Implementation of Fuzzy K-NNC Algorithm for Classification of Thesis Proposal Based on Scientific Group in Informatics Engineering UNIKOM

Rozesh Eco Cevin Sagala[1], Kania Evita Dewi[2]

Teknik Informatika-Universitas Komputer Indonesia
Jalan Diaptiukur No.112-116 Bandung. 40132
Email : rozesh.cevin.sagala@email.unikom.ac.id[1], kania.evita.dewi@email.unikom.ac.id[2]

ABSTRACT

Yosep amin's research on the classification of thesis proposals using the fuzzy k-nearest neighbour (FK-NN) method still has a low accuracy, which resulted in an accuracy of 51% at K = 7 and K = 8. In a study conducted by Eko prasetyo on "Fuzzy K-Nearest Neighbors in every class for data classification", it shows that the FK-NNC method is better in comparison with the FK-NN method, so the FK-NNC method will be attempted to be implemented in the proposal classification Thesis in this research.

This research will perform a summary of the text first before the calcification process, which is to reduce the size of the existing matrix, after which the extraction results will be included in the data train or test data. Parameters used as assessment factor that is seen from background of thesis proposal.

Based on the results of testing of FK-NNC method in the thesis proposal classification obtained the greatest accuracy value that is at K = 8 and K = 10 with an accuracy of 77.7%. Therefore it is concluded that FK-NNC method can be used in classification of thesis proposal.

Keywords: classification, Fuzzy K-Nearest Neighbors, Fuzzy k-nnc, extraction text.

1. INTRODUCTION

Thesis is a closing course that must be taken by students / college students who are taking a strata-level education in college. In Informatics Engineering UNIKOM, students / students who will submit a thesis, first make a research or called a proposal. In the proposals they will ask, students / students must determine the group of scholars they will take in the thesis. Informatics Engineering UNIKOM has 5 (five) scientific groups that are group A, B, C, D and E. Each group will be handled by professors who are experts in their field. The development of science and technology today, making the variety of thesis themes proposed by students / coeds, thus resulting in greater irisian each group of scholars. This can sometimes lead to errors, when students will choose a group of scholarships that will be taken in the thesis. At the time students are wrong in choosing a group then the handling of thesis proposals such as input for direction in the work of the thesis will not be maximal, because thesis proposal is not handled by testers who are experts in the field. To help in identifying the thesis proposal, we can use the classification method.

Research relating to the classification of documents has been done before. Yosep Amin's research on "The Implementation of the Fuzzy K-nearest neighbor (Fuzzy K-NN) algorithm for the Classification of Thesis Proposals Based on Scientific Groups in Informatics Engineering UNIKOM", shows that the results of the research, analysis, training and testing algorithm of Fuzzy K-nearest neighbor For the classification of Proposal obtained the greatest accuracy value of 51% at K = 7 and K = 8 [1]. From the results of the research obtained the accuracy is still low, and also still has a matrix large enough so that other solutions are needed in order to improve the accuracy of the more optimal.

This research will use an algorithm Fuzzy K-Nearest Neighbor in every class (FK-NNC) in the classification of the thesis proposal and also keyword extraction to minimize test and training data matrix [2]. The FK-NNC method is the development of the Fuzzy K-Nearest Neighbour Method. Research conducted by Achmad Syaiful Aziz in the classification of types of steels based on chemical composition using Fuzzy k-Nearest Neighbor in every class show performed 3 (three) times the experiment using training data different obtain the level of data accuracy by an average of 96, 7% and average error rate of 3.3% [3]. Eko prasetyo research entitled "Fuzzy K-Nearest neighbor in every class for the classification of data" indicate where the FK-NNC method is better and accurate in classifying the data than the method of FK-NN and K-NN[4].Based on the explanation above, in this research will be implemented Fuzzy K-

NNC algorithm in solving the problem of classification of thesis proposal. Taken a case study in the study program Informatics Engineering UNIKOM.
The expected goal of this study was to determine the accuracy of Fuzzy K-Nearest Neighbor in every class within the classification proposal thesis in Informatics UNIKOM.

2. RESEARCH CONTENTS

2.1 Description system

In designing the system to be built, will be made an application to classify the proposal thesis in Informatics Engineering UNIKOM. Where on the program will be built will use fuzzy algorithm K-Nearest Neighbour in every clas. There are two stages in the classification that will dilakakukan the stage of training data and test data. Before the data will be trained or tested, the input data will be filtered / eliminated the less important point, where this is done so that the preprocessing data to produce a good output. After enter data is done preprocessing process then next will be done compression process sbesar 50%, this is done to minimize existing matrix.

2.2 Summary

A text summary is an activity of summarizing the text of a document, in which the purpose of this summary is to present the main idea in a document in less space / storage [5]. On the text summary there are two methods: manually and automatically. In automatic text summary there are two procedures that can be done that is extraction and abstraction. In the extraction process using a keyword extraction algorithm whereby this algorithm copies the text units that are considered important or most informative in the text to be summarized, whereas the abstraction is to construct an internal senabtical representation such as extracting the essence of the text, then creating a summary by creating Aru sentences that present the essence of the text. Text summary can be done on one document and multiple documents (multy document).

2.3 Keyword Extraction

In this keyword extraction stage, it is done to minimize the matrix that will be stored as training data and that will be used as test data. The summary step of the keyword extraction method consists of two stages: first using the key extraction algorithm, then the compression ratio selector [5]. In the first stage the document is given a benchmark with the keyword extraction algorithm and then calculates the TF / IDF as it has done before, ie at the calculation phase of the TF / IDF sentence. From the result of Tf / Idf, add up all TF / IDF for each sentence and from the result is sorted from the largest, once sorted then do the compression of 50% from the original document. After the selection of 50%, the next process is to rearrange one sentence from the largest to the lowest, and this new text with the composition of the new sentence will be the final text that is summarized Compression is done as much as 50% because in the previous study, obtained more accuracy Both compared with compression of 20 and 30%. [5].

2.4 Term Weighting/ TF-IDF

Stage Term weighting / TF-IDF is the method used in calculating the weight of each word. The most commonly used method is using TF-IDF (Term Frequency - Inverse Document Frequency) method. In this method, the calculation of term weight (word) in a document is done by calculating the value of Term Frequency (TF) and Inverse Document Frequency (IDF) by multiplying it [6]. Term Frequency (TF) is to show the number of words that often appear in the document. The number of occurrences of words in the document will affect the weight of the word. Inverse Document Frequency (IDF) is a reduction in term dominance that often appears in various documents. Sometimes a lot of words appear in the document, is a word that is not important, but the word appears on a few documents can be said as a more important word. Weighting will take into account the inverse document frequency (inverse document frequency) frequency factor. To calculate the weight of words (Wi) in a document is calculated using the formula:

$$W_i = TF \times IDF \qquad (2.1)$$

where :
W_i *= The weight of words in a document.*
i *= index term .*
TF *= The number of words that appear in a document.*
IDF *= inverse document frequency from term.*

To calculate the IDF can use the formula:

$$IDF = log\left(\frac{D}{DF}\right) \qquad (2.2)$$

where :
IDF *= The result of the calculation of the total number of words divided by the number of documents contained words in the search.*
D *= Number of existing documents.*
DF *= Number of documents containing the word searched.*

2.5 Fuzzy K- Nearest Neighbour in Every Class

The Fuzzy K-Nearest Neighbor in every Class (FK-NNC) method is a classification method using the nearest neighbor K on each class that exists from a test data and will be added to the neighboring class x found by the FK-NNC method [7] . The FK-NNC method is the development of the FK-NN method, in which the FK-NNC method develops by

doing the nearest Cxk neighbor. Nevertheless, the framework of the FK-NNC still uses the FK-NN framework where the membership value is interrupted [0.1] where the total membership value of a data in all classes equals 1, as in the following formula 2.3 :

$$\sum_{j=1}^{c} \boldsymbol{U_i} = \mathbf{1},\ \mathbf{0} \leq \boldsymbol{U_j} \leq \mathbf{1} \qquad (2.3)$$

Where :
:
U_i = Value of the training data membership.
U_j = The value of the UI test data to the j^{th} class.

In the next stage will be calculated the distance between data testing with data traning, this is done to calculate the proximity anatar data to be tested and train data. To calculate the distance will be done by using euclidean distance method. The following is an equation for assigning a membership value to the Wj test data by using the euclidean distance method. The distance of eucidean distance between the test data vector Wj with the Wii data train vector is calculated using the following formula:

$$\|\boldsymbol{W_i} - \boldsymbol{W_j}\| = \left(\sum_{i=1}^{n} |\boldsymbol{W_i} - \boldsymbol{W_j}|^2\right)^{1/2} \qquad (2.3)$$

Where :

$\|W_i - W_j\|$= Euclidean distance of test data vector and train document.
W_j = Weighting results TF-IDF summary test data.
W_i = The weight of the TF-IDF results summary of the i-th exercise data. i = index term ke-i.
n = Total sum of result of preprocessing text.

After obtained the distance between the test data and training data with eucidean distance method then the next calculation of accumulated test distance to each class. Where for each class will be multiplied by the class of the neighboring K obtained. Furthermore the Uj test data distance to all neighboring K's of each class will be summed, by the formula:

$$\boldsymbol{S_i} = \sum_{j=1}^{K} \boldsymbol{d}(\boldsymbol{W_i}, \boldsymbol{W_j})^{-2/m-1} \qquad (2.4)$$

Dimana:
S = Cumulative distance ke-i.
i = 1, 2, 3, 4, 5.
d = Accumulated distance of test data Wj to neighbor in class j as much as C class.

m = Weight rank.
If the distance calculation using Manhattan method then m = 1.
If on distance calculation using Euclidean method then m = 2.
On distance calculation using Chebyshev method m = ∞ then.

Furthermore, after obtaining the accumulated test distance to each class will be the sum of the accumulated test distance to all existing classes. Where the accumulated distance of the Wj test data to all classes is denoted by D, the formula is as follows:

$$\boldsymbol{D} = \sum_{i=1}^{c} (\boldsymbol{S_i}) \qquad (2.5)$$

Where :
D = The number of accumulations of all classes.

The next stage is to calculate the value of membership. In order to get the value of the Wj test data membership on all i-th classes (there are C classes), using the following formula:

$$\boldsymbol{U_i} = \frac{\boldsymbol{S_i}}{D} \qquad (2.6)$$

where :
U_i = Membership of k-i class.

Then to determine the predicted data class Wj test data, selected the class with the largest membership value of data Ui. By the formula:

$$\boldsymbol{y'} = \boldsymbol{arg\,max}(\boldsymbol{U_i}) \qquad (2.7)$$

2.6 Analisis System

The system to be built has several stages, such stages as the stage of training and testing phase. In the training and testing phase is divided into several stages again that is preprocessing stage, kemdian stage weighting the value of the word, the text extraction phase and the test phase. An overview of the applications to be built can be seen in the following figure.

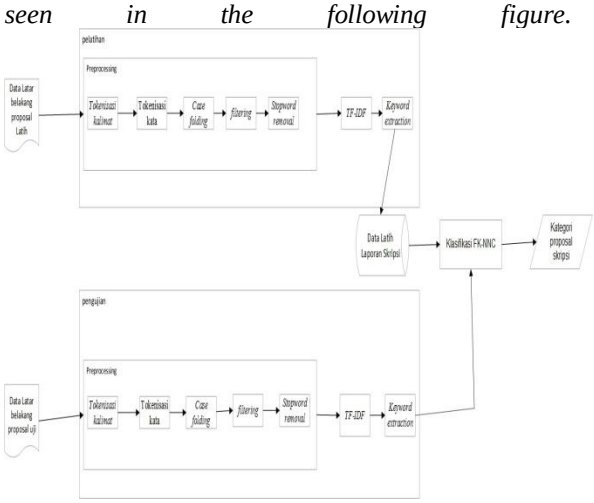


Figure 1. System overview

***2.6** Input Data Analysis*

The input data that is inputted is a background document of a thesis proposal which is in .txt format, where on each input data will be done to check and deletion of point, this is done so that the input data can produce good preprocessing result, location of point deleted as in word PT , CV, jl, no, percent and degree. In the data input proposal of this thesis, there are 5 classes of scientific groups, namely class A, B, C, D, and E. There are 55 data propsal background documents will be used as train data and 18 documents as will be used as test data, 73 input data.

2.6 Training Stage Analysis

The training stage is conducted to obtain learning data that will be used for comparison data on the classification process in the testing phase. This stage is included in the preprocessing stage, the TF-IDF weighting stage and the keyword extraction stage.

2.6.1 Preprocessing

Preprocessing is the process of processing the original data before being processed in the next stage. Enter the initial data in the form of background documents proposal thesis student informatics engineering UNIKOM. Preprocessing stages in this study include: tokenisasi sentence, tokenisasi word, case folding, filtering, and stopword removal The following Preprocessing steps performed can be seen in Figure 2.

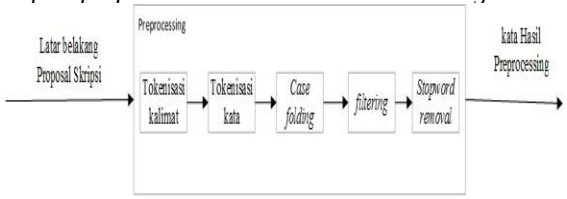


Figure 2. Preprocessing

For the preprocessing phase the data will be described in the following steps.

a. Sentence Tokenization Phase

Tokenisasi kaliamat used to separate the text into sentence unit form. The sentence phenomenon in this system uses a delimeter (".). The following steps can be seen in Figure3.



Figure 3. tokenisasi kalimat

b. stage tokenezing

This stage checks the first character up to the last character. When a delimeter character is found like a space, the text will be separated and inserted in the set of sentences. The following steps performed can be seen in Figure 4.

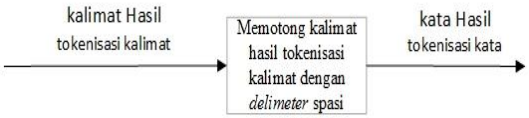


Figure 4. tokenisasi kata

c. stage case folding

The case-folding stage is the process whereby each word is converted into lowercase or all capital letters. In the case of case folding in this study, will do the conversion of each word to all lowercase letters. Following the steps performed, can be seen in Figure 5.

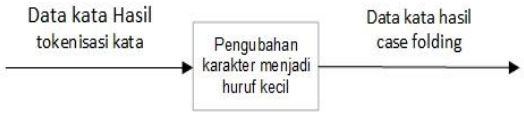


Figure 5. case folding

d. stage filtering

Filtering stage is a process where characters other than "a" to "z" and spaces will be omitted and only receive spaces. The following steps performed, can be seen in Figure 6.

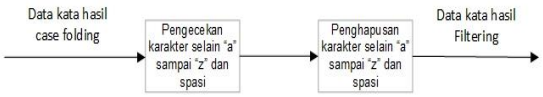


Figure 6. filtering

e. stage Stopword Removal

Is the phase of eliminating words that are less important or not describe the contents of the existing documents. In the Stopword Removal process that will be done, will use a stoplist from the tuning research, but there is addition and subtraction of words. Where this is done so that the important word in the research but contained in the stoplist contained in the dictionary of the tuning is not removed. The following steps performed can be seen in Figure 7.



Figure 7. Stopword removal

2.6.2 Analyze/ TF-IDF

The TF-IDF stage is performed for word weighting calculations (wi) in the document. The first step is to find the value of TF and IDF of each word, to get the value of IDF calculated by equation (2.2). Then after getting the value of TF and IDF then we calculate the weighting (wi), to get the

weight (wi) of each document on each word, This is done to find the weight of each word can be done with equation calculation (2.1). The following steps performed in the calculation of weighted words, can be seen in Figure 8.



Figure 8. TF-IDF

2.6.3 analyze Keyword Extraction

This stage uses a summary of keyword extraction methods consisting of several stages: first using the key extraction algorithm and the compression ratio selector [10]. In the first stage, the document is given a benchmark with the keyword extraction algorithm and then calculates the TF / IDF as it has done before, ie at the calculation phase of the TF / IDF sentence. Then from the TF / IDF results, will add up all TF / IDF for every sentence. Then after the sum of each sentence (wi), then the result is sorted from the largest to the smallest. Once sorted, then do the compression of 50% of the original document. So from the compression results that have been done, data compression results will be stored in the database that will be used as training data.

2.7 Analyze Tahap Pengujian

At the testing stage, the steps performed for the word processing stage are the same as those done in the training phase, which include preprocessing process, word weighting/TF-IDF and keyword extraction. After obtained keyword extraction results in the form of text compression, then the compression results will be used as test data. Then to get the category it will perform the process of classification using FK-NNC method. The first process in the testing phase is to perform preprocessing of the test data entered, the preprocessing process is done the same as the training stage ie tokenisasi sentence, tokenisasi kata, case folding, filtering and stopword removal. Then after the results can be preprocessing it will be done the process of keyword extraction of test data and do the classification.

*2.7.1 Analyze **stage clasifition***

The classification stage is done after the TF-IDF weighting process is completed, it will be classified the test document using FK-NNC method. The following explanation of the steps in the FK-NNC algorithm for a test data, presented as follows:

1. Calculates the TF-IDF of the said test document against the training document (2.1).
2. Calculate the distance of the test document and train using the euclidean distence method (2.3)
3. Sort the distance of euclidean distence by ascending with a predetermined value of K.
4. Calculate the accumulated test distance of each class (2.4)
5. Calculate test distance to all neighboring k from each class (2.5)
6. Calculate the membership value of each class (2.6)
7. Selecting classification prediction results (2.7).

The above explanation is the stage that dilakuka FK-NNC method in classifying the test data to get the category results. The above process will be displayed in the form of images that can be seen in Figure 9.

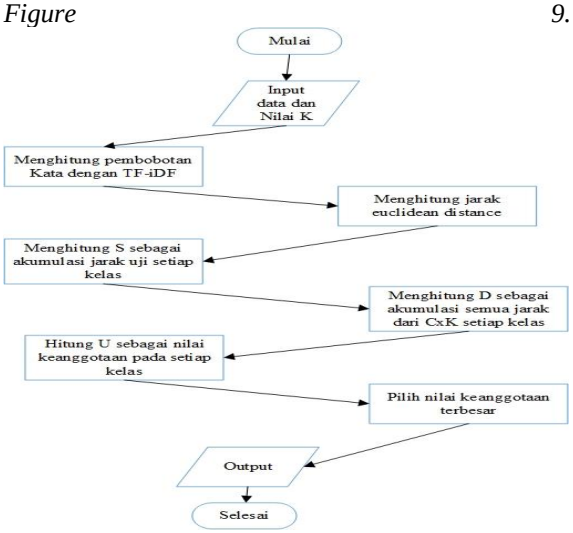


Figure 9. FK-NNC Stage

2.8 Testing Accuracy

Assessment of accuracy in research of skirpsi proposal classification using fuzzy K-Nearest Neighbors method in every class is done to know the accuracy value to be able to determine whether fuzzy K-Nearest Neighbors method in every class is suitable for classification of thesis proposal.

The amount of data used in this test is as much as 73 background data proposal thesis student of Informatics Engineering UNIKOM, where 55 data is used as training data and 18 data as test data.

In this test there are 11 data from each scientific group used as training data, where there are 5 groups of scholars. In this test it will also use the K value of K = 2 samapai with K = 10, then the following is the result of the background test of the scrispy proposal which is done by the fuzzy K-Nearest Neighbour method in every class.

a. Testing data with value K =2.

Tabel 1. confusionmatrix Testing with K =2

Kelompok keilmuan		PERKIRAAN				
		A	B	C	D	E
Tujuan	A	**1**	0	0	0	0
	B	0	**1**	0	0	0
	C	2	1	**3**	0	0
	D	0	1	1	**2**	1
	E	1	1	0	2	**1**

$$accuracy = \frac{8}{18}\ x\ 100\%$$
$$accuracy = 44{,}4\%$$

b. Testing data with value K =3.

Tabel 2. confusionmatrix Testing with K =3

Kelompok keilmuan		PERKIRAAN				
		A	B	C	D	E
Tujuan	A	**1**	0	0	0	0
	B	0	**1**	0	0	0
	C	2	0	**2**	1	0
	D	0	1	2	**2**	0
	E	0	2	0	2	**1**

$$accuracy = \frac{7}{18}\ x\ 100\%$$
$$accuracy = 38{,}8\%$$

c. Testing data with value K =4.

Tabel 3. confusionmatrix Testing with K =4

$$accuracy = \frac{11}{18}\ x\ 100\%$$
$$accuracy = 61{,}1\%$$

d. Testing data with value K =5.

Kelompok keilmuan		PERKIRAAN				
		A	B	C	D	E
Tujuan	A	**1**	0	0	0	0
	B	0	**1**	0	0	0
	C	3	0	**1**	2	0
	D	0	0	1	**4**	0
	E	1	1	0	1	**2**

Tabel 4. confusionmatrix Testing with K =5

$$accuracy = \frac{9}{18}\ x\ 100\%$$
$$accuracy = 50\%$$

e. Testing data with value K =6.

Tabel 5. confusionmatrix pengujian dengan K=6

$$accuracy = \frac{9}{18}\ x\ 100\%$$
$$accuracy = 50\%$$

f. Testing data with value K =7.

Tabel 6. confusionmatrix Testing with K =7

$$accuracy = \frac{12}{18}\ x\ 100\%$$
$$accuracy = 66{,}6\%$$

g. Testing data with value K =8.

Tabel 7. confusionmatrix Testing with K =8

Kelompok keilmuan		PERKIRAAN				
		A	B	C	D	E
Tujuan	A	**1**	0	0	0	0
	B	0	**1**	0	0	0
	C	4	0	**1**	1	0
	D	0	0	1	**4**	0
	E	1	1	0	0	**2**

Kelompok keilmuan		PERKIRAAN				
		A	B	C	D	E
Tujuan	A	**1**	0	0	0	0
	B	0	**1**	0	0	0
	C	1	0	**4**	1	0
	D	0	0	0	**5**	0
	E	1	0	0	1	**3**

$$accuracy = \frac{14}{18}\ x\ 100\%$$
$$accuracy = 77{,}7\%$$

h. Testing data with value K =9.

Tabel 8. confusionmatrix Testing with K =9

Kelompok keilmuan		PERKIRAAN				
		A	B	C	D	E
Tujuan	A	**1**	0	0	0	0
	B	0	**1**	0	0	0
	C	2	0	**3**	1	0
	D	0	0	0	**5**	0
	E	0	0	1	0	**4**

$$accuracy = \frac{13}{18}\ x\ 100\%$$
$$accuracy = 72{,}2\%$$

i. Testing data with value K =10.

Tabel 9. confusionmatrix Testing with K =10

Kelompok keilmuan		PERKIRAAN				
		A	B	C	D	E
Tujuan	A	**1**	0	0	0	0
	B	0	**1**	0	0	0
	C	2	0	**3**	1	0
	D	0	0	0	**5**	0
	E	0	0	1	0	**4**

$$accuracy = \frac{14}{18}\ x\ 100\%$$
$$accuracy = 77{,}7\%$$

3. Closing Conclusion

3.1 conclusion

Based on the result of research that has been done, it can be seen that the classification of thesis proposal by using Fuzzy K-Nearest Neighbors in every class (FK-NNC) method has better accuracy than using Fuzzy K-Nearest Neighbour method. Testing of classification of thesis proposal by using FK-NNC method with 73 data, where 55 data used as training data and 18 data as test data, yielding 77,7% biggest accuracy at K = 8 and K = 10.
Accuracy results can be influenced by the use of appropriate K values in each class and the same words that the test data has with the train data.

3.2 Suggestion

From the existing limitations, the suggestion for the development of Fuzzy K-Nearest Neighbors in every class (FK-NNC) algorithm on the thesis proposal classification is for further research to use the help application to perform automatic checking on typing or in the elimination of unnecessary point , So that on the train data or test data to produce a better preprocessing process.

DAFTAR PUSTAKA

[1] Amin, Yosep. 2016. "Implementasi Algoritma *Fuzzy K-Nearest Neighbour (Fuzzy K-NN)* Untuk Klasifikasi Proposal Skripsi Berdasarkan Kelompok Keilmuan Di Teknik Informatika UNIKOM*".Skripsi*. Universitas Komputer Indonesia.

[2] Darmawan, Rizki; Wahono, Romi Satria;, "Jornal of Intelligent System," *Hybrid Keyword Extraction Algorithm and Cosine Similarity for Improving Sentences Cohesion in Text Summarization*, vol. 1, Desember 2015.

[3] Rosyid, Harunur; Prasetyo, Eko; Agustin, Soffiana;, "seminar nasional teknologi informai dan multi media," *PERBAIKAN AKURASI FUZZY K-NEAREST NEIGHBOR IN*, Januari 2013.

[4] Achmad Syaiful Aziz. 2017. "*APPLICATION OF CLASSIFICATION OF STEEL BY USING CHEMICAL COMPOSITION FK-NNC (FUZZY K-NEAREST NEIGHBORIN EVERY CLASS)*".*Skripsi.* Universitas Muhammadiyah Gresik.

[5] Darmawan, Rizki; Wahono, Romi Satria;, "Jornal of Intelligent System," *Hybrid Keyword Extraction Algorithm and Cosine Similarity for Improving Sentences Cohesion in Text Summarization*, vol. 1, Desember 2015.

[6] Suliantoro, Deddy Wijaya;, *INTEGRASI PEMBOBOTAN TF IDF PADA METODE K-MEANS UNTUK*, April 2012.

[7] Kembang Rinci Hapsari and Yunus Juli Santoso, "Prosiding Seminar Nasional Manajemen Teknologi XXII," *STEMMING*

ARTIKEL BERBAHASA INDONESIA DENGAN PENDEKATAN CONFIX-STRIPPING, Januari 2015.