AUTOMATION OF FORMATION THE ABSTRACT INDONESIAN BASED ON THE OVERALL DOCUMENT USING TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF)

Arie Prima Anggara[1]

[1]Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
Email : arieprimaa@gmail.com

ABSTRACT

Abstract is a summary of the overall research documents so that readers know the purpose of the study. From the abstract, the reader can know the mind of the writer reports / scientific articles and know the general picture of the complete article. In forming an abstract paragraph in a document journal of each section would require finding short sentences and clear. Search sentence on each section in the document that the journal would deprive many authors report / scientific articles in abstract form with limited formation of the abstract should be brief, sections should be balanced and avoid long sentences.

On the problems it is constructed abstract formation system automatically applies the method Term Frequency – Inverse Document Frequency. TF-IDF method will seek a sentence on the subtitles introduction, analysis, and conclusions that will form the abstract sentence automatically with emphasis on the content of the sentence that is on the subtitles. Formation abstract automated analysis system that is built has three stages, namely the input journal, preprocessing, abstract formation. Input journal chosen in the form of a journal format doc. preprocessing is done in this phase is the breakdown of the sentence, case folding, filtering, tokenizing, synonim checking, and stopword. Hereinafter stage performing calculations using tf-idf and last stage of determining the outcome abstract cosine similarity method.

Based on test results, then do the testing phase of the system and manually. Then the distance between the results obtained manually by an average system differs only 2 to 4 sentences. 100 % accuracy with the percentage obtained by the journal to 7 and 9 while pesentase accuracy with the lowest 60 % earned by the journal to 1.

Keywords: Term Frequency – Inverse Document Frequency (TF-IDF), Artificial Intelligence, Automation Formation Abstract.

1. INTRODUCTION

Abstract is a summary of the entire research document that the reader knows the purpose of the study. Abstract is used as a bridge to understanding the description that will be presented in an essay typically reports or scientific articles mainly to understand the ideas of the problem. From the abstract, the reader can know the mind of the writer reports / scientific articles and know the general picture of the complete article. For that abstract maker should be able to represent the content of scientific articles covering the gamut of background, analysis / analysis, and the results / conclusions. [1]

In forming an abstract paragraph in a document journal of each section would require finding short sentences and clear. In the academic world, this short article is used by the institution / agency / organization of education as early information on a study when incorporated in journals, conferences, workshops, or the like. Search sentence on each section in the document journal that many would deprive authors of the report / scientific articles in abstract form with limited formation of the abstract should be brief, parts must be balanced and avoid long sentences. Making the abstract can be done automatically by using methods Term Frequency - Inverse Document Frequency (TF-IDF).

To be able to form sentences abstract automatically will be applied methods Term Frequency - Inverse Document Frequency (TF-IDF), as TF-IDF is a weighting measure how important a word in the document when viewed globally throughout the document. [5] In making abstracts of the research will look for subtitles is needed, namely in the introduction or background, analysis or analysis of the data and conclusions. Where TF-IDF will measure the similarity between two or more sentences in the sentences that are in the subtitles required. In measuring a sentence of similarity of each section in the document will use text mining techniques.

From the above problems, this study will apply the methods Term Frequency - Inverse

Document Frequency (TF-IDF) on the establishment of a paragraph abstract automatically.

1.1 Purpose and objectives

The purpose of the thesis is to make the system automatically abstract formation on the overall document research journals using Term Frequency - Inverse Document Frequency (TF-IDF). The purpose of this research is to manufacture good abstract sentence automatically to complete a scientific paper with a short and clear, thus helping to complete the report / scientific articles in the establishment of abstract phrases quickly and accurately.

1.2 Summary

The summary is an effective way to present a long essay in short form. Therefore, it makes a summary of a lengthy essay that can be compared to prune a tree so stay trunk, branches and twigs are important along with the leaves as needed, so it seems that the essence of each tree is maintained. In summary the beauty of the style of the language, illustrations, as well as detailed explanations omitted, while the juice is left tanap kerangaknya decoration. Although the shape is compact, but still retain the mind of the author and original approach [11].

In summarizing we recommend using a single sentence rather than compound sentences. Compound sentences shows that there are two or more ideas that are parallel. Where possible summarize the sentences into phrases, then phrase being said. Similarly, a series of ideas that penjang should be replaced by a central idea only. In general, a summary of the specified length of summaries finally, there is the word that sums up the total of one-half, one-third of total words, up to a tenth of the word. For articles which have a number of words that are very much of course the summary results of a tenth be the right choice. As for the article which he said did not amount too much then the summary results as much as half of total words is the right choice so that the summary results will still be read with the structure of the summary results good.

1.3 Abstract

Abstract is a summary of the contents of a scientific paper is intended to help the reader to easily and quickly view the purpose of writing. In the academic world, this short article is used by the institution / agency / organization of education as early information on a study when incorporated in journals, conferences, workshops, or the like. In the virtual world (internet), an abstract is used as a quick overview on a scientific paper / research to be read, as well as a "display" models display clothes to wear seen or tested before purchase. Furthermore, complete part of a research sold to those who are interested to get it.

To make the abstract, the things that need to be considered are as follows:

1. The structure of paragraphs

An abstract is written in a paragraph that explains the entire contents of text briefly and clearly. Writing does not do indentation in the first sentence of the paragraph. Single space is an option that is owned by the author to construct a sentence in the first paragraph. Deeper, sometimes a mentor Thesis / Thesis / Dissertation set up on the use of a specific typeface and size.

2.Number of words

Ideally, a paragraph consists of 150 to 200 words. However, consideration of the most appropriate number of words in writing Theses or dissertations usually rely on the judgments views supervisor (supervisor) who accompanies a student in his writing. A supervisor should not consider the number of words as the main reference for writing paragraphs, because the main part is precisely the content (content) paragraph.

3. Fill a paragraph

First, identify the focus of the research are described briefly so that readers understand what is observed by a researcher in the study. Second, the author should clearly describe the design of the research conducted in the process of finding answers or solutions to the issues raised in the research. Design steps to resolve this problem by students commonly known as the Research Methods. Third, then the author will explain its findings to the reader. Some researchers consider the findings do not necessarily need to open the discussion revealed that done because it will actually make content writing repetition. Clearly meant as part of the discussion of the research findings also parsed in conclusion. Fourth, the need for conclusion in an article also seen in an abstract still get a critical concern as the final part of the paragraph. In this section, sometimes a number of researchers insert recommendations were yet without discussion or long descriptions. Furthermore, it is not unusual an abstract filled by the name of the author and his counselors, moreover it is written in bold font

Google Translate for Business:Translator ToolkitWebsite TranslatorGlobal Market Finder.

1.4 Stages of *Text Mining*

At this stage of text mining, there are five steps:

a. Case Folding, that is, change all text to lowercase.

b. Filtering, removing characters that are not in use. As for the characters to be eliminated are: '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ',' , '' ',' - ',' / ',' { ','} ',' + ',' _ ','! ',' @ ',' # ',' $ ','% ',' ^ ',' & ',' * ',' (',') ','? ','

< ','> ',' [','] ',' | ',' ~ ',' ` ' , ';', ':', '=' and " 's" for English documents will be deleted unless the colon.

c. Tokenizing, which broke the existing sentences into words.

d. Synonim Checking, which replace the existing words that have the same meaning. Replacement of the word that is based on a database created by the author.

e. Stopwords, namely the process stowords this removal process is done is remove the less relevant word or words that do not have a meaning that is so important and related to that of the sentence in the document with a list of matching words stopword on the database to be tested. The database is used for both Indonesian and English as well as taken from a site that provides berbagain types stopword list for a variety of languages.

In general, the preprocessing stage there are five steps, namely Case Folding, Filtering, tokenizing, stemming, and stopwords. There is one step that the author did not use in the manufacture of this esktraktif document peringkasan application that processes stemming. Stemming is the process of removing the suffix of a word and turn it into a base word. The purpose of the removal process Stemming this is because the author has the opinion that in an extractive document peringkasan additive contained in a word has a different meaning. As examples of the sentence is: "You make a home for his wife", "Father made a cup of coffee by the mother", "father always be kind to neighbors". The word "create" in the first sentence and the third sentence has meaning father of an activity but on a two-word phrase "create" means you get a treatment. If the stemming process is still being done then it will get the difference between one word and another.

At this stage of filtering author adds character " 's" which will be used to remove punctuation for documents that use the English language. As an example sentence is: "that is the President's car" turned into "that is President car".

1.5 *Term Frequency – Inverse Document Frequency (TF-IDF)*

Methods Term Frequency-Inverse Document Frequency (TF-IDF) is a way of assigning weights to the relationship of a word (term) of the document. For a single document each sentence is considered as a document. This method combines two concepts for weight calculation, namely Term Frequency (TF) is the frequency of words (t) of the sentence (s). Document frequency (DF) is the number of sentences in which a word (t) appears. Frequency of words in a given document shows how important word in the document. [5]

The frequency of sentences containing the word indicates how common the word is in the test documents. The weight of the bigger words if often appear in a sentence and the smaller if it appears in many sentences. In this method the weighting of words in a document is done by multiplying the value of TF and IDF. Weighting is obtained based on the number of times a term in the sentence (TF) and the number of occurrences of the term in the entire sentence in the document (IDF). The weight of a term greater if those terms appear frequently in a document and the smaller if the term appears in many documents. IDF value of a term calculated using equation (1) below:

$$IDF = log(\frac{N}{Df}) \qquad (1)$$

information :
N = total number of words in the document
Df = Number of sentences that contain the keyword t

Calculating the weight (W) each sentence with equation (2) below:

$$W_{d.t} = TF_{d.t} * IDF_t \qquad (2)$$

with:
d = the sentence to-ds
t = word (term) ket
TF = Term Frequency
W = weight of the sentence to-d of the word (term) all t
IDF= *Inverse Document Frequency*

Then just do the sorting (sorting) the cumulative value of W for each sentence. Three sentences with the biggest W value used as a summary or as a result of the output of the automatic text peringkasan. [5]

1.6 Pengujian Akurasi

In this study, the evaluation method used is the method of calculating intrinsic F-Measure is based on the calculation of Precision and Recall that according to Zaman B. and E Winarko [12] is an evaluation standard for the calculation information retrieval. evaluation of information retrieval calculation using the Precision and Recall also be used in the evaluation of the calculation of the automatic text peringkas.

$$recall = \frac{\#summary\ sentence\ summary \cap system\ manual}{sentence\ summary\ manual} \qquad (4)$$

which, recall = success rate

Precision is the degree of precision of the results summary. Calculation precision can be seen in the following equation:

$$Precision = \frac{\#sentence\ summary \cap ingkasan\ manual\ system}{sentence\ summary\ of\ the\ system} \qquad (5)$$

dimana, *precision*= tingkat ketepatan

F-measure is a combination of recall and precision. F-measure calculation can be seen in the following equation:

$$F - measure = \frac{2 \times recall \times precision}{recall + precision} \quad (6)$$

2. CONTENTS OF RESEARCH

2.1 Problem Analysis

In forming a paragraph abstract of a document journal of each section would require a search phrase that clearly and concisely, the search phrase on each section in the document journal that many would deprive authors in forming abstract with limited formation of the abstract should be brief, parts must be balanced and avoid long sentences.

From the above problems, it takes methods Term Frequency - Inverse Document Frequency (TF-IDF) which will measure the similarity between two or more sentences in the sentences that are in the subtitles required.

2.2 Analysis of Feedback

The input data used in the form of a journal file with a .doc format. The document as a whole entered into the application, the application will be made later by the decomposition of the contents of the document in the form of a few sentences that refer to certain rules, such as taking only sentence in the introduction, analysis, and conclusions.

2.3 Analysis of Data preprocessing

Preprocessing analysis stage is used to analyze what data is actually required in the Term Frequency - Inverse Document Frequency (TF-IDF). The stages of prepocessing analysis of the data in this study are as follows:

1. Add File Data

First select the document first. File documents in select must be a file extension * .doc.

2. Data Extraction

After a journal file that is intended to get, then do text extraction process. In this process done taking the text of the journal with the help of the library in .net which Microsoft.Office.Interop.Word.

3. Decision Text

Making stage is a stage penggambilan sentence required text of the entire text contained in the journal. Text that is needed is a text in the Introduction or Background, Analysis or Data Analysis and Conclusions. The flow of the process undertaken to obtain the required texts can be seen in the figure below:

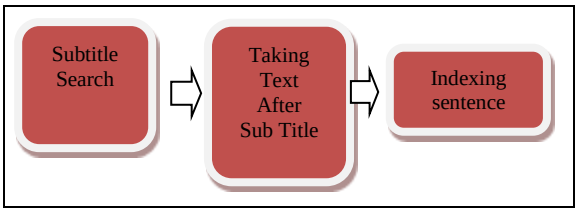


Figure 1. Flow Selection Sentence

a. Search Subtitle:
Searching for a title is needed, namely: introduction, analysis and conclusions.
b. Text retrieval:
Text retrieval index is after the title, namely the introduction or background, analysis, or analysis and conclusions will be saved.
c. Indexing sentence:
Having obtained the text above, the separation of each sentence that is bounded by the dot in each text.

4. Case Folding

At this stage, will be checked against the capitals that are in each sentence. If found the capital letters, it will be lower, that is, change to lowercase.

Table 1. Casefolding

	deskripsi kalimat
kalimat 1	di era otonomi daerah saat ini, sistem pemerintahan daerah sudah berbeda dibandingkan dengan sistem pemerintah diera orde baru.
Kalimat 2	kalau diera orde baru, organisasi pemerintah dan sistem informasinya ditentukan oleh pemerintah pusat, di era otonomi daerah ini pembentukan instansi pemerintah daerah termasuk sistem informasinya ditentukan oleh pemerintah daerah setempat
Kalimat 3	oleh karena itu sistem informasi pada setiap daerah bisa berbeda sesuai dengan perkembangan yang terjadi / kebutuhan di daerah masing-masing
Kalimat 4	pada awal otonomi daerah, pemerintah di daerah bisa membentuk dinas, badan dan lembaga tehnis sesuai dengan kebutuhan daerah setempat
Kalimat 5	adanya ketentuan ini membuat berbagai daerah membentuk dinas secara berlebihan untuk menampung sebanyak mungkin pejabat struktural
Kalimat 6	ketentuan mengenai pembentukan dinas dan lembaga tehnis tersebut kemudian disusul peraturan baru yang memberikan batasan jumlah dinas yang boleh dibentuk di pemerintah daerah.
Kalimat 7	bahkan penanganan informasi di suatu daerah cukup hanya dimasukkan dalam suatu seksi/bagian dari dinas dan setiap daeah menggunakan istilah yang berbeda seperti : hubungan masyarakat (humas) informasi komunikasi (infokom), badaninformasi komunikasi telematika (bikt).
…	…
Kalimat 21	3.kegiatan humas yang banyak berhubungan dengan unit kerja di pemda yang memiliki tingkat eselon yang lebih tinggi bisa menjadi hambatan bagi humas meski gengsi posisi humas cukup tinggi

5. Filtering

Removing special characters that are not used. In this test the entire punctuation characters such as exclamation points, question marks, quotation marks, and so forth will be deleted unless unless the dot. Steps to be done is to eliminate some of the characters that are not needed during the process perangkuman. As for the characters to be eliminated are: '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ',' , " ',' - ',' / ',' {','} ',' + ',' _ ','! ',' @ ',' # ',' $ ','% ',' ^ ',' & ',' * ',' (',') ','? ',' <','> ',' [','] ',' | ',' ~ ',' `' , ';', ':', '=' and " 's" for English documents.

Disappearance of characters such as numbers will be performed by the system, because the system can not recognize the description of the figures contained in the document.

Table 2. Filtering

	deskripsi kalimat
kalimat 1	di era otonomi daerah saat ini, sistem pemerintahan daerah sudah berbeda dibandingkan dengan sistem pemerintah diera orde baru.
Kalimat 2	kalau diera orde baru, organisasi pemerintah dan sistem informasinya ditentukan oleh pemerintah pusat, di era otonomi daerah ini pembentukan instansi pemerintah daerah termasuk sistem informasinya ditentukan oleh pemerintah daerah setempat
Kalimat 3	oleh karena itu sistem informasi pada setiap daerah bisa berbeda sesuai dengan perkembangan yang terjadi / kebutuhan di daerah masing-masing
Kalimat 4	pada awal otonomi daerah, pemerintah di daerah bisa membentuk dinas, badan dan lembaga tehnis sesuai dengan kebutuhan daerah setempat
Kalimat 5	adanya ketentuan ini membuat berbagai daerah membentuk dinas secara berlebihan untuk menampung sebanyak mungkin pejabat struktural
Kalimat 6	ketentuan mengenai pembentukan dinas dan lembaga tehnis tersebut kemudian disusul peraturan baru yang memberikan batasan jumlah dinas yang boleh dibentuk di pemerintah daerah.
Kalimat 7	bahkan penanganan informasi di suatu daerah cukup hanya dimasukkan dalam suatu seksi/bagian dari dinas dan setiap daeah menggunakan istilah yang berbeda seperti : hubungan masyarakat (humas) informasi komunikasi (infokom), badaninformasi komunikasi telematika (bikt).
…	…
Kalimat 21	3.kegiatan humas yang banyak berhubungan dengan unit kerja di pemda yang memiliki tingkat eselon yang lebih tinggi bisa menjadi hambatan bagi humas meski gengsi posisi humas cukup tinggi

6. Tokenizing

In the process of tokenizing this thing done is to break the existing sentences into words.

Table 3. Tokenezing

Kal1	Kal2	Kal3	…	Kal22
Era	Diera	Karena	…	Kegiatan
Ekonomi	Orde	Itu	…	Humas
Daerah	Baru	Sistem	…	yang
Saat	Organisasi	Informasi	…	Banyak
Ini	Pemerintah	Pada	…	Berhubungan
Sistem	Dan	Setiap	…	Dengan
Pemerintahan	Sistem	Daerah	…	Uniot
Daerah	Informasinya	Bisa	…	Kerja
Sudah	Ditentukan	Berbeda	…	Di
Berbeda	Oleh	Sesuai	…	Pemda
Dibandingkan	Pemerintah	Dengan	…	Yang
Dengan	Pusat	Perkembangan	…	Memiliki
Sistem	Di	Yang	…	Tingkat
Pemerintah	Era	Terjadi	…	Eselon
Diera	Otonomi	Kebutuhan	…	Yang
Orde	Daerah	Di	…	Lebih
baru	Ini	Daerah	..	Tinggi
	pembentukan	masingmasing	…	bisa

7. *Synonim Checking,*

In the process of checking synonyms process is done is replace the word that has the same meaning as a word that has been created in the master database. At this time the test document words that fall into a process that is suppose synonymous to the word "structural, universal, schematic" which will be replaced by the word "systematic". [9]

Table 4. Synonim Checking

Kal1	Kal2	Kal3	…	Kal22
Era	Diera	Karena	…	Kegiatan
Ekonomi	Orde	Itu	…	Humas
Daerah	Baru	Sistem	…	yang
Saat	Organisasi	Informasi	…	Banyak

Ini	Pemerintah	Pada	…	Berhubungan
Sistem	Dan	Setiap	…	Dengan
Pemerintahan	Sistem	Daerah	…	Uniot
Daerah	Informasinya	Bisa	…	Kerja
Sudah	Ditentukan	Berbeda	…	Di
Berbeda	Oleh	Sesuai	…	Pemda
Dibandingkan	Pemerintah	Dengan	…	Yang
Dengan	Pusat	Perkembangan	…	Memiliki
Sistem	Di	Yang	…	Tingkat
Pemerintah	Era	Terjadi	…	Eselon
Diera	Otonomi	Kebutuhan	…	Yang
Orde	Daerah	Di	…	Lebih
baru	Ini	Daerah	..	Tinggi
	pembentukan	masingmasing	…	bisa

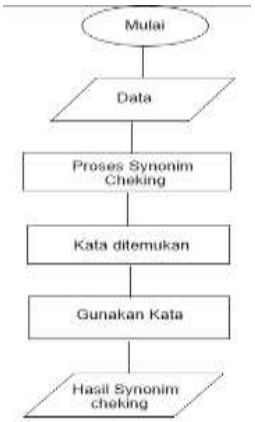


Figure 2. The process of checking Synonym

8. *Stopwords*

In stopwords removal process is a process that is done is remove the less relevant word or words that do not have a meaning that is so important and related to that of the sentence in the document with a list of matching words stopword on the database to be tested. The database is used for both Indonesian and English taken from a site that provides berbagain types stopword list for a variety of languages.

Table 5. Stopword

Kal1	Kal2	Kal3	…	Kal22
Era	Diera	Karena	…	Kegiatan
Ekonomi	Orde	Itu	…	Humas
Daerah	Baru	Sistem	…	yang

Saat	Organisasi	Informasi	…	Banyak
Ini	Pemerintah	Pada	…	Berhubungan
Sistem	Dan	Setiap	…	Dengan
Pemerintahan	Sistem	Daerah	…	Uniot
Daerah	Informasinya	Bisa	…	Kerja
Sudah	Ditentukan	Berbeda	…	Di
Berbeda	Oleh	Sesuai	…	Pemda
Dibandingkan	Pemerintah	Dengan	…	Yang
Dengan	Pusat	Perkembangan	…	Memiliki
Sistem	Di	Yang	…	Tingkat
Pemerintah	Era	Terjadi	…	Eselon
Diera	Otonomi	Kebutuhan	…	Yang
Orde	Daerah	Di	…	Lebih
baru	Ini	Daerah	..	Tinggi
	pembentukan	masingmasing	…	bisa

4.3 Application Analysis TF-IDF

Methods Term Frequency-Inverse Document Frequency (TF-IDF) is a way of assigning weights to the relationship of a word (term) of the document. For a single document each sentence is considered as a document. This method combines two concepts for weight calculation, namely Term Frequency (TF) is the frequency of words (t) of the sentence (s). Document frequency (DF) is the number of sentences in which a word (t) appears. [3] The frequency of occurrence of the word in a given document shows how important word in the document.

The steps are:

1. 1. Count the number of occurrences of each word on every sentence in the document to be tested. At this time of testing, a document that has 22 words (N).
2. 2. After calculating the number of occurrences of each word in each sentence contained in the test documents, the next step is to calculate the total appearance of each word on the entire sentence there. The result is stored in the DF column or Document Frequention.
3. 3. The next step is to divide the total value of the existing line in the test documents with the total value of the number of occurrences of each word in the whole sentence contained in the test documents.

The results of the division is stored in the "N / DF".
N = Many Sentence = 22
DF = Total occurrence of the word
Then
N/DF [problems] = 22/1 = 22
N/DF [happens] = 22/1 = 22
Up N/DF[n]..

4. The next step is to do mathematical calculations using the formula that the IDF are in equation 1.

$$IDF = log(\frac{N}{Df}) \quad (1)$$

5. Calculation :
IDF(1) = log (15/4) = 0.57403
IDF(2) = log (15/1) = 1.17609
IDF(3) = log (15/1) = 1.17609
…… Up IDF(idx of n)

6. The final step of the TF-IDF method is to calculate the weight of each sentence, using equation 2.

$$W_{d,t} = TF_{d,t} * IDF_t \quad (2)$$

dengan :

d = kalimat ke-d

t = kata (*term*) ke-t

TF = *Term Frequency*

W = bobot kalimat ke-d terhadap kata (*term*) ke-t

IDF= *Inverse Document Frequency*

Then the calculation is :
Sentence to 1, according to 1 => W(1.1) = 2 * 0.176091259 = 0.3521825
Sentence to 2, according to 1 => W(2.1) = 0 * 0.176091259 = 0
Sentence to 3, according to 1 => W(2.1) = 0 * 0.176091259 = 0
Dst.. to
Sentence to (1,2,3), according to n => W(n.n) = x

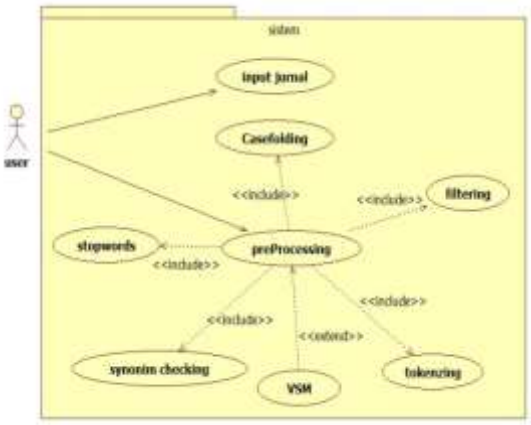
After completing the weighting using the Term Frequency - Inverse Document Frequention it will form a matrix of term-by-sentence where the values that have been obtained in this process will be used for calculating the value of cosine.

4.4 Modeling System

The modeling system is a section describing the functions - functions that can be performed by the application based on the analysis method that has been done before and mendeksripsikannya them into the form of a diagram usecase diagrams, scenario, activity diagram, sequence diagram, and class diagram.

4.4.1 *Use Case Diagram*.

Use case diagram shows the interaction between the group process with a group of actors, describe the functionality of a system built and how the system interacts with the outside world. Use case diagrams can be used during the analysis process to capture system requirements and to understand how the system is supposed to work. In Figure 3.4. Use Case Diagram showing the application of information retrieval are words or phrases in a document using TF-IDF



Gambar 3. ***Use Case Diagram***

4.5 Scenario Testing

Test scenario describes an action to be performed when testing of the system lasts. In the study abstract formation of this automated system, will be testing the accuracy of the results and the results of the abstract on abstract manual system. The results obtained from the manual abstract 3 people who are putting together the final project requiring an abstract sentence, given a case study taken in this research is the formation of abstract document journal. 3 will examine the comparison of the results with the results of the abstract abstract manually by the system.

The data used in this thesis research amounted to 15 general-format document journal. Abstract formation is done by taking the sentence on each piece that background, analysis, and conclusions in the document journal, either for results is done manually and the results of the system. If> 2 people choose the same sentence, the conclusion of the sentence was elected, and if <1 then the sentence is not selected.

4.5.1 Conclusion Testing Results

Conclusion of the test results stating that the results of the abstract formation manually by percentage system generates the same sentence is not too far away. Where the distance between the

user results with average system differs only 2 to 4 sentences. 100% accuracy with the percentage obtained by the journal to 7 and 9 while pesentase accuracy with the lowest 60% earned by the journal to 1.

Table 6. Conclusions Testing Accuracy

Journals	Accuracy of the same sentence	Persentase
1	1,14,16,19,20,22	60%
2	2,4,7,12,16,17,19	80%
3	1,3,5,7,12,14,16,17,19	90%
4	2,5,11,16,21,23,24	70%
5	4,6,14,18,20,22,25,26	70%
6	1,2,5,11,14,20,21	70%
7	3,5,8,9,11,12,14,16,18	100%
8	1,3,8,12,15,19,20,22	90%
9	2,4,6,10,11,14,15,18,20,22,25,27	100%
10	1,7,11,12,16,21,23,24	70%
11	1,7,11,12,16,18,19,21,22	90%
12	3,4,7,9,11,13,16,17,18	80%
13	1,4,7,11,14,16,17	70%
14	1,3,7,8,14,16,17,20,22	90%
15	3,8,9,12,13,15,16,18,19	90%

3. CLOSING

Based on the results obtained in the research and preparation of this paper it can be concluded that the formation of abstract sentences automatically using the Term Frequency - Inverse Document Frequency (TF-IDF) capable of forming an abstract paragraph in a document automatically journal. Conclusion of the test results stating that the results of the abstract formation manually by percentage system generates the same sentence is not too far away. Where the distance between the user results with average system differs only 2 to 4 sentences. 100% accuracy with the percentage obtained by the journal to 7 and 9 while pesentase accuracy with the lowest 60% earned by the journal to 1. As for suggestions for the development of the system of formation of this abstract is perfecting the application is still to be done, the current document is used only in DOC format.

BIBLIOGRAPHY

[1] Gary Blake and Robert W. Bly, 1993 . *The Elements of Technical Writing*, pg. 117. New York:Macmillan Publishers, ISBN 0020130856

[2] Taufiq M. Isa 1) dan Taufik Fuadi Abidin 2)., "Mengukur tingkat kesamaan paragraf menggunakan vector space model untuk mendeteksi plagiarisme" Seminar Nasiaonal dan ExpoTeknik Elektro, 1) jurusan matematika. FMIPA, Universitas Syiah Kuala, 2) Jurusan informatika, FMIPA, Universitas Syiah Kuala, 2013.

[3] R. S. Pressman, Software Engineering: A Practitioner's Approach.

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

[5] P. P. Widodo dan Heriawati. "Menggunakan *Unified Modeling Language* (UML)". Bandung Informatika, 2011.

[6] Wahana Komputer, Shortcourse Series: Microsoft Visual C# 2010. Andi, 2011

[7] Deny Ocr.(2014, Aug.10) *Pengenalan DevExpress* [online]. Available : http://www.jagocoding.com/tutorial/600/Pengenalan_DevExpress_Windows_Form

[8] Sholechul Aziz, Jurus Andalan Menguasai EYD (Ejaan yang Disempurnakan). IV. Jakarta : Kunci Komunikasi, 2015.

[9] S Mohd. Ehmer Khan. "Different Approaches to White Box Testing Technique for Finding error", 2011

[10] Many, I. and Maybury. 1999. *Advance in Automatic Text Summarization*. The MIT Press: Cambrige.

[11] Zaman B. dan E Winarko. 2011. Analisis Fitur Kalimat untuk Peringkas Teks otomatis pada Bahasa Indonesia. *Indonesian Journal of Computing and*.