is useful
to calculate the probability of the sequence of events that can be observed. To find out the sequence of events that is hidden using the algorithm of Maximum Entropy Markov Model.

Maxium Entropy Markov Model or MEMM was a sequence of models adapted from MaxEnt (multinomial logistic regresstion) classifier. Based on logistic regression, the model is a sequence MEMM discriminatory while the HMM (Hidden Markov models) model of generative sequence [17].

Using Maximum Entropy MEMM on to its classification process so that the training formula to determine the best target class P (T | W) formula (2.4) as follows [17]:

$$\hat{T} = arg\max_{T} P(T|W)$$
$$= arg\max_{T} P(t|t', w)$$
$$= \operatorname{argmax}_{T} \frac{\exp(\sum_j \theta_j f_j(w_i, t_i))}{\sum_{t' \in seluruh\ kelas\ T} \exp(\sum_j \theta_j f_j(w_i, t'))}$$

t_i is class of i from the designated data. W is a Word from the whole dataset. T' is a whole class early. θ is the initial weight. $f_i(w_i, t_i)$ alah fungsi indikator yang menghasilkan nilai 0 jika syarat tidak terpenuhi dan 1 jika syarat terpenuhi [17]. Indikator yang diambil berdasarkan kelas ke-i dan ekstraksi fitur ke-i.

Multinomial Logistic Regression or also called Maximum Entropy Model of part of pengklasifikasi exponential or log-linear, extract a set of words, combining linearly (times every word with weights and sum it) then applying a function on that combination [17]. The training process in the logistic regression using the objective function of minimizing the value of training data error, is cross entropy loss function (2.5) [17]:

$$L_{CE}(\hat{y}, y) = -\sum_{k=1}^{K} 1\{y = k\} \log p(y = k|x)$$

1as a function to determine if the index class sebenernya to k of the same training data with the index to k then the softmax function value is 1 if not eligible then the function value 0. The value of the cross entropy calculated the average of the entire data to get the value of the overall error. To optimize and update the value of θ weighting based on the cross entropy if not close to 0 then calculated the value of the gradient descent. Following the process of calculation of gradient descent on the formula (2.6) [18]:

$$\frac{\partial L_{CE}}{\partial W_k} = (1\{y = k\} - p(y = k|x))X_k$$

$$\theta' = \theta - \eta \frac{\partial L_{CE}}{\partial W_k}$$

Same with the cross entropy for function 1{}, X_k based on function matrix feature for each of the k class, η learning rate weighting value if the value is too high then the training process will be weighted θ will be too big and have the value of the cross entropy. If the value is too small then make proes training become very slow and very small weight changes. θ' renewal of weights [18]. The process to determine the sequence of classes based on the data of observation called decoding and as a method of testing. This research uses the Viterbi algorithm decoder for it. The following formula (2.7) for the process of testing the Viterbi [18]:

$$arg\max_{T} P(T|W) = \prod_{i=1} \mathrm{P}(t_i|w_1 \dots w_n, t_i\,; \theta)$$

1. Initialization:

   $\pi(*,*,0) = 1$

2. Recursion:

   For every classes $j \epsilon \{1 \dots k\}$, for every tokens $t \epsilon [1 \dots n]$ dan for all classes $i \epsilon [1 \dots k]$

   $v_t(j) =$
   $\max_{i=1\dots k}(v_{t-1}(i) * \mathrm{P}(\mathrm{t}_j|w_1 \dots w_t, t_i\,; \theta))$

3. Termination:
   Nilai tertinggi : P* = $\operatorname{argmax}_{i\,\epsilon 1\dots k}\, v_t(i)$

2.8. The Test Results

Analysis of the testing plan is carried out through two stages, including the stages of token-class.
Will be calculated the level of truth of the classification that has been done by the system, resulting in the value of accuracy and error.

2.8.1. Token-Class Testing

Each token words, numbers, and symbols that has possess class validated the truth, and then calculated the level of truth and faults, resulting in a value of accuracy and error. Following the results of a test token-class classification in table 3.

Table 3 Token-Class Testing Result

No	Document Name	Accuracy	Error
1	1.csv	77%	23%
2	2.csv	13%	88%
3	3.csv	0%	100%
4	4.csv	0%	100%
5	5.csv	13%	88%
6	6.csv	6%	94%
7	7.csv	19%	81%
8	8.csv	0%	100%
9	9.csv	19%	81%
10	10.csv	6%	94%
11	11.csv	6%	94%
12	12.csv	0%	100%
13	13.csv	0%	100%
14	14.csv	6%	94%
15	15.csv	6%	94%
16	16.csv	13%	88%
17	17.csv	0%	100%
18	18.csv	6%	94%
19	19.csv	6%	94%
20	20.csv	13%	88%
21	21.csv	0%	100%
22	22.csv	0%	100%
23	23.csv	0%	100%
24	24.csv	0%	100%
25	25.csv	6%	94%
26	26.csv	0%	100%
27	27.csv	0%	100%
28	28.csv	6%	94%
29	29.csv	13%	88%
30	30.csv	13%	88%
31	31.csv	6%	94%
32	32.csv	6%	94%
33	33.csv	6%	94%
34	34.csv	6%	94%
35	35.csv	6%	94%
36	36.csv	0%	100%
37	37.csv	13%	88%
38	38.csv	13%	88%
39	39.csv	13%	88%
40	40.csv	13%	88%
Average		6%	94%

2.9. Testing Result Analysis

The value of high accuracy and low error obtained has several causes. Following the analysis of the test results with the concepts of accuracy and error token-class. The analysis of the test results with the concept of token-class Because the input data used by the system are text files, so there are limitations on the features that were used, the increase or decrease in the value of accuracy and error caused by the quality of a PDF file generated on the scanning process before, the value of the low accuracy obtained because there are few testing document symbol irregular generated by previous conversion process, based on the observations made, the impact influence on the drop in the value of accuracy is always caused by features of the ORGANIZATION dedicated to provide weighting on the category of courses, faculty, and the University. Therefore, the algorithm MEMM has a network of competitive, so if the Program of study is assumed to be won with the same weighting between the Faculty and the University, the faculties and the University will be terklasifikasikan in the same class as a course, The lack of use of the extraction of features that can recognize a few token specifically against a large number of classes, namely the classes, so the 17th to give weights on each token is not significant
.

3. CONCLUSION

Based on testing the functionality of the system and measurement accuracy has been made, information extraction system using an algorithm of Maximum Entropy Markov Model (MEMM) has successfully built with token-class accuracy tally of 77%. Here the cause against the accuracy obtained. Token-class accuracy gains caused by the use of a weighting feature is not significant, so the algorithm MEMM cannot do the classification properly On this research still has some flaws, so the value of accuracy in the concept of the token-class or grade-token cannot be obtained with the maximum. Thus, some suggestions will be presented for further development regarding information extraction system using machine learning, including using the data input with the docx format or. html in order to use some of the features extras such as detecting the bold, italic, underline, font style and some others. Required a significant additional feature for weighting the categories of Faculty, courses, and universities. It is done so that when the process is done, the testing algorithm MEMM can determine the appropriate classes for faculty, courses, or University. Need a proven algorithm parameters in comparison with MEMM performs the extraction of information in scientific paper documents. It needs checking and checking misspelling or typo on the outcome of conversion of PDF files.

REFERENCES

[1] D. Mustaqwa and N. Indriani, "Implementasi Ekstraksi Informasi Pada Dokumen Teks Skripsi Menggunakan Metode Ruled Based," 2017.

[2] A. I. Riaddy, S. M. Yulianti Sibaroni and S. M. Annisa Aditsania, "Ekstraksi Informasi pada Makalah Ilmiah dengan Pendekatan Supervised Learning," *e-Proceeding of Engineering*, vol. 3, no. 1, pp. 1184-1190, 2016.

[3] E. Susanti and K. Mustofa, "Ekstraksi Informasi Halaman Web Menggunakan Pendekatan Boostraping pada Ontology-Based Information Extraction," *IJCCS*, vol. 9, no. 2, pp. 111-120, 2015.

[4] D. Jurafsky and J. H. Martin, Speech and Language Processing; An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition.

[5] S. Mengel and Y. Jing, "Extracting Structured Data from Web Pages with Maximum Entropy Segmental Markov Model," *WISE*, no. LNCS 5802, pp. 219-226, 2009.

[6] S. S, Jiljo and P. P. V, "A Study On Named Entity Recognition For Malayalam Language Using TnT Tagger & Maximum Entropy Markov Model," *International Journal of Applied Engineering Research ISSN 0973-4562*, vol. 11, no. 8, pp. 5425-5429, 2016.

[7] A. T. Nedjo, D. Huang and X. Liu, "Automatic Part-of-speech Tagging for Oromo Language Using Maximum Entropy Markov Model (MEMM)," *Journal of Information & Computational Science* , vol. 11, no. 10, pp. 3319-3334, 2014.

[8] P. D. Sugiyono, Metode Penelitian Kombinasi (Mixed Methods), Bandung: Alfabeta, 2014.

[9] I. Sommerville, Software Engineering, 9th ed., Pearson, 2011.

[10] Tellez-Valero, Alberto, M. Montes-y-Gomez and L. V. Pineda, "A Machine Learning Approach to Information Extraction," *International Conference on Intelligent Text Processing and Computational Linguistics CICLing*, pp. 539-547, 2005.

[11] C. Siefkes, "An Overview and Classification of Adaptive Approaches to Information Extraction," *Lecture Notes in Computer Science 3730, Journal Semantics IV Springer-Verlag Berlin Heidelberg*, pp. 510-521, 2005.

[12] C. Siefkes, "An Incrementally Trainable Statistical Approach to Information Extraction Based on Token Classification and Rich Context Models," 16 Februari 2007. [Online]. Available: http://www.siefkes.net/talks/disputation-ie.pdf. [Accessed April 2018].

[13] E. Setiawan, "Kamus Besar Bahasa Indonesia," Kemdikbud Pusat Bahasa, 2018. [Online]. Available: https://kbbi.web.id. [Accessed 05 2018].

[14] P. D. E. Z. Arifin, Dasar-Dasar Penulisan Karya Ilmiah, Jakarta: Grasindo.

[15] B. Liu, Web Data Mining, University of Illinois, Chicago: Springer, 2011.

[16] Z. Markov and D. T. Larose, Data Mining The Web : Uncovering Patterns In Web Content, Structure, and Usage, New Jersey: John Wiley & Sons, Inc, 2007.

[17] D. Jurafsky and J. H. Martin, "Speech and Language Processing 3rd ed. draft," 28 August 2017. [Online]. Available: http://web.stanford.edu/~jurafsky/slp3/. [Accessed April 2018].

[18] M. Collins, "NLP Edu ~mcollins," 2014. [Online]. Available: http://www.cs.columbia.edu/~mcollins/fall2014-loglineartaggers.pdf. [Accessed July 2018].

[19] D. L. Olson and D. Delen, Advanced Data Mining Techniques, Berlin: Springer, 2008.

[20] S. Iqbal, S. A. Qureshi, T. H. Rizvi, G. Abbas and M. M. Gulzar, "Concept Building Through Block Diagram," *Journal of The Institution of Electrical and Electronics Engineers Pakistan*, Vols. 66-67, pp. 30-34, 2010.

[21] A. Kadir, Dasar Pemrograman Java 2, Yogyakarta: Andi, 2004.