Implementation Of Fuzzy K-Nearest Neighbour (fuzzy K-NN) For Classification Of Proposals Thesis Based On A Group Of Scholarly In Informatics Engineering UNIKOM

Yosep Amin

Informatic Engineering – Indonesian Computer University
Dipatiukur Street 112-114 Bandung
Email : yosep.amin@gmail.com

ABSTRACT

Thesis is a cover subjects taken by students who is currently studying education strata 1 at public universities or private universities, the thesis became one of the requirements for graduation. As science every time create more diverse themes that emerged thesis, causing their diversity thesis proposal. So that the academic authorities to make a grouping of the themes of the thesis Nevertheless there still misclassification thesis proposal. Factors triggering one of a select group of science because science is still much to do a group with other scientific groups. To identify a thesis proposal can be done by utilizing the method of classification.

In this research method fuzzy k-nearest neighbour (FK-NN) is implemented for the classification of thesis proposal by the scientific group. By doing the previous word processing and using TFIDF as weighting method so that an input word can be processed according to the needs. Parameters used as a factor in the assessment visits of the title and description of the thesis proposal.

Based on test results obtained value of the accuracy of the biggest ie with a percentage of 51% on the value of K = 7 and K = 8, and therefore can be concluded that in a study conducted grades K that produce the greatest accuracy that the value of K = 7 and K = 8.

Keywords: proposals skripsi , fuzzy k-nearest neighbour , tf-idf , classification.

1. INTRODUCTION

1.1 Background

Thesis is a cover subjects taken by students who is currently studying education strata 1 at public universities or private universities thesis became one of the requirements for graduation. Thesis is a term used in Indonesia to illustrate a scientific paper in the form of exposure to undergraduate research results S1 posts that discuss problems / phenomena in a given field of science by using the rules that apply [1]. Writing / scientific research will be reported into the thesis report, prior to the thesis, usually the student / student will make the first proposals for research or normally called a thesis proposal will be submitted to the committee hearing thesis proposal for trial.

As science every time create more diverse themes of the thesis that sprung proposed as a research theme, which is then poured in the thesis proposal, the difference theme thesis proposal made by students, raises their diversity thesis proposal both in terms of the background, purpose, methodologies, or research data. So that the academic authorities to make a grouping of the themes of the thesis, is done for the thesis proposal can be identified and help put the thesis proposal to the appropriate scientific groups. Nevertheless, there still misclassification thesis proposal, sometimes one of a select group of students in science. If this happens then the handling of the thesis proposal as input for the direction of the thesis workmanship will not be the maximum, because the thesis proposal is not handled by examiners who are experts in their field. Factors triggering one of a select group of science because science is still much to do a group with other scientific groups. To identify a thesis proposal can be done by utilizing the method of classification.

Fuzzy k-Nearest Neighbour is one method that is used for classification, this method was first developed by James M. Keller basis of this algorithm is the provision of membership value as a function of the values of a number of patterns within the set of k-Nearest Neighbour and provision of membership value Neighbour the class particular [2]. In studies that have been done relating to this case is titled Web Document Classification Based on Fuzzy k-NN Algorithm conducted by Zhang, Yi Niu, and Huabei nie [2]. Said the advantages of the method fuzzy k-NN is the level of accuracy that is better than the other classification methods such as k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM). In the study also said that the shortage of which is owned by Fuzzy k-NN, namely the classification lies in the speed slower than the above comparison of two methods of classification [2].

Based on the above explanation, in this study will be the implementation of fuzzy algorithms k-Nearest

Neighbour to solve the problems of classification of cases thesis proposal. Case studies taken in the course of Informatics UNIKOM.

1.2 Formulation of the problem

Based on the above background exposure can be formulated the problem, such as: How is the implementation Fuzzy k-Nearest Neighbour to classify the thesis proposal.

1.3 Purpose

purpose of this study was to implement Fuzzy k-Nearest Neighbour within the classification Thesis Proposal.

1.4 Objectives

Objectives of this study was to determine the accuracy of Fuzzy k-Nearest Neighbour within the classification thesis proposal.

1.5 Research methodology

The method used in this research is descriptive research method. Is a method in researching the status of a group human, an object, a set of conditions, a system of thought, or a class of events in the present. The purpose of this descriptive study was to create a description, picture or painting in a systematic, factual and accurate information on the facts, nature and the relationship between the phenomenon investigated. As well as having the characteristics of explaining the relationship, test hypotheses, make predictions, and get the meaning (conclusion) and the implications of a problem to be solved [3]. groove research to be conducted in this study can be seen in figure 1.1 as follows:

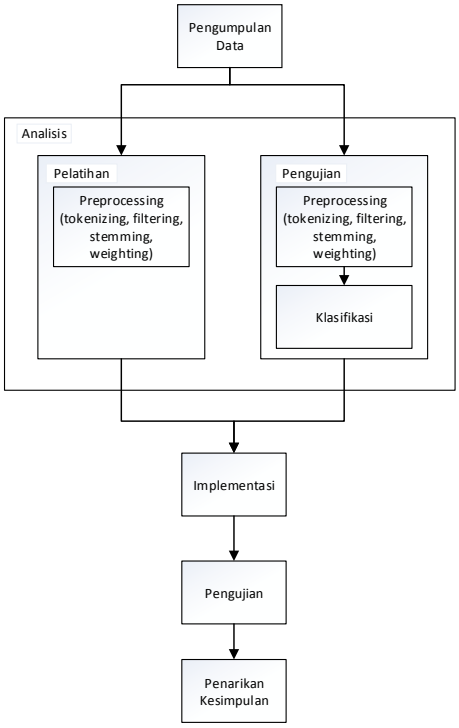


Figure 1 Flow Research

2.CONTENT OF RESEARCH

2.1 Theoretical basis

2.1.1 Thesis proposal

Thesis is a term used in Indonesia to illustrate a scientific paper in the form of exposure to writing research scholar S1 which addresses an issue / phenomenon in a given field of science by using the rules that apply [1]. Thesis proposal itself is a request, submission of ideas to others to obtain licenses topics, place of study, and others filed by a student in his attempt to make as a condition of graduation thesis as an undergraduate.

2.1.2 Text Mining

Text Mining is the application of concepts and data mining techniques to find patterns in the text, the process of analyzing the text in order to extract useful information for a particular purpose. The process of data mining for data or text documents require more stages, considering the text data has characteristics that are more complex than ordinary data. The goal is finding words - words that can represent the contents of the document so it can analyze the connectivity between document processing [7]. Steps - steps are performed on text mining in this study include text preprocessing, term weighting, and classification.

2.1.2.1 Text Preprocessing

In Phase text preprocessing do some process of preparing a thesis proposal's title and description in order to be a text document in phase ready to be

processed further, generally at this stage there are several processes, including: tokenizing, filtering, stemming.

2.1.2.2 Tokenizing

Tokenizing stage is the stage of deduction based on the input string that compose each word. Every sentence in the document will be compiled into a word-for-word, punctuation and hyphens will be abolished.

2.1.2.3 Filtering

Filtering stage is the stage to take the word - an important word from the token, in this process could use the algorithm stop list (discard the less important word) or word list (save important words). Stop list is a list of words that are often used and does not explain the contents of the document or stopword.

2.1.2.4 Stemming

Stemming stage is the stage where the word filtering the results will be converted into its basic form (Stem). For the word in English, stemming most popular method is to use the algorithm Martin porter's stammer or so-called porter stemmer. This algorithm can also be applied to languages other than English, in this case will be applied to Indonesian as well. Porter stemmer for Indonesian developed by Z. Fasillah Tala in 2003. Implementation of the porter stemmer for English Indonesian based Porter Stemmer developed by W.B Frake in 1992 [8].

2.1.2.5 Term Weighting

Term weighting stage is the stage of the weighting word resulting from the process of stemming. The method often used is using TF-IDF (Term Frequency - Inverse Document Frequency). In this method, the weight calculation terms (words) in a document is done by multiplying the value of the Term Frequency Inverse Document Frequency.

Term Frequency (TF) is a determining factor weighting term in a document based on the amount they appear in the document. The value of the number of occurrences of a word (term frequency) taken into account in assigning weights to a word. The greater the number of occurrences of a term in the document, the greater the weight of the document or the suitability will provide greater value.

Inverse Document Frequency (IDF) is the reduction of the dominance of terms that often appear in various documents. This is necessary because the terms that have appeared in various documents, can be considered as a general term (common term) so unimportant value. Conversely factor kejarangmunculan word (term scarcity) in a collection of documents should be considered in assigning weights. The word that appears on a few documents should be regarded as a more important word (Uncommon term) rather than the word that appears on many documents. Weighting will take into account factors inverse document frequency that contains a word (inverse document frequency). It was proposed by George Zipf. Zipf observed that the frequency of something tends to reverse proportionally with the order [2].

To calculate the weight of words (Wi) in a document is calculated using the formula:

$$w_i = TF(t_i, d) \times IDF(t_i) \quad (1)$$

Where:

w_i = the weight of the word (term) in the document d.

i = word index (term).

$TF(t_i, d)$ = number of words (term) t_i that appears in the document d

$IDF(t_i)$ = inverse document frequency of the word (term) t_i.

To calculate the IDF (t_i) can use the formula:

$$IDF(t_i) = \log\left(\frac{D}{DF(t_i)}\right) \quad (2)$$

Where:

$IDF(t_i)$ = inverse document frequency of the word (term) t_i.

D = the amount of the overall document.

$DF(t_i)$ = number of documents that have the word (term) (t_i).

2.1.2.6 Fuzzy K-nearest neighbour

This algorithm provides a membership value as a function of the distance of a number of the set of K-nearest neighbor and the neighbor granting membership value in a particular class. So the algorithm testing the data to be classified will have a value of membership in all classes. The classification algorithm will select the class membership value in the data the highest testing [2].

Here's a formula to calculate the distance between the document testing and training documents using euclidean distance:

$$\|x - x_j\| = \left(\sum_{i=1}^{n} |N_i - N_j^i|^2\right)^{1/2} \quad (3)$$

Where:

$\|x - x_j\|$= jeuclidean distance data vector testing And k-j training documents.

Ni = the weight of the term in the document testing

N_i^j = weighting term in the jth training document.

i = i-th index term.

n = number of terms the overall results of text preprocessing.

To calculate the value of the data testing membership in each class, as follows:

$$\mu(x) = \frac{\sum_{i=1}^{k} \mu_{ij} \left(\frac{1}{\|x - x_j\|^{2/(m-1)}} \right)}{\sum_{i=1}^{k} \left(\frac{1}{\|x - x_j\|^{2/(m-1)}} \right)} \quad (4)$$

Where:

$\mu(x)$ = the value of the i-th class membership on The data testing x.

M = weight rank (weight exponent)

μij = value of the i-th class membership on the Neighbors j.

To calculate the value of the i-th class membership on the neighbor j is calculated using the following formula:

$$\mu ij = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \quad (5)$$

Where:

μij = value of the i-th class membership on neighbors all j.

I = class genre.

J = nearest neighbor (K).

2.1.2.6 Fuzzy K-nearest neighbour

Confusion matrix is a calculation method that is used to search for accuracy in the classification results. Here is the confusion matrix table:

Table 1 *Confusion Matrix*

Kelas		Prediksi	
		1	0
Kelas sebenarnya	1	TP	FN
	0	FP	TN

Information:

a. True Positive (TP), a number of documents from grade 1 which was classified as grade 1.

b. False Positive (FP), a number of documents from grade 0 incorrectly classified as Class 1.

c. False Negative (FN), a number of documents from one class incorrectly classified as grade 0.

d. True Negative (TN), a number of documents of class 0 is correctly classified as class 0. [12]

Then it will be calculated accuracy, the accuracy of the calculation of the number of correct predictions and distributed by the number of predictions made with the purpose to find out what percentage of the success of a system of classification, used to calculate the accuracy of the following formula:

$$akurasi = = \frac{\text{THE NUMBER OF CORRECT PREDICTIONS}}{AROUND\ THE\ TOTAL\ PRODUCTION\ PERFORMED.} \times 100\ \% \quad (6)$$

2.2 Analysis and Design System

2.2.1 Proces Analisys

In the analysis process will be dijelas process - any process that is used for word processing and can later be used for the classification process. In general, the process can be seen in the following figure:

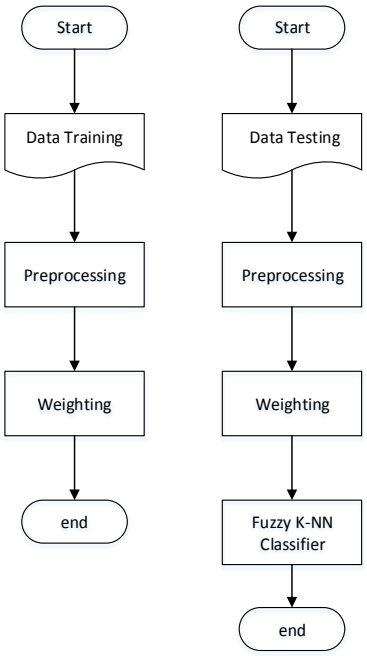


Figure 2 Process Flow Analysis

2.2.2 Classification Process Analysis

At this stage it will be determined how close connectedness between words - said to data on the training data as baseline data that serves as a testing information and data that serves as the data to be searched for conformance to training data. This study used an algorithm fuzzy k-nearest neighbor (FK-NN). Fknn combines capabilities and algorithms fuzzy k-nearest neighbor. The ability of fuzzy taken that membership function, with this function will map an element x in the universe of membership using a form of function theory, in order to obtain the possibility of membership value is closer dann as a function of the distance / similarity, and algorithms k-nearest neighbor will serve to measure how close distance from the membership value that no funds will choose the value of membership in the data the highest testing.

For more details, step - step method of fuzzy k-nearest neighbor is as follows:

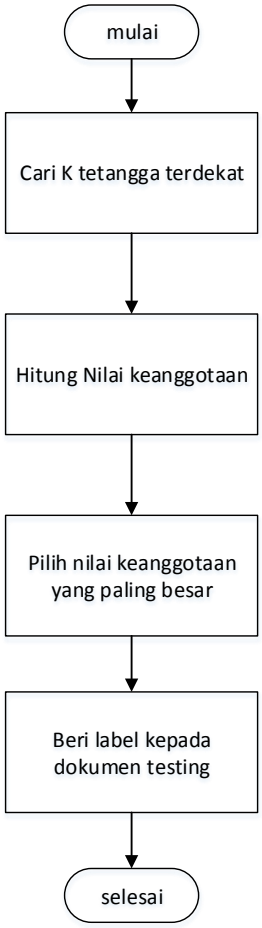


Figure 3 FKNN Flow

2.2.3 Data analysis

The data analysis consisted of analysis of data input and output data analysis, sample data such as document title and description thesis proposal have been provided previously.

2.2.3.1 Input Data Analysis

`Data input in the form of a document that contains the title and description of the thesis proposal. Then we will perform an tokenizing to break a string in the document being said - said constituent and eliminate punctuation, then the results will go to the next process which is the process of filtering is the process whereby words - words that are less important will be removed, and then after filtering, process which will be done next is stemming process, the purpose of this process is to get a basic form or basic word filtering of results. After that will go to the next stage. the term weighting stage, at this stage will be weighted words. This is done as a provision for the classification process.

2.2.3.2 Output Data Analysis

Data output of the classification of the document name thesis proposal, the data obtained from a series of processes that have been done earlier, from preprocessing stage, the stage of term weighting, and the training, testing. So from the process - a process that is performed classification with fuzzy k-neareast neigbor do then produced a document classified.

2.2.3 Analysis Methods

The analysis method is an explanation will be made, in the form of stage-by-stage are there to resolve the issue of the document classification proposal this thesis, the analysis algorithm that will be investigated in the classification of thesis proposal based on the title and description is to analyze how the fuzzy k-nearest neighbor in classifying documents thesis proposal based on the title and description. Before you can classify, input documents must go through several stages, namely the training phase and testing phase.

2.2.3.1 Analysis of Training Phase

The training stage is done to get the learning data that is useful for benchmarking data on the classification process in the testing phase. Training phase consists of the preprocessing stage, the stage of term weighting.

2.2.3.2 Analysis Testing Phase

In the testing phase. Steps - steps taken for the same word processing stage as the training phase which includes process text preprocessing (tokenizing, filtering, stemming) and weighting. Documents testing to be classified must pass through the process. After the text preprocessing and weighting process has been done, will be done the classification process. Step-steps are as follows:

1. Find the nearest K neighbors
   a. Calculate the distance testing documents with each document training, using the formula 3.
   b. Sort the results by ascending
   c. download the appropriate amount of data within the document K.
2. Membership Value Calculation
   a. Calculate the value of membership using formula 4.
3. Select Value Largest Membership
4. Add a Label on testing documents.

2.2.4 Analysis of Non-Functional Requirements

Non-functional requirements needed for the implementation of the algorithm fuzzy k-nearest neighbor (FK-NN) for classification systems into a thesis proposal consists of two things namely the need for hardware and software.
a. Hardware Requirement

Specifying the hardware used to build this system are as follows:
1. Processor Intel Core i3 2.2 GHz
2. Ram 2 GB
3. The hard drive capacity of 500 GB
4. VGA Nvidia GeForce 1GB

b. Software Requirements
The software used in building this system are as follows:
1. The operating system windows 7 home premium
2. Visual Studio 2010

2.2.5 Functional Requirements Analysis

Analysis of functional requirements is a step - a step that will identify the processes running in the system to be developed and explain the necessary requirements. So that the system can run as expected. Stages of modeling in the analysis as follows:

2.2.5.1 Entity Relationship Diagram

Modelling the initial database for Classification system thesis proposal using fuzzy K-nearest Neighbor built are as follows:

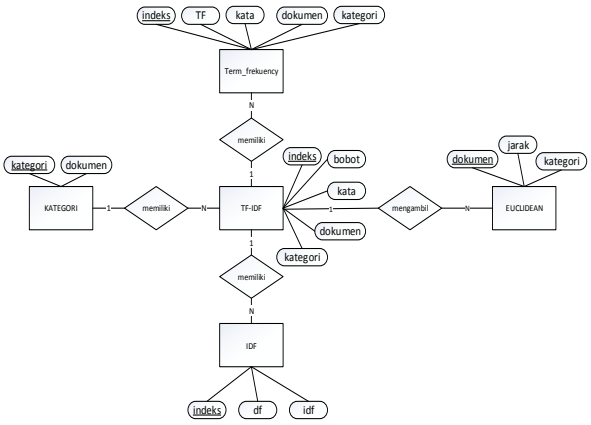


Figure 4 ERD

2.2.5.2 diagram Context

Context diagram describes the system that will be created as a single entity that interacts with people or other systems.

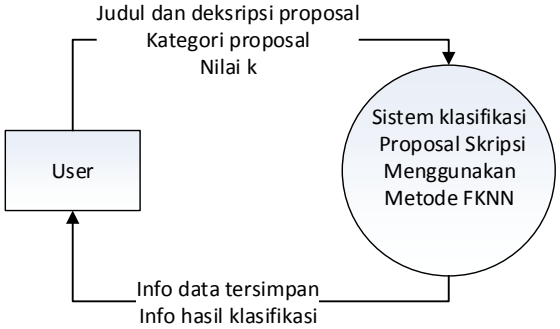


Figure 5 diagram Context

2.2.5.3 Data Flow Diagram

DFD is a graphical representation that illustrates the flow of information and information transformation that is applied as data flowing from input to output. DFD can be seen in Figure 6.

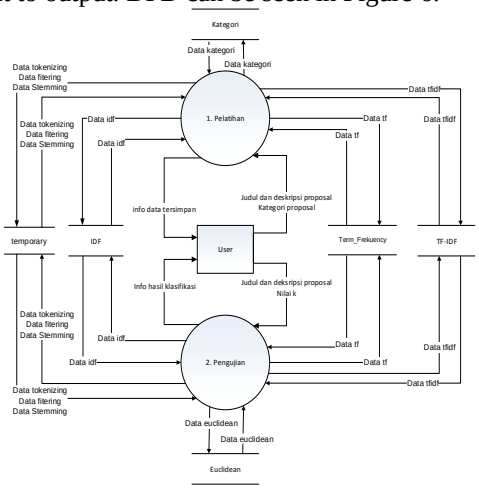


Figure 6 DFD

2.3 Performance Testing

Performance testing is the stage that has the aim to determine the performance of the classification method used in a system built the fuzzy k-nearest neighbor. Performance testing in this research is to measure the level of accuracy, the test methods used in this study the method of confusion matrix.

In this study, testing is done by using one of the methods. Here is the test to be performed:

1. Test the document that are not included into the training data

This testing is done by testing data is not included in the database, this test aims to determine the level of recognition of test data outside the database on the training data contained in the database. Training data are used as much as 150 and as many as 67 pieces of test data, the value of K be the contents of a value of 6 to 30. The result can be seen in Figure 7, as follows:

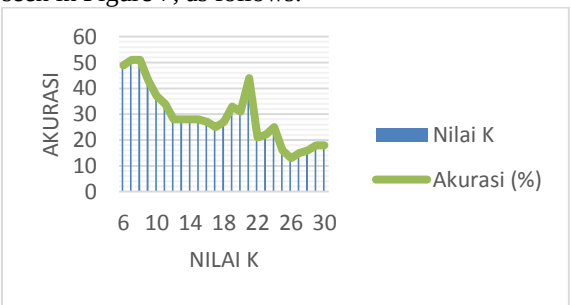


Figure 7 performance

Based on the results of the test scenario 1 is test of test data that is not contained in the training data with the amount of training data as many as 150 pieces and the value of K which starts from 6 to 30, the method fuzzy k-nearest neighbor may classify the value of the percentage of accuracy greatest obtained at K = 7 and K = 8 ie by 51%, to K = 6 to 49%, K = 9 by 43%, K = 10 by 37%, K = 11 by 34%, K = 12,13,14,15 at 28 %, K = 16.18 by 27%, K = 17 at 25%, K = 19 to 33%, K = 20 to 31%, K = 21 at 44%, K = 22 to 21%, K = 23 for 22%, K = 24, 25%, K = 25, 16%, K = 26, 13%, K = 27 to 15%, K = 28, 16%, K = 29.30 by 18%. The accuracy of the results can be affected by the use of appropriate K value in each - salty class, the amount of training data is used, and the similarity of words that are owned by test data with training data so that the difference distance difference between the test data and training data will be smaller.

3.CLOSING

3.1Conclusion

Based on the results of research, analysis, training and testing algorithms Fuzzy K-nearest neighbor classification thesis proposal by a group of knowledge in Information Engineering Digital, obtained the value of the accuracy of the biggest ie with a percentage of 51% on the value of K = 7 and K = 8, then from it can be concluded that in a study conducted grades K that produce the best accuracy in the value of K = 7 and K = 8.

3.2Advice

In making this final task, there are still many deficiencies that can be remedied for future development. Some advice that can be given is as follows:

1. Using the method of calculating the distance such as cosine similarity.
2. Using other stemming algorithms such as algorithms Nazief & Adriani for stemming Indonesian for porter stemming algorithm not show accurate results.
3. Use the application help to justify writing the wrong (tipo) contained in the training data or test data in order to make the results of tokenizing, filtering, and stemming better.
4. Using the training data are many and varied for each class. So the percentage the better classification accuracy.
5. Description thesis proposal described as scientific groups.

4.REFERENCES

[1] Farid Hamid,S.Sos.,M.Si., dan Drs.A.Rachman,MM.,M.Si., Panduan Skripsi FIKOM UMB, hal 21

[2] Prasetyo, Dimas. Muflikhah, Laili. Ridok, Achmad. "*klasifikasi genre film berdasarkan judul dan sinopsis menggunakan fuzzy k-nearest neighbour (Fuzzy K-NN)*", Prodi ilmu komputer, Program teknologi informasi dan ilmu komputer, Universitas Brawijaya

[3] Nazir, Moh. Ph.D, "*Metode Penelitian*", Edisi Ketujuh, Bogor : Ghalia Indonesia, 2011.

[4] Pressman, Roger S, Ph.D,"*Rekayasa Perangkat Lunak*", Edisi 7,Yogyakarta: ANDI, 2012.

[5] S.A, Rosa, Shalahudin.M. "*Rekayasa Perangkat Lunak: Tertruktur dan Berorientasi Objek*". Bandung: Informatika. 2015.

[6] Sianipar, R., H. "*Panduan Praktis Pemrograman C# Bagi Pemula*". Yogyakarta: ANDI. 2014.

[7] Raymond J. Mooney. CS 391L: *Machine Learning Text Categorization.* University Of Texas at Austin, 2006.

[8] Gregorius, S. Gunawan, Ibnu. Yunono, Ferry. "*Algoritma Porter Stemmer For Bahasa Indonesia untuk pre-processing text mining berbasis Metode Market Basket Analysis*". Jurusan Teknik Informatika, UK.

[9] Lasmedi, Afuan."*Stemming Dokumen Teks Bahasa Indonesia Menggunakan Algoritma Porter*", Prodi Teknik Informatika, Fakultas Sains dan Teknik, Universitas Jendral Sudirman.

[10] Budiharto,W., and Suhartono, D.,"*Artificial Intelligence*", Yogyakarta : Andi, 2014.

[11] Keller JM, Gray MR, Givens J.A. A Fuzzy K-Nearest Neighbor .Algorithm IEEE Transaction on Systems, Man, and Cybernetics. Vol SMC-15. No 4. 1985.

[12] Prasetyo, Eko. *Fuzzy K-Nearest Neighbor In Every Class Untuk Klasifikasi Data*. Seminar Nasional Teknik Informatika. Jurusan Teknik Informatika Fakultas Teknologi Industri Universitaas Pembangunan Nasional Veteran Jawa Timur. Pp 57-60. 2012.

[13] Wisdarianto, Ardhy. Ridok, Achmad. Rahman, Arif, Muh. "*Penerapan Metode fuzzy k-nearest neighbour (Fuzzy K-NN) untuk Pengklasifikasian Spam Email*", Prodi ilmu komputer, Program teknologi informasi dan ilmu komputer, Universitas Brawijaya.

[14] W., Teguh, Hardika. Mardji. Furqon, Tanzil, M. "*Penerapan Metode fuzzy k-nearest neighbour (Fuzzy K-NN) untuk Diagnosa Penderita Liver*

Berdasarkan Indian Patient Dataset(ILPD)", Prodi ilmu komputer, Program teknologi informasi dan ilmu komputer, Universitas Brawijaya.

[15] M.,Selly, Yanita. Ridok, Achmad. Muflikhah, Laili."*Perbandingan K-Nearest Neighbor dan fuzzy k-nearest neighbour (Fuzzy K-NN) Pada Diagnosis Penyakit diabetes Melitus*", Prodi ilmu komputer, Program teknologi informasi dan ilmu komputer, Universitas Brawijaya.

[16] S., Amiratus, Rahmi. Ridok, Achmad. Muflikhah, Laili."*Penerapan Metode fuzzy k-nearest neighbour (Fuzzy K-NN) untuk menentukan Kualitas Hasil Rendemen Tanaman Tebu*", Prodi ilmu komputer, Program teknologi informasi dan ilmu komputer, Universitas Brawijaya

[17] Puspasari, Maslikha. Dewi, Candra. Rahman, Arif, Muh. "*Prediksi Tingkat Resiko Penyakit Jantung Koroner(PJK) Menggunakan Metode fuzzy k-nearest neighbour (Fuzzy K-NN)*", Prodi ilmu komputer, Program teknologi informasi dan ilmu komputer, Universitas Brawijaya.

[18] Lestiyanto, Yudha, Dony. "*Stemming Bahasa Indonesia Sebagai Media Belajar Siswa Sekolah Menggunakan Algoritma Porter*". Sistem Informasi. Universitas Dian Nuswantoro. Semarang. 2014.

[19] Agusta, Ledy, "*Perbandingan Algoritma Stemming Porter dengan algoritma nazief & adriani untuk stemming dokumen teks bahasa indonesia"* konferensi nasional sistem dan informatika 2009. Fakultas Teknologi Informasi. Universitas Kristen Satya Wacana. KSN&109-036. 2009.

[20] Herman; Andani Achmad; dan Amil ahmad ilham, "*klasifikasi dokumen naskah dinas menggunakan algorita term frequency-inversed document frequency dan vector space model*", balai besar pengkajian dan pengembangan komunikasi dan informatika makassar, kemkominfo & Jurusan elektro, prodi informatika, fakultas teknik, UNHAS.

[21] Indriani, Aida, "*Klasifikasi Data Forum dengan menggunakan Metode Naive Bayes Classifier*", Seminar Nasional Aplikasi Teknologi Informasi(SNATI) 2014, Program studi Teknik Infomatika, STMIK PPKIA Tarakan Rahmawati. ISSN: 1907-5022.