APPLICATION OF THE SUPPORT VECTOR MACHINE METHOD ON INDONESIAN PART OF SPEECH TAG

Asep Sumpena Nugraha[1], Ken Kinanti Purnamasari[2]

[1,2]University Computer Indonesia
Jl.Dipati Ukur No.112-116, Lebakgede, Coblong, Bandung, West Java 40132
E-mail: asepsn7@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

POS Tag is the classification of words into classes. In making POS Tags there are difficulties related to ambiguation because the structure or grammar of a language is often different or diverse. This problem can be overcome with machine learning. The method used in this study is Support Vector Machine or what is often referred to as SVM. In the SVM process to limit between classes using a Linear Kernel. Basically SVM is a machine learning that can only classify two classes but after being developed, SVM can classify more than two classes with a technique called multiclass. After seeing several studies on multiclass, it was decided in this study to use the Multiclass One Versus All (OVA). The input data used was 50,005 tokens in the training process while the overall testing process totaled 12,111 tokens resulting in an accuracy of 54.29%. This is caused by features when information retrieval is still less effective.

Keyword: Part of Speech Tag, Support Vector Machine, Ambiguation, Kernel Linier, Multiclass One Versus All.

1. INTRODUCTION

Part of Speech tags or often called POS Tag aims to classify each word into a class corresponding word (verb, adjective, etc.). POS Tag NLP contribute to other fields such as Question-Answering, speech recognition, information retrieval systems, machine translation systems, word sense disambiguation system, etc [1].

POS Tag research has been done on the Indonesian language text. In the study, the method used is rule based with word class numbered 23 pieces and has an accuracy rate of 79% [2]. Based on the results of these studies, Rule Based have ambiguation related problems resulting in decreased levels of accuracy. Ambiguation is a word that has more than one word class [3]. This is caused by the difficulty of the rule-making Rule Based in dealing with the structure and grammar. These problems can be dealt with machine learning. There are several methods that have not been applied to the POS Indonesian Tag including the SVM, KNN, and Decision Tree. In a comparative study of methods at POS Tag thailand language, the best method is the accuracy of the SVM followed by Decision Tree and lastly, KNN [4]. Other research shows SVM has a better accuracy in comparing these three methods is the research conducted by Amr [5].

Research on ambiguation handling has been done using SVM can take a look at the research done by Gimenez for POS Tag English and Spanish. Values obtained very good accuracy that is equal to 97.16% for English and 98.86% for the Spanish language [6] [7]. In addition to the research conducted by Gimenez, SVM at POS Tag English language has also been carried out by Nakagawa. The data is used as much as 2,416 words or 56 684 words. The study resulted in an accuracy of 97.11% [8].

Based on these studies, POS Indonesian Tag that will be created using SVM is likely to have a high accuracy as well. With other reasons also that the SVM method has never been applied to the POS Tag Indonesian, it was decided to make the study of Indonesian tagger POS system using SVM.

2. RESEARCH CONTENT

On the contents of this study will be explained on the POS Tag, research method, system architecture, the input data, tokenization, feature extraction, training SVM, SVM testing, and test results.

2.1 POS Tag

POS Tag can be interpreted to determine the morphosyntactic category of each word in the sentence given [7]. POS Tag can also be defined by the classification by categorizing each word into his class.

Table 1. Examples of POS Tag

Text	Token	Class
Saya memegang tongkat.	Saya	PRP
	memegang	VB
	tongkat	NN
	.	Z

In Table 1 contained the text "Saya memegang tongkat.". Before it is classified, the text is divided into several tokens. Once the tokens classified by grade appropriate word. "Saya" belong to a class PRP word in which the class is defined as a personal pronoun, another example is she, you, me, them, you guys. The second word "memegang" belong to a class

VB words (verb) which means doing an activity or employment. The third word is "tongkat" which is a word class NN (noun) which is a form of objects or frequently used as subject or object. Last token is "." That belong to the class of words punctuation or Z.

Number of POS Tag each class is usually different languages, this is caused by a variety of grammatical forms. Classes are used in this study adapted to the existing classes in the training data that has been created by the UI POS Tag in the form of a corpus with TSV format [9] which amounted to 23, The classes include.

Table 2. Target Class

Tag	Description
CC	Conjunctive or can be called also coordinating conjunction, usually used to connect one sentence to another sentence.
CD	*Cardinal Number*, A numeric or words that show a numeric. An example is the one, two, a few.
OD	*ordinal Number*, A word or value that indicates the position. This class differences by class CD, if the CD is to numeric or quantity but more to show the position OD. Examples are the 1st, second, last.
DT	Normally kept in front of the noun to mark it. Can be a definite or indefinite. This class can also be called determiner
FW	Foreign Word, a word from a foreign language that does not exist and has not been adapted and absorbed by or KBBI Indonesian dictionary.
IN	This word almost the same class with CC word class which is conjunctive but are usually kept in front of the preposition and produce preposesi said.
JJ	Is a word which the descriptors, modification, or some specific properties of a noun phrase. These classes are usually referred to the adjective.
MD	Capital and the auxiliary verb
NEG	Is a word that is negative or rejection
NN	Noun is a word that shows humans, animals, concepts, direction, related to time, and currency.
NNP	Is the specific name of a person, geography, countries, organizations, institutions or companies, day, month, competition, and the symbol of the stock. The word class may also be referred to proper nouns.
NND	Is a word that indicates the scales said, making a case which was not calculated to be numbered.
PR	In English is often called the Demonstrative pronouns, designation implies place of the object.
PRP	The personal pronoun, both for individuals and more than one person as our
RB	Words that describe adjectives and verbs.
RP	Particle, a particle that is usually found in declarative sentences, interrogative, or imperative.
SC	*subordinating conjunction* or also called subordinator, is a link between two or more clauses are usually divided into the main clause and the clause supporters.
SYM	*Symbol*, Tokens included in class this term are the currency symbol such as "$".
UH	The word for calling.
VB	The verb or activities undertaken.
WH	Words that are asked or Question word.
X	Type a word that belongs to a class of this word is a misspelling or typo or words that are not included in the class of other words.
Z	Punctuation or often referred to Punctuation.

2.2 Research Method

The method used is descriptive [10]. This method has seven stages among which the identification of problem, goal research, collect datasets, software development, testing, and conclusion.

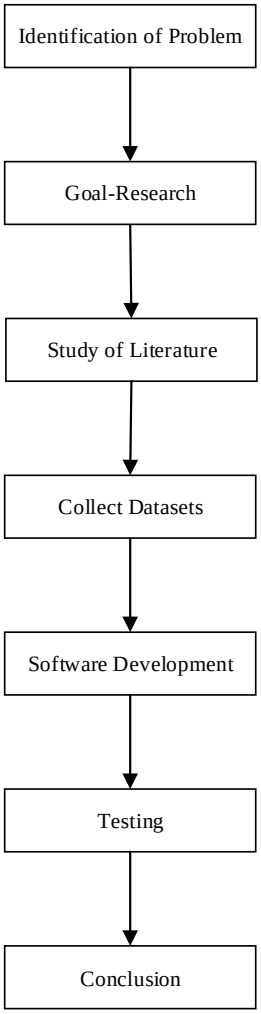

Figure 1. Research Method

2.3 System Architecture

In the construction of a system for implementing Support Vector Machine on Part of Speech Indonesian Tag has several stages. These stages can be seen in Figure 2.

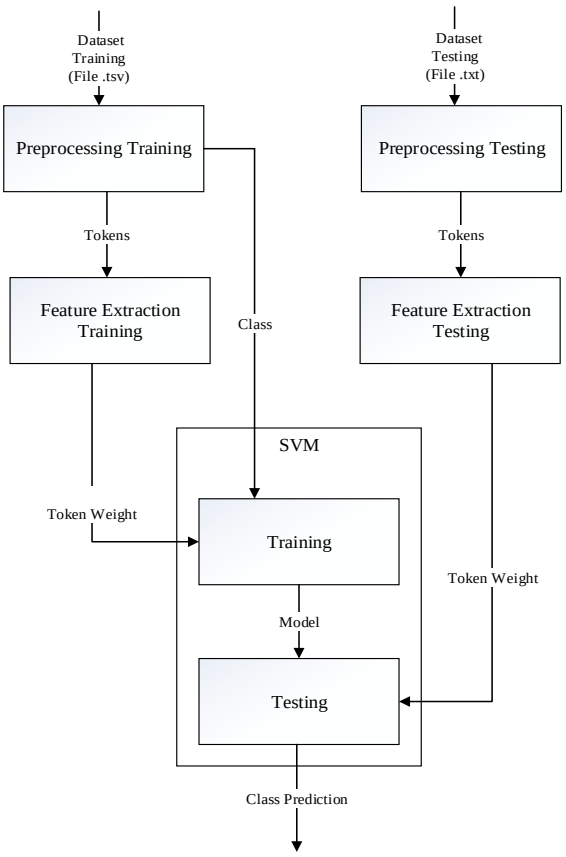

Figure 2. System Architecture

2.4 Data Input

Training data taken from previous research based rule which consists of a token and a class with a separator in the form of tabs in .tsv file format.

Jamrud	NNP
Bahia	NNP
kemudian	RB
dibawa	VB
ke	IN
Las	NNP
Vegas	NNP
untuk	SC
dilihat	VB
oleh	IN
calon	NN
pembeli	NN
namun	CC
disita	VB

oleh	IN
polisi	NN
Los	NNP
Angeles	NNP
.	Z

Figure 3. Training Data

As for its testing of data taken from Panlocalization. The file format used is .txt. The contents of the file in which the text Indonesian.

Korea Selatan dan Jepang terus membawa keuntungan.

Figure 4. Testing Data

2.5 Tokenization

In the preprocessing process used tokenization (tokenizing), this process of dividing the text into tokens. Punctuation and symbols are categorized into classes of words 'Z' or 'SYM' [11]. Tokenization on training and testing have a different process. In tokenization training, the data is separated by a tab for files that are used in the form of .tsv file.

Figure 5. TokenizationTraining

While on tokenization testing every text separated by punctuation although the punctuation is taken as the word class as well. Here groove tokenization testing.

Figure 6. Tokenization Testing

2.6 Feature Extraction

Could be interpreted as characteristic features / information, so the process of feature extraction is the process of taking the characteristics / information contained in a data [12]. Extraction of useful features to distinguish between one class with another class. These features are used in this study, there were 14 features [3] [6] [7] [13] among others.

Table 3. Features

features	provision	Example
caps	Is True if the initial letter of the token that examined the form of capital	***G**ates*, ***D**ia*
In_Cap	Is True if the token that examined containing capital letters except for initial	*i**P**hone*
All_Cap	Is True if all the letters of the token that examined the form of capital	***SBY***
All_Low	Is True if all the letters of the token that examined a small letter	***Selalu, bekerja***
Num	Is True if the token that examined a numerical number.	***1, 20***
Hyp	Is True if the token that examined the symbolic ("-").	*Bolak-balik, kejar-kejaran*
Me-	Is True if the token is checked to contain the suffix "me"	***Me**lakukan, **me**melihara*
Pe-	Is True if the token is checked to contain the suffix "pe"	***Pe**laksana-an, **pe**meriksaan*
Ke-	Is True if the token is checked to contain the suffix "to"	***Ke**laparan, **ke**keliruan*
Se-	Is True if the token is checked to contain the suffix "se"	***Se**ekor*
Be-	Is True if the token is checked to contain the suffix "be"	***Ber**main, **ber**napas*
In-	Is True if the token is checked to contain the suffix "in"	***Di**diamkan, **di**rendam*
-an	Is True if the token is checked to contain the suffix "an"	*Pengaku**an,** kekeliru**an***
-kan	Is True if the token that examined containing additive kan	*Panas**kan,** kobar**kan***

2.7 SVM

Support Vector Machine (SVM) is a learning machine that was first introduced by Vapnik. SVM is a supervised machine learning to solve problems such as the category of text, handwritten digit recognition, tone recognition, image classification and object detection, and classification of data [14].

In the SVM method, the main point is to optimize the name hyperplane. Hyperplane used as a boundary separating one class support vector with more class support vector. By optimizing Support vector particularly adjacent support vector between one class with another class used as a benchmark for the classification limit in order hyperplane to be made to be optimal. This vector is derived from a dataset that

has been converted to a vector through vectorization after feature extraction process and serve as a support vector. For example in the training dataset consists of x and y in the form of$\{(x_1, y_1), \dots, (x_n, y_n)\}$ Where called a vector x and y is the class label [15].

2.7.1 Kernel

Kernel on Support Vector Machine method is the separation between the first class with another class. There are a few kernel to support vector machine such as Linear, polynomial, and Radial Basis Function (RBF). In previous studies, the kernel is frequently used Linear and Radial Basis Function (RBF) [16].

a. Linear

Linear kernel using the straight line as a barrier / hyperplane between classes. Linear kernel requires only two variables such x_i and x_j. The variable is the vector of the result of vectorization to the weight values of feature extraction. x_j in the calculations, the value of vector x_i which ditranspose before multiplied by. As with linear kernel to label labeling result in the target class.

$$k\left(x_i, x_j\right) = x_i^T x_j \qquad (1)$$

b. RBF

In the settlement RBF require parameter gamma and C. Gamma function as boundary decisions and decision areas, for example if gamma is small then the decision will be a small but limit the decision region will be broad and vice versa. Gamma value used must be greater than zero. C serves as a penalty to errors in classification. x is taken from vector vectorization results. Exp is the exponent of the result of the calculation of x and gamma.

$$exp(-\gamma||x_i - x||^2), \gamma > 0 \qquad (2)$$

c. Polynomial

Polynomial kernel has two different parameters of another kernel. Parameter r is a free parameter which if filled r equals zero then it is called homogeneous. While the parameter d is the degree / quadrat are generally filled with d equal to 2.

$$\mathrm{k}(x_i, x) = \left(y.\ x_i^T\ x + r\right)^d \qquad (3)$$

In this study, the kernel used is a linear kernel with reference to the research conducted by Gimenez who have an accuracy above 90% results.

2.7.2 Multiclass

There are two multi-class technique often used in SVM One Versus One (OVO) and One Versus All (OVA). OVA is comparing one with all other than himself who is considered to be one unit. Multi-class SVM is used because his life is the only machine learning classifying only two classes in a linear fashion.

In previous studies, it is known that OVA the accuracy is better than the OVO, although OVO in terms of a faster pace [17]. Therefore, the research to be conducted on the application of SVM against Indonesian Tag POS uses the OVA. With this method is expected to provide high accuracy.

Table 4. Multiclass One Versus All

$h_i = 1$	$h_1 = -1$	hypothesis
Class 1	Not Class 1	$f^1(g) = (w^1)g + b^1$
Class 2	Not Class 2	$f^2(g) = (w^2)g + b^2$
Class 3	Not Class 3	$f^3(g) = (w^3)g + b^3$
Class 4	Not Class 4	$f^4(g) = (w^4)g + b^4$

In applying this method to be built z OVA fruit binary SVM models. *z* is the number of classes. In classifying it can be seen in the following equation [18].

$$Kelas\ g = arg\ \max_{r=1..z}((w^{(r)})^T . \varphi \begin{bmatrix} p_i \\ q_i \end{bmatrix} + b^{(r)}) \qquad (4)$$

The equation used in step final testing process after testing support vector of the data obtained. is a hyperplane which number as many as *z* or the number of classes of training data. Hyperplane is a model or class boundaries from the calculation of SVM training. In addition, there are also other calculation results SVM training or bias. Support vector testing of data entered into the equation as a component of testing the data to be classified. The results of the classification determined by arg max that tertinggilah value to be taken from the calculation all the support hyperplane with testing data vector as target class or classes can also be referred to as a prediction.

$$\left(w^{(r)}\right)^T b^{(r)} \varphi \begin{bmatrix} p_i \\ q_i \end{bmatrix}$$

2.7.3 Training SVM

In the training phase, there are several processes. Such processes include vectorization, Labeling, and SVM. Here groove SVM training.

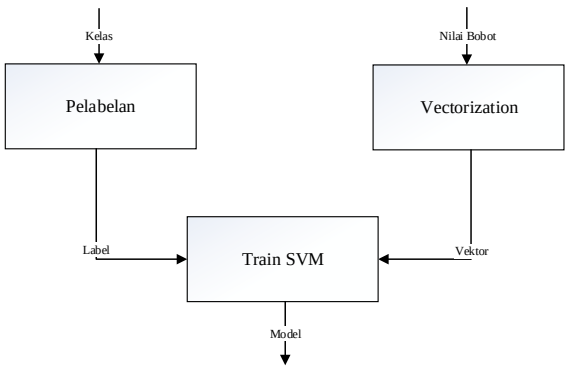

Figure 7. Training SVM

In the SVM training process is the first step to change the weighting feature of feature extraction results into a format that can be accepted by SVM value in the form of vector. The process is called vectorization. The next step is to label classes with 1 or -1. In this study the process of labeling using multiclass One Versus All (OVA). After that input to the kernel, the kernel used this research is linear kernel.

$$\sum_{i=1}^{n} x_i x_j = x_i^T x_j (i,j = 1,\dots,n) \quad (5)$$
$$\sum_{i=1}^{n} y_i y_j = y_i^T y_j (i,j = 1,\dots,n) \quad (6)$$

The value of x is taken from vector vectorization results. While the value of y is taken from the class labeling. Complete these steps of i equal to 1 to n-th, as well as with j. n here is the number of tokens or data. After that create a matrix using the following equation.

$$C = \begin{bmatrix} x_i^T x_j & \cdots & x_i^T x_{jn} \\ \vdots & \ddots & \vdots \\ x_{in}^T x_j & \cdots & x_{in}^T x_{jn} \end{bmatrix} \quad (7)$$

$$D = \begin{bmatrix} y_i^T y_j & \cdots & y_i^T y_{jn} \\ \vdots & \ddots & \vdots \\ y_{in}^T y_j & \cdots & y_{in}^T y_{jn} \end{bmatrix} \quad (8)$$

In this equation, the values of the matrix C are taken from Equations 5 and the values of D are taken from Equation 6. i starting from 1 up to-in, while j starting from 1 until all jn. in and jn is the amount of data. after completion of adding each row of the matrix in order to get the value of k and l.

$$k_i = x_i^T x_1 + \cdots + x_i^T x_j \quad (9)$$
$$l_i = y_i^T y_1 + \cdots + y_i^T y_j \quad (10)$$

Do this from i equals 1 to n-th. n is the total number of training data Next do the following equation.

$$\varphi \begin{bmatrix} k_i \\ l_i \end{bmatrix} = \begin{cases} \sqrt{{k_n}^2 + {l_n}^2} > 2, maka \begin{bmatrix} \sqrt{{k_n}^2 + {l_n}^2} - k_i + |k_i - l_i| \\ \sqrt{{k_n}^2 + {l_n}^2} - l_i + |k_i - l_i| \end{bmatrix} \\ \sqrt{{k_n}^2 + {l_n}^2} \leq 2, maka \begin{bmatrix} k_i \\ l_i \end{bmatrix} \end{cases} \quad (11)$$

Calculations start by checking whether the result will be worth more than 2 or may be equal to or less than 2. If the result is more than 2, the formula used is that is above, but if not then the result is equal to $\begin{bmatrix} k_i \\ l_i \end{bmatrix}$ [19]. *K* value obtained from Equation 9 while the value obtained from Equation 10. l Do this from i equals 1 to i equals n. n is the total number of training data

Equation 11 will result in the value of α_i training support vector. In search of value, value added support vector perpendicular bias by 1 so perfect. Then do the following equation.

$$\sum_{i=1,j=1}^{n} \alpha_i s_i^T s_j = l_i \quad (12)$$

Complete these steps of i equal to 1 to n-th, as well as with j. By using software mapple then get the value a. If the value has been obtained, the next step is to find the value of w and b are new as a hyperplane.α_i

$$w = \sum_{i=1}^{n} \alpha_i s_i \quad \text{Where } \alpha_i \geq 0 \quad (13)$$

In calculating the hyperplane, which is used must be positive. By using these equations will be obtained value w and b simultaneously. This value is considered as the hyperplane, the hyperplane collection called the model to be used in the process of testing the SVM.α_i

2.7.4 Testing SVM

Testing is the testing data entry process into the model that has been made during the training phase. At this stage, the data in the form of weight testing the feature after preprocessing and feature extraction.

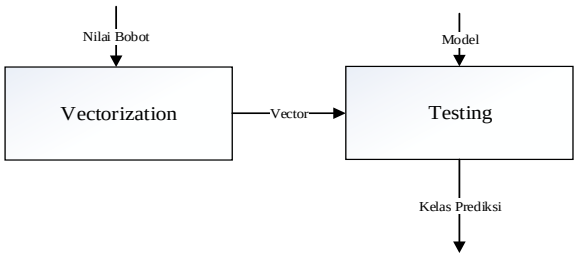

Figure 8. Testing SVM

Similar to the training process, the weight should be converted into a vector, whereas for the class labeled. In the process of labeling a class for testing the data being equal to 0, while the label for the class of other data the same as in the training process in which the testing of data tokens combined with training data to find the value of support vector testing [15]. After that search the variables needed in the calculation kernel. Perform the following calculation with i from 1 to n-th, as well as with j. The value of n is the number of data.

$$\sum_{i=1}^{n} g_i g_j = g_i^T g_j (i,j = 1,\dots,n) \quad (14)$$
$$\sum_{i=1}^{n} h_i h_j = h_i^T h_j (i,j = 1,\dots,n) \quad (15)$$

After that mapped into the matrix. The matrix E for a weight vector matrix testing and class F for testing.

$$E = \begin{bmatrix} g_i^T g_j & \cdots & g_i^T g_{jn} \\ \vdots & \ddots & \vdots \\ g_{in}^T g_j & \cdots & g_{in}^T g_{jn} \end{bmatrix} \quad (16)$$

$$F = \begin{bmatrix} h_i^T h_j & \cdots & h_i^T h_{jn} \\ \vdots & \ddots & \vdots \\ h_{in}^T h_j & \cdots & h_{in}^T h_{jn} \end{bmatrix} \quad (17)$$

Then look for value of p_i and q_i by calculating Equation 18 and Equation 19. $g_i^T g_j$ taken from each line matrix E and $h_i^T h_j$ is taken from each line matrix F.

$$p_i = g_i^T g_j + \cdots + g_i^T g_j \quad (18)$$
$$q_i = h_i^T h_j + \cdots + h_i^T h_j \quad (19)$$

Having obtained the value of the kernel p_i and q_i. then find support vector. The following equation to find support vector testing.

$$\varphi \begin{bmatrix} p_i \\ q_i \end{bmatrix} = \begin{cases} \sqrt{{p_n}^2 + {q_n}^2} > 2, maka \begin{bmatrix} \sqrt{{p_n}^2 + {q_n}^2} - p_i + |p_i - q_i| \\ \sqrt{{p_n}^2 + {q_n}^2} - q_i + |p_i - q_i| \end{bmatrix} \\ \sqrt{{p_n}^2 + {q_n}^2} \leq 2, maka \begin{bmatrix} p_i \\ q_i \end{bmatrix} \end{cases} \quad (20)$$

Similarly, the process of training before starting to count the support vector, first check whether the results of $\sqrt{{p_n}^2 + {q_n}^2}$ more than 2 or less than or equal to 2. If the result is more than 2 then use the formula above otherwise the same as the support vectornya $\begin{bmatrix} p_i \\ q_i \end{bmatrix}$, The last stage of the testing process that is input support vector testing of data into the model hyperplane that Equation 4. Results of the largest value in the calculation of the grade predictions. This prediction classes will be compared with the original class to get the number of correct predictions class in calculating the value of accuracy.

2.8 Test Result

Tests carried out by the number of correct predictions of results of the classification of the token and token classes with training data which amounts to 50,005 tokens and data testing, amounting to 12,111 tokens.

Table 5. Test Result

Name	Total tokens	Number of Class Prediction Right
Testing1.txt	1,012	561
Testing2.txt	1,581	978
Testing3.txt	1,001	522
Testing4.txt	1,254	654
Testing5.txt	1,364	718
Testing6.txt	1,110	635
Testing7.txt	1,010	506
Testing8.txt	1,607	796
Testing9.txt	1,025	572
Testing10.txt	1,147	633
Total	12,111	6,575

Based on the results of testing that has been done, then count accuracy value by dividing the total of the number of tokens with the total number of correct predictions class and multiply the result by one hundred percent. Values obtained accuracy is at 54.29%.

3. CONCLUSION

The results obtained through testing showed that the average accuracy of POS Tagger Indonesian obtained using SVM in this study was 54.29%. It is influenced by all_low feature extraction features that are not effective in differentiating between the class that is required of other features to help feature 'all_low' is to be more effective.

Because there are still many shortcomings in this study, there are some suggestions that can be done for future development. The following suggestions can be taken.

a. Checking basic words before recharge feature information retrieval process. If the token is not the root then all the features of a particle is False. For example, the word 'I' is a basic word it does not have any affix. If the token is checked not including the base word tokens are likely to have a particle like the word 'damage' that has a suffix 'me'.
b. The addition or alteration of features in the feature extraction process. For example, adding a feature affixes 'tar', 'sasi', and 'iah'.

BIBLIOGRAPHY

[1] R. J. P. M. C. Padma, "Morpheme Based Parts of Speech Tagger For Kannada," 2016.

[2] F. Rashel, A. Luthfi, A. Dinakaramani dan R. manurung, "Building an Indonesian Rule-Based Part-of-Speech Tagger," 2014.

[3] K. K. Purnamasari dan I. S. Suwardi, "Rule-based Part of Speech Tagger for Indonesian Language," *IOP Conference Series: Materials Science and Engineering*, vol. 407, 2018.

[4] T. Nomponkrang dan C. Sanrach, "The Comparison of Algorithms for Thai-Sentence Classification," 2016.

[5] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," 2017.

[6] J. Gimenez dan L. Marquez, "SVMTool: A general POS tagger generator based on Support Vector Machine," 2004.

[7] J. Gimenez dan L. Marquez, "Fast and Accurate Part-of-Speech Tagging: SVM Aproach Revisited," 2004.

[8] T. Nakagawa, T. Kudo dan Y. Matsumoto, "Revision Learning and its Application to Part-of-Speech Tagging," 2002.

[9] F. R. A. L. R. M. Arawinda Dinakaramani, “Designing an Indonesia Part of Speech Tagset and Manually Tagged Indonesia Corpus,” 2014.

[10] M. Nazir, Metode Penelitian, Bogor: Ghalia Indonesia, 2014.

[11] F. Saefulloh, “Part Of Speech Untuk Bahasa Indonesia Menggunakan Conditional Random Field(CRF),” 2017.

[12] A. L. Sari, “Coreference Resolution Dengan Menggunakan Metode SVM Pada Novel Bahasa Indonesia,” 2017.

[13] H. Ramza, M. S. Kadar, F. Abdurrahman dan M. S. Ab-Rahman, “Analisis Hubungan Imbuhan Me dan Ber pada Setiap Perenggan dalam Tulisan Cerita Pendek “Kembali Kasih”,” 2013.

[14] L. B. Durgesh K. Srivastava, “Data Classification Using Support Vector Machine,” 2005.

[15] A. N. Amalia, “Implementasi Support Vector Machine (SVM) Pada Klasifikasi Laporan Skripsi (Studi Kasus: Teknik Informatika UNIKOM),” 2016.

[16] N. Guenther dan M. Schonlau, “Support Vector Machine,” 2016.

[17] J. Milgram, M. Cheriet dan R. Sabourin, “"One Against One" or "One Again All": Which One is Better for Handwriting Recognition with SVMs?,” 2006.

[18] G. M. Foody dan A. Mathur, “A Relative Evaluation Of Multi-Class Image Classification By Support Vector Machines,” 2004.

[19] B. P. Utama, “Support Vector Machine Dalam Sistem Pendeteksi Kepribadian Berdasarkan Pola Tanda Tangan,” 2017.