Measurement of the Similarity of Indonesian Papers on One Journal Topic with the Naive Bayes Algorithm and Vector Space Model

Ni Luh Wiwik Sri Rahayu Ginantra[1]
Information Technology Department
STMIK STIKOM Indonesia
Denpasar, Indonesia
wiwik@stiki-indonesia.ac.id

Ni Wayan Wardani[2]
Information Technology Department
STMIK STIKOM Indonesia
Denpasar, Indonesia
niwayan.wardani@stiki-indonesia.ac.id

***Abstract*—One way to maintain the quality of scientific work in Indonesia is by checking articles before they are published. Checking before the publication was done so that the similarity level is not high because the published papers can be quoted to cause a high level of similarity. The next problem is the importance of grouping topic papers, where papers to be checked should have the same category as comparative papers. In this study, to classify the topic of the journal using the Naïve Bayes algorithm and to measure the similarity of papers using the Vector Space Model method. Naïve Bayes algorithm can better classify the test data with the .docx file format than to the test data in the .pdf file format. The results of the calculation of text similarity detection by the Vector Space Model can reach 90% and above for test data with the .docx file format, while for test data with the .pdf file format the calculation results by the Vector Space Model are on average less than 90%. The results of the calculation of text similarity detection by the *V*ector *S*pace *M*odel method are also strongly influenced by training data. The more complete and complex of the training data, then more valid the results of the Vector Space Model performance testing**

***Keywords*—similarity; classification; naïve bayes; vector space model**

I. Introduction

At present, one of the essential points in carrying out the functions of the Tridharma of Higher Education by lecturers is conducting research and publishing the results of their thoughts and analyzes. The performance of lecturers which subsequently became the performance of departments, faculties and universities was greatly influenced by the extent and quality of the publications of the permanent lecturers.

Publication demands made by the academic community of universities have a considerable impact on the awareness of the lecturers of the importance of conducting studies, research, and writing scientific works. The development of scientific work in Indonesia has been relatively good, especially since the enactment of government regulations, which required S1, S2 and S3 students to write articles in scientific journals as one of the prerequisites for graduation. For lecturers, of course, there will be higher demands for active writing in scientific journals at the accredited national level and reputable international journals

In line with these government regulations, there will be an increase in the number of scientific publications by academics. With the increasing number of publications, the quality of scientific work is also very important. One way to maintain the quality of scientific work in Indonesia is by checking articles before they are published. Checking before the publication was done so that the similarity level is not high because the published papers can be quoted to cause a high level of similarity.

In addition to the need to check articles before they are published, the next problem is the importance of grouping topic papers, where papers to be checked should have the same category as comparative papers.

Based on these problems, then in this study, The Naïve Bayes algorithm will be used to classify articles into one topic. The workings of the Naïve Bayes algorithm are using probability calculations. The basic concept is to calculate the opportunities of a class from each group of attributes that exist and determine which class is the most optimal. The grouping or classification process is divided into two phases, namely learning / training and testing / classify. In the learning phase, part of the data that has known the data class is likened to forming an approximate model, then in the testing phase, the model that has been formed is tested with some data.

Based on these problems, then in this study, The Naïve Bayes algorithm will be used to classify articles into one topic. The workings of the Naïve Bayes algorithm are using probability calculations. The basic concept is to calculate the opportunities of a class from each group of attributes that exist and determine which class is the most optimal. The grouping or classification process is divided into two phases, namely learning / training and testing / classify. In the learning phase, part of the data that has known the data class is likened to forming an approximate model, then in the testing phase, the model that has been formed is tested with some data.

II. Method

A. *Dataset*

The dataset used in this study is PDF documents in Indonesian, which are from the repository of the journal neliti.com. Journal topics used in this study also took the topic of journals contained in neliti.com totaling 67 topics or fields of study.

Neliti is a research search engine that helps research institutions and universities in Indonesia to rediscover research results, primary data, and facts. Neliti indexes scientific journals, books, research reports, policy papers, conference papers, and primary data from universities, research bodies, government institutions, and publishers [1].

Neliti as a single repository that contains many research results that were previously spread on various websites so that it is difficult to find. Through the process of gathering this content into one database, Neliti strives to support researchers in producing research that improves the quality of life for the Indonesian people [1].

B. *Research Design*

The following are the stages of the research design :

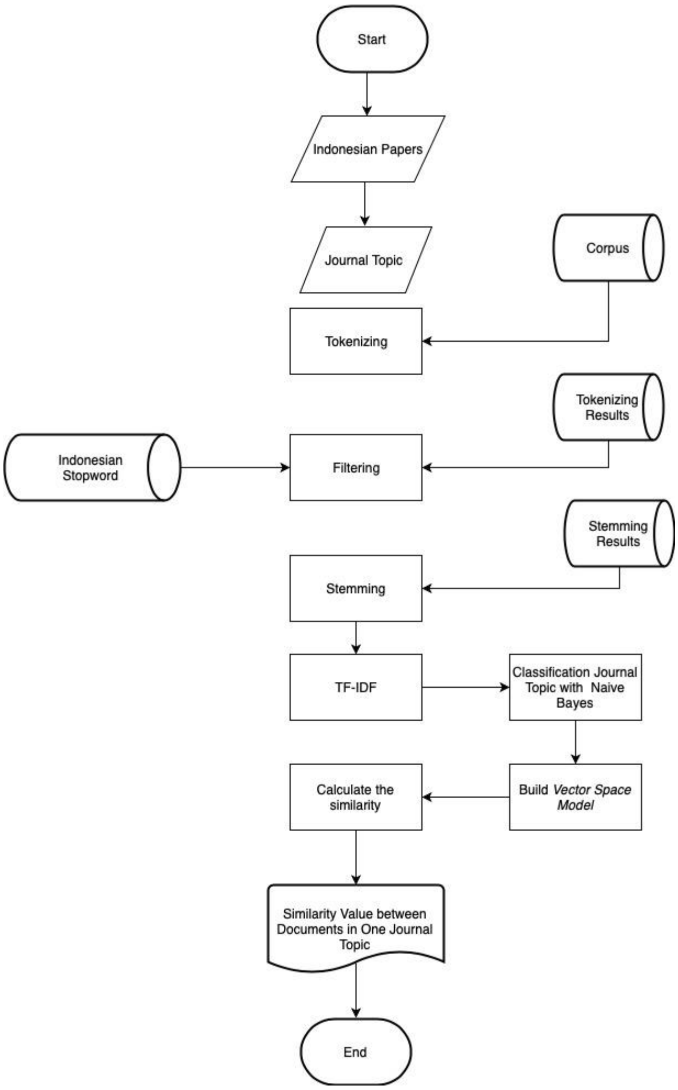

Fig. 1. The stages of the research design

C. *Text Processing*

Text mining is the process of retrieving data in the form of text from a source; in this case, the source is a document. With text mining, can search for keywords that can represent the contents of a document and then analyze and do the matching between documents and database keywords that have been made. Text processing is part of text mining. The stages of text processing, in general, are tokenizing, stopping, and stemming[2].

1) Tokenizing

Tokenizing is a process carried out on documents to get terms. The process that is done is to cut the words that build a document, and the results of the pieces are called tokens, and maybe in the same process throw various characters such as punctuation [3].

2) Stopping / Filtering

Stopping is a process that is carried out after tokenizing on text processing. The process of stopping is to eliminate words that often appear in general, called stop words. Stop word tends to have a low weight, so it almost does not affect the calculation if the stop word is deleted. One technique commonly used to reduce word index is by stemming or removing stop words [4].

3) Stemming

Stemming is the process of getting root words from a term. The purpose of this process is done so that the meaning of a term from one document is the same as other documents because the term is already in the basic form. For reasons of word transformation, a document usually uses a different form of the word, even though the word has a meaning that is not much different. In many situations, it will be helpful if the different forms of the word are considered the same.

D. *Term Frequency-Inversed Document Frequency Algorithm (TF-IDF)*

One way to give word weight (term) t of a document (document) d is to calculate the number of t words in document d; this weighting is called the term frequency TF. The weakness of the term frequency is that all words have an equally important weight. One solution to this weakness is to give high weight to words that appear slightly in many documents. This is because words that appear a little on many documents are considered necessary. To weigh the weight of words that appear slightly on many documents generally use inversed document frequency (IDF). Combining the weight of TF and IDF is done by multiplication, merging is done so that we get the mixed weight of a term from each document [3].

TF: Term Frequency, which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (the total number of terms in the document) as a way of normalization[5]:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

IDF : Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important.

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

E. *Naïve Bayes Classifier*

Naive Bayes is a simple probabilistic classification that calculates a set of probabilities by summing frequencies and combinations of values from a given dataset. The algorithm uses the Bayes theorem and assumes all independent or non-interdependent attributes given by values on class variables. Another definition says Naive Bayes is a classification with probability and statistical methods presented by British scientist Thomas Bayes, namely predicting future opportunities based on previous experience

Naive Bayes is based on the simplifying assumption that attribute values are conditionally independent when given output value. In other words, given an output value, the probability of observing together is a product of individual probabilities. The advantage of using Naive Bayes is that this method only requires a small amount of training data to determine the estimation of parameters needed in the classification process. Naive Bayes often works much better in most complex real-world situations than expected, so the NB method is the method used for the text classification process in this study. There are two stages in the text classification process. The first step is training the sample article set (training example). While the second stage is the process of classification of documents whose topic is unknown

Theorema Bayes :

$$P(Ci \mid X) = \frac{P(X|Ci)\ X\ P(Ci)}{P(X)}$$

Information:

P (Ci | X) : The probability of occurring Ci class with X

P (X) "constant" for all classes so only formed P (X | Ci) x P (Ci) which is necessary maximized.

X : event X

Ci : available class (C1, C2, ... Ci)

P (Ci) : probability of occurring Ci class.

P (X) : The probability of occurrence of event X.

P (X | Ci) : The probability of occurrence of event X with condition

$$P(X|Ci) = P(Xt\ |Ci)$$

Information:

Xt : attribute values in sample X

P (Xt | Ci) : the probability of occurrence Xt with the condition of Ci, can calculated from database training

F. *Vector Space Model*

The Vector Space Model or Term Vector Model method is an algebraic model for describing text documents (several objects) as vectors of identifiers. It is usually used in information filtering (information filtering), information discovery (information retrieval), indexing, and ranking that are mutually relevant. The process of calculating this method is document indexing, term weighting, and similarity calculations. The document indexing process is the process through stages in text mining. The next process is weighing the term using the TF / DF algorithm. The last process is the calculation of similarity with the Cosine approach, which is stated in the formula[6]:

$$Similarity(dj, qk) = \frac{\sum_{i=1}^{n}(tdjj\ \mathrm{X}\ \mathrm{tqik})}{\sqrt{\sum_{i=1}^{n} tdjj\ X \sum_{i=1}^{n} tqik}}$$

Keterangan:

Similarity(dj,qk) : level of approval of a document with specific requests

tdij : *i-term in vector for j-document*

tqik : *i-term in vector for k-query*

n : the number of terms that are unique in the data set

G. *Classification Journal Topic with Naïve Bayes*

The classification steps use the *Naive Bayes Classifier* method as follows:

TABLE I. DOCUMENTS SAMPLE

1	Penelitian ini berupa pengembangan Sistem Pendukung Keputusan (SPK) untuk perencanaan kebutuhan bahan baku yang mempunyai batas masa kadaluarsa dan adanya ketentuan diskon bagi pembelian dalam jumlah tertentu.
2	Pemeriksaan pajak merupakan serangkaian kegiatan untuk mencari, mengumpulkan, mengolah data dan atau keterangan lainnya untuk menguji kepatuhan pemenuhan kewajiban perpajakan dan untuk tujuan lain dalam rangka melaksanakan ketentuan peraturan perundang-undangan perpajakan.

1. Stop word Removal

Removal of conjunctions that exist in the document and calculate the frequency of occurrence of the conjunctions that will be deleted in the example document:

TABLE II. STOPLIST

No	*Stoplist*	Frekuensi
1	ini	1
2	berupa	1
3	untuk	3
4	yang	1
5	dan	3
6	adanya	1
7	bagi	1
8	dalam	2
9	tertentu	1
10	mencari	1
11	atau	1
12	lainnya	1
13	lain	1

2. Tokenizing

After removing the conjunctions. Example document that has deleted the conjunctions and changed all uppercase letters after a collection of characters in a document into units of words.

TABEL III. DOCUMENT AFTER STOPWORD REMOVAL

1	Penelitian berupa pengembangan Sistem Pendukung Keputusan SPK perencanaan kebutuhan bahan baku mempunyai batas masa kadaluarsa ketentuan diskon pembelian jumlah tertentu
2	Pemeriksaan pajak merupakan serangkaian kegiatan mencari mengumpulkan mengolah data keterangan menguji kepatuhan pemenuhan kewajiban perpajakan tujuan rangka melaksanakan ketentuan peraturan perundang undangan perpajakan.

TABLE IV. CHANGE ALL UPPERCASE LETTERS TO LOWERCASE

1	penelitian berupa pengembangan sistem pendukung keputusan spk perencanaan kebutuhan bahan baku mempunyai batas masa kadaluarsa ketentuan diskon pembelian jumlah tertentu
2	pemeriksaan pajak merupakan serangkaian kegiatan mencari mengumpulkan mengolah data keterangan menguji kepatuhan pemenuhan kewajiban perpajakan tujuan rangka melaksanakan ketentuan peraturan perundang undangan perpajakan

TABLE V. TOKENIZING PROCESS FROM THE DOCUMENT SAMPLE

No	Term	No	Term
1	penelitian	23	merupakan
2	berupa	24	serangkaian
3	pengembangan	25	kegiatan
4	sistem	26	mencari
5	pendukung	27	mengumpulkan
6	keputusan	28	mengolah
7	spk	29	data
8	perencanaan	30	keterangan
9	kebutuhan	31	menguji
10	bahan	32	kepatuhan
11	baku	33	pemenuhan
12	mempunyai	34	kewajiban
13	batas	35	perpajakan
14	masa	36	tujuan
15	kadaluarsa	37	rangka
16	ketentuan	38	melaksanakan
17	diskon	39	ketentuan
18	pembelian	40	peraturan
19	jumlah	41	perundang
20	tertentu	42	undangan
21	pemeriksaan	43	perpajakan
22	pajak		

3. Determining the IDF Value

After tokenizing, the results of tokenizing, data is checked on each topic to see the appearance of the word. Then the appearance of the word (df) is used as a reference in finding the value of idf with the log formula (number of topics/df in each word).

4. Determining TD-IDF Value

The result of Idf then looks for the value of tf-idf with the formula (occurrence of the word for each topic * idf value).

TABEL VI. IDF VALUE

appearance of the term	d1	d2	d3	d4	d5	df	idf
penelitian	1	0	0	0	0	1	0,69897
berupa	1	0	0	0	0	1	0,69897
pengembangan	1	0	0	0	0	1	0,69897
Sistem	1	0	0	0	0	1	0,69897
pendukung	1	0	0	0	0	1	0,69897
keputusan	1	0	0	0	0	1	0,69897
spk	1	0	0	0	0	1	0,69897
perencanaan	1	0	0	0	0	1	0,69897
kebutuhan	1	0	0	0	0	1	0,69897
Bahan	1	0	0	0	0	1	0,69897
Baku	1	0	0	0	0	1	0,69897
mempunyai	1	0	0	0	0	1	0,69897
Batas	1	0	0	0	0	1	0,69897
Masa	1	0	0	0	0	1	0,69897
kadaluarsa	1	0	0	0	0	1	0,69897
ketentuan	1	0	0	0	0	1	0,69897
Diskon	1	0	0	0	0	1	0,69897
pembelian	1	0	0	0	0	1	0,69897
Jumlah	1	0	0	0	0	1	0,69897
Tertentu	0	0	1	0	0	1	0,69897
pemeriksaan	0	0	1	0	0	1	0,69897
Pajak	0	0	1	0	0	1	0,69897
merupakan	0	0	1	0	0	1	0,69897
serangkaian	0	0	1	0	0	1	0,69897
kegiatan	0	0	1	0	0	1	0,69897
mencari	0	0	1	0	0	1	0,69897
mengumpulkan	0	0	1	0	0	1	0,69897
mengolah	0	0	1	0	0	1	0,69897
data	0	0	1	0	0	1	0,69897
keterangan	0	0	1	0	0	1	0,69897
menguji	0	0	1	0	0	1	0,69897
kepatuhan	0	0	1	0	0	1	0,69897
pemenuhan	0	0	1	0	0	1	0,69897
kewajiban	0	0	1	0	0	1	0,69897
perpajakan	0	0	1	0	0	1	0,69897
tujuan	0	0	1	0	0	1	0,69897
rangka	0	0	1	0	0	1	0,69897
melaksanakan	0	0	1	0	0	1	0,69897
ketentuan	0	0	1	0	0	1	0,69897
peraturan	0	0	1	0	0	1	0,69897
perundang	0	0	1	0	0	1	0,69897
undangan	0	0	1	0	0	1	0,69897
perpajakan	0	0	1	0	0	1	0,69897

The calculation example of tdf-idf_d1 is the multiplication of idf values with the value of category d1 where
idf * d1 = (1 * 0.39794) which produces a value of 0.39794.

5. Determining the Identification Word (Feature)

Identifying the appearance of words from the results of IDF then counting the number of words on each topic, words that have the highest tf-idf value from tfidf_d1, and tfidf_d2, are the words that identify the topic.

TABLE VII. TD-IDF VALUE

tfidf_d1	tfidf_d2	tfidf_d3	tfidf_d4	tfidf_d5
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0

TABLE VIII. IDENTIFICATION WORD

d1	d3	d1	d3
penelitian	pemeriksaan	baku	menguji
berupa	pajak	mempunyai	kepatuhan
pengembangan	merupakan	batas	pemenuhan
sistem	serangkaian	masa	kewajiban
pendukung	kegiatan	kadaluarsa	perpajakan
keputusan	mencari	ketentuan	tujuan
spk	mengumpulkan	diskon	rangka
perencanaan	mengolah	pembelian	melaksanakan
kebutuhan	data	jumlah	ketentuan
bahan	keterangan	tertentu	peraturan
	undangan		perundang
	perpajakan		

6. The result of classification calculation

TABEL 9. THE RESULT OF CLASSIFICATION CALCULATION

Topic	Calculation
d1	0,333333
d2	0,291667
d3	0,3125
d4	0,291667
d5	0,291667

H. *Calculation of Document Similarity*

1. Calculate the root of the total term keywords in all documents and the root of the terms of each document from the results of tokenizing.

   Formula: sqrt (number of terms or number of term documents)

   example:
   q (number of keywords) = 5
   d1 (number of documents 1) = 19
   d2 (number of documents 2) = 24

   q: sqrt (5) = 2,23606
   d1: sqrt (19) = 4.35890
   d2: sqrt (24) = 4.898979

2. After getting the root of the document and keyword, then calculate the similarity.

   Formula: (number of keyword terms in document * number of term documents) / (the root of a keyword * the root of a keyword)

   Example:
   d1 = keyword appears in document 1 by 4
   d2 = keyword appears in document 2 as many as 1

   d1: (4 * 19) / (2,23606 * 4,35890) = 7.79743
   d2: (1 * 24) / (2,23606 * 4,89897) = 2,19089

Calculation of the similarity search above between document and document, so that it can sort which documents are most similar to the test document. From these results, it can be seen that d1 has a greater result. The result d1 is above the search ranking and d2 below it.

III. RESULT AND DISCUSSION

The application of the Naïve Bayes classification model and the similarity detection of paper with the Vector Space Model are tested with several papers that have .pdf and .docx file formats. The following is a trial conducted:

Table 10. DATA TRAINING

Data Training	
Papers	Journal Topic
Paper1.pdf	Computer Science & Information Technology
Paper2.pdf	Computer Science & Information Technology
Paper3.pdf	Computer Science & Information Technology
Paper4.pdf	Computer Science & Information Technology
Paper5.pdf	Computer Science & Information Technology
Paper6.pdf	Computer Science & Information Technology
Paper7.pdf	Computer Science & Information Technology
Paper8.pdf	Computer Science & Information Technology
Paper9.pdf	Computer Science & Information Technology
Paper10.pdf	Computer Science & Information Technology

The data in table 10 are papers that become training data. These papers have journal topics in the fields of computer science and information technology. The file format used as training data is .pdf.

The data in table 11 are papers that become test data. These papers have journal topics in the fields of computer science and information technology. The file format used as test data is .pdf.

In table 12, below shows the results of the trial. Test 1 uses 10 test data. All test data documents are the same as documents on training data with the same file format, namely .pdf. Test 2 uses 10 test data. All test data documents are the same as documents in training data but with different file formats, namely, file formats .docx.

TABLE 11. TEST DATA

Data Training	
Papers	Journal Topic
Paper1.pdf	Computer Science & Information Technology
Paper2.pdf	Computer Science & Information Technology
Paper3.pdf	Computer Science & Information Technology
Paper4.pdf	Computer Science & Information Technology
Paper5.pdf	Computer Science & Information Technology
Paper6.pdf	Computer Science & Information Technology
Paper7.pdf	Computer Science & Information Technology
Paper8.pdf	Computer Science & Information Technology
Paper9.pdf	Computer Science & Information Technology
Paper10.pdf	Computer Science & Information Technology

TABLE 12. RESULT OF TRIALS

Trials		1	2	3	4	5	6	7	8	9	10
Trials 1	Classification	5.53E-288	-	-	1.15E-128	-	-	2.42E-206	7.28E-254	-	8.54E-218
	Similarity	91.52%	86.73%	76.54%	94.79%	-	62.61%	90.38%	90.84%	71.74%	96.52%
Trials 2	Classification	2.07E-288	3.86E-141	2.27E-255	1.15E-128	6.21E-153	1.89E-219	2.42E-206	7.28E-254	7.35E-288	1.52E-211
	Similarity	93.38%	95.88%	94.24%	94.66%	-	81.68%	90.38%	93.13%	84.12%	96.55%

IV. CONCLUSION

A. *Conclusion*

1. Naïve Bayes algorithm can better classify the test data with the .docx file format than to the test data in the .pdf file format.
2. In some test data documents with the .pdf file format, the naïve Bayes algorithm cannot classify into one journal topic precisely but classifies several journal topics so that it affects the performance of the Vector Space Model method.
3. The results of the calculation of text similarity detection by the Vector Space Model can reach 90% and above for test data with the .docx file format, while for test data with the .pdf file format the calculation results by the Vector Space Model are on average less than 90%.
4. The results of the calculation of text similarity detection by the Vector Space Model method are also strongly influenced by training data. The more complete and complex of the training data, then more valid the results of the Vector Space Model performance testing.

B. *Future Work*

1. The next study of the performance testing of the Naïve Bayes algorithm and Vector Space Model method can be tested on paper documents in various file formats.
2. Training data can be connected directly to the database from Neliti.com

ACKNOWLEDGMENT

We gratefully thank DRPM Ditjen Penguatan Risbang and BOPTN for financial support. We also thank you for all those who have helped a lot in this research.

REFERENCES

[1] M. W. Berry and J. Kogan, "Text Mining Applications and Theory". Wiley, 2010.

[2] P. Manning, Christopher D and Raghavan, "Introduction to Information Retrieval". California: Stanford University, 2008.

[3] M. and Chenoweth and M. Song, "Text Categorization dalam Encyclopedia of Data Warehouse & Data Mining". IGI Global, 2009.

[4] A. Ryansyah and A. P. K. Dokumen, “Implementasi Algoritma TF-IDF Pada Pengukuran Kesamaan Dokumen,” Jurnal Sistem & Teknologi Informasi Komunikasi, vol.1, pp. 1–10, 2016.

[5] T. M. Isa and T.F Abidin., “Mengukur Tingkat Kesamaan Paragraf Menggunakan Vector Space Model untuk Mendeteksi Plagiarisme,” Seminar Nasional dan Expo Teknik Elektro, pp-229-234, 2013.

[6] https://www.neliti.com, Reporitory Ilmiah Indonesia.