TF-IDF Method and Vector Space Model Regarding the Covid-19 Vaccine on Online News

By Bita Parga Zen

WORD COUNT 5275 TIME SUBMITTED 08-AUG-2023 12:23PM

PAPER ID 101834967

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

TF-IDF Method and Vector Space Model Regarding the Covid-19 Vaccine on Online News

Bita Parga Zen[1)*], Irwan Susanto[2)], Dian Finaliamartha[3)]
[1)2)3)]Informatics Engineering Department, Faculty of Informatics, Telkom Institute of Technology Purwokerto
[1)]bita@ittelkom-pwt.ac.id, [2)] irwansusanto_yk@ittelkom-pwt.ac.id, [3)] 217102131@ittelkom-pwt.ac.id

Submitted : Oct 2, 2021 | **Accepted** : Oct 10, 2021| **Published** : Oct 10, 2021

Abstract: Advances in information and technology have caused the use of the internet to be a concern of the general public. Online news sites are one of the technologies that have developed as a means of disseminating the latest information in the world. When viewed in terms of numbers, newsreaders are very sufficient to get the desired information. However, with this, the amount of information collected will result in an explosion of information and the possibility of information redundancy. The search system is one of the solutions which expected to help in finding the desired or relevant information by the input query. The methods commonly used in this case are TF-IDF and VSM (Vector Space Model) which are used in weighting to measure statistics from a collection of documents on the search for some information about the Covid 19 vaccine on kompas.com news then tokenizing it to separate the text, stopword removal or filtering to remove unnecessary words which usually consist of conjunctions and others. The next step is sentence stemming which aims to eliminate word inflection to its basic form. Then the TF-IDF and VSM calculations were carried out and the final result are news documents 3 (DOC 3) with a weight of 5.914226424; news documents 2 (DOC 2) with a weight of 1.767692186; news documents 5 (DOC 5) with weights 1.550165096; news document 4 (DOC 4) with a weight of 1.17141223;, and the last is news document 1 (DOC 1) with a weight of 0.5244103739.

Keywords: TF-IDF, Vector Space Model, Covid 19, Vaccine, Online News

INTRODUCTION

Advances in information and technology currently cause the use of internet to have become the attention of the general public because its ease of accessing a site on the internet, nowadays people can easily find information. Especially on online news as one of the technologies that has now developed as a means to provide information ranging from the latest news, politics, health, inspiration, education, technology, automotive etc. The development of the website as one form of mass media which resulted in a sharp increase the amount of information escpecially a news article. As for the results of observations from news sites (Tribunnews.com, Detik.com, Liputan6.com, and others) using the scraping technique, in 2018 it was found that there were approximately 109,061 published news (F.Wiranto 2019). When viewed from the side of the number, news readers are very sufficient to get the desired information. However, with this, the amount of information that is accommodated will result in an explosion of information and the possibility of information redundancy. The reader must search for all documents on existing news sites and determine whether the news document is in accordance with the topic intended by the reader or not. Readers will spend a lot of time reading various news from the same or different online news sites but having the same information core. From these problems, a program is needed to make it easier for readers to avoid getting information with the same core (F.Wiranto 2019).

Seeing the rapid spread of COVID-19 and the dangers that will arise is not treated immediately, one very possible way to prevent the spread of this virus is to develop a vaccine because the virus spreads quickly, so a vaccine is needed that can be applied in a short time so as to minimize its impact. The community gives their responses and opinions in various media. One of the media that is widely used by the public to give their opinion on something is social media. Social media now seems to be something that must be owned by all people (Rachman, Ff, 2020).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In information retrieval, TF -IDF is a term of inverse-frequency document frequency, is a numerical statistic intended to reflect how important a word to a document in a collection or corpus. TF-IDF is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases in proportion to the number of times a word appears in the document, but is offset by the frequency of words in the corpus, which helps adjust for the fact that some words appear more frequently in general. A variation of the TF-IDF weighting scheme is often used by search engines as the primary tool in assessing and ranking document relevance based on user queries. TF–IDF can be used for stop-words filtering in various subject areas including text summarization and classification (Abo El-Khair, I 2017).

Furthermore, the Vector Space Model (VSM) is one of the models used to determine the similarity of documents, which is used in generating automatic FAQs. VSM modeling has an efficient way of working, easy to represent and can be implemented in document-matching (Aziz, Abdul, 2015).

LITERATURE REVIEW

Data Types and Sources

The type of primary data used in this study is information data on the kompas.com site because the kompas.com site provides updated information about the covid 19 vaccine. Through online media, in the beginning of the Covid-19 vaccine, when Kompas media reported that President Jokowi was reviewing the implementation of candidate injections. The inaugural Covid-19 vaccine for 20 volunteers from a target of 1,620 volunteers located at the Faculty of Medicine, Padjadjaran University, Bandung, West Java, on August 11, 2020. The injection is a series of phase III clinical trials of vaccine candidates developed by Sinovac Biotech, China (Kompas Gramedia, 2020). The vaccine candidate is named CoronaVac. PT Bio Farma (Persero) as a Pharmaceutical BUMN is collaborating with Sinovac Biotech in the CoronaVac phase III clinical trial in Indonesia through technology transfer and knowledge transfer (Kompas Gramedia, 2020).

Information Retrieval

Information Retrieval is the representation, storage, organization and access to information items. The representation and organization of items is arranged in such a way that it is easily accessible if you want to get the information needed by a user (Baeza-Yates, Ricardo 1999). Information Retrieval has three main components , namely input , processor , and output. According to (Manning et al 2009), Information Retrieval is a way to find unstructured documents (usually in the form of text) that can meet the information needs that are usually stored in computers (Manning et al 2009).

Text Mining

Text mining is a process of extracting information where a user interacts with a set of documents by using several tools used in data mining to perform several components, namely categorization (R. Melita Et Al 2018).

The development in the fields of web, digital libraries, technical documentation, medical data has made it easier to access a larger amount of a textual documents, which come together to develop useful data resources. Therefore, it makes text mining (TM) or the knowledge discovery from textual databases a challenging task owing to meet the standards of the depth of natural language which is employed by most of the available documents. The available textual information in the form of databases and online sources (Salloum, Al-Emran, Monem, & Shaalan, 2018).

METHOD

TF-IDF (Term Frequency-Inverse Document Frequency) weighting

TF-IDF is a type of weighting for a document in a corpus used in Information Retrieval. TF-IDF weighting is often used for statistical measurements in measuring how important a word is in a document. V Variations of TF-IDF weighting scheme is often used by search engines (engines) as a primary tool to assess and rank the relevance and offset by the frequency of words in a document. TF-IDF is formulated in the equation below (Ifa Musfiroh Nurjannah, et all 2013).

$$TF - IDF(tk,dj) = TF(tk,dj)*IDF(tk,dj) \quad (2.1)$$

Information :
W = TF-IDF = Document weight
dj = j document
tk = k-th term

Where previously calculated in advance on the Term Frequency (TF), which is a frequency of words that exist in a term occurrence in each document. Then calculated Inverse Document Frequency (IDF) is the weight of a term calculated value of the frequency of a term that appears in several documents. The more often a term appears in many documents, the smaller IDF value will be. The TF and IDF formulas are as follows (Ifa Musfiroh Nurjannah, et all 2013).

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$TF(tk, dj) = f(tk, dj) \quad (2.2)$$

Information :
TF = Total frequency term
f = Number of occurrences
dj = j document
tk = k-th term
Then calculate the IDF value using the following equation (Ifa Musfiroh Nurjannah, et all 2013):

$$IDF(t\_k) = 1/(df(t)) \quad (2.3)$$

Or,

$$IDF(t\_k) = \log N/(df(t)) \quad (2.4)$$

Information :
IDF = term weight
N = Total number of documents
df = Number of occurrences of the document
dj = j document
tk = k-th term
Equation 2.3 can only be used if there is only one document to be processed while for equation 2.4 can be used in processes that involve many documents (Ifa Musfiroh Nurjannah, et all 2013).

VSM (Vector Space Model)

Vector Space Model is a method to see the similarities between documents that are represented as a collection of words and can be converted into VSM (Ah Dwijawisnu B 2015). In terms carried out by weighting of a term document which is seen as a vector that has the distance and direction. In addition , the Vector Space Model is a term represented by the dimensions of the vector space. The relevance of a document to a query is based on the similarity between the query vector and the document vector. While the vector space model is made based on the idea that the content of a document is determined by the words used in the document (C Slamet Et All 2018).

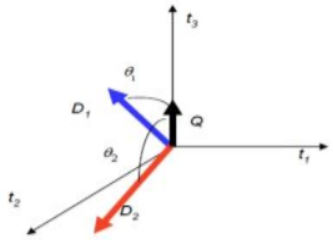

Figure 1. Vector Space Model (VSM)

Information :
Figure above explain that "ti" is the word in the database, "di" is a document, and Q is a word key .
The vector space model (VSM) is a conversion of the bag-of-words representation that defines the similarities between documents. On VSM. Multidimensional vector that represents the documents in the database and user queries. Dimensions according to the number of terms in the documents involved in this model (F.Amin 2016):
1. Vocabulary is a collection of all the different terms that exist after preprocessing and has a term index
2. The weight in the document is given to each term i or query j which has a real value of Wij.
3. Query and document are represented as vector t dimension dj = (W1, W2 , ..., Wtj) and there are n documents in it, namely j = 1, 2, ..., n. Examples of three-dimensional vector space model on two documents, namely D1 and D2, a user query Q1, and three terms T1, T2 and T3 shown in Figure 2.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

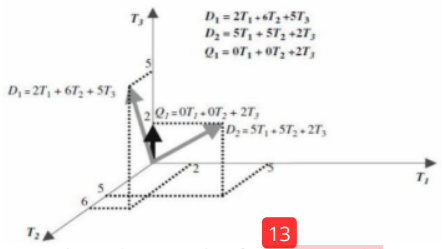

Figure 2. Example of a Space Model

A collection of documents in the vector space model is represented by a term-document matrix (or term-frequency matrix). Figure 2.2 zero value is the value if that term does not exist in the document. Each cell in the matrix corresponding to the weighting given of a term in the documents that have been determined. An example of a term-document matrix for a database with a number of n documents and t terms is a term-document matrix image (Saptono Et All 2018).

	T_1	T_2	T_3	$T_{\_}$	T_t
D_1	W_{11}	W_{21}	W_{31}	□	T_{t1}
D_2	W_{12}	W_{22}	W_{32}	□	T_{t2}
D_3	W_{13}	W_{23}	W_{33}	□	T_{t3}
$D_{\_}$	□	□	□	□	□
D_n	W_{1n}	W_{2n}	W_{3n}	□	T_{tn}

Figure 3. Matrix *Vector Space Model*

Calculation of query distance using equation (1) and document using equation (2)

$$| q | = \sqrt{\sum_{j=1}^{t} (W_{iq})^2} \quad (1)$$

It is explained that |ql is the query distance , while Wiq is the weight of the 1st document query , thus, to get the query distance based on the document query weight (Wiq) from the system, it can be calculated by the query distance (|ql) . The query distance can be calculated using the equation of the square root of the query (Saptono Et All 2018).

$$| d_j | = \sqrt{\sum_{i=1}^{t} (W_{ij})^2} \quad (2)$$

It is explained that |dj| is the distance between documents, while Wij is the weight of the i document, thus , to get the distance of the document based on the weight of the document (Wij) from the system, it can be calculated by the distance between documents |dj|. The distance between documents can be calculated using the root equation of the sum of the squares of the documents. Next to p CALC similaritas query document (inner product), can use the equation (Saptono Et All 2018).

$$sim(q, d_j) = \sum_{i=1}^{t} W_{iq} . W_{ij} \quad (3)$$

Explained that Wij is the weight of term contained in the document, Wiq a weighting query, while Sim (q, dj) is the similarity between the query and document the air in order to get the weights based on the weight of the terms contained in the document (Wij) and weighting query (Wiq) or it can be done by adding up the weight of q multiplied by the weight of the document (F Amin 2013).

Cosine Similarity (cosine value of the angle between two vectors) can be calculated using equation (4) .

$$Sim(q, d_j) = \frac{q . d_j}{|q| * |d_j|} = \frac{\sum_{i=1}^{t} W_{iq} . W_{ij}}{\sqrt{\sum_{j=1}^{t} (W_{iq})^2} * \sqrt{\sum_{i=1}^{t} (W_{ij})^2}} \quad (4)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Sim(q,dj) or Similarit as between query and document is directly proportional to the total weight of the query (q) then multiplied by the weight of the document (dj) and inversely proportional to the square root of q (|q|) then multiplied by the square root of the document (| dj|). The value which approaching document weight value of 1 or larger document than the value resulting from the calculation of the inner product derived from p calculation similarity (F Amin 2013).

System Schematic

Based on the process to be used, as shown in Figure 2.3, the steps taken are explained. The first step is to look for online news sourced from Kompas.com. Then it is done by case folding which functions to change all letters in a document to lowercase, then the tokenizing stage to separate the text, then stopword removal or filtering is carried out to remove unimportant words which usually consist of conjunctions and others. The next step is to do sentence stemming which aims to eliminate word inflection to its basic form. Then the TF-IDF and VSM calculations are performed, so that the final document ranking results are obtained.

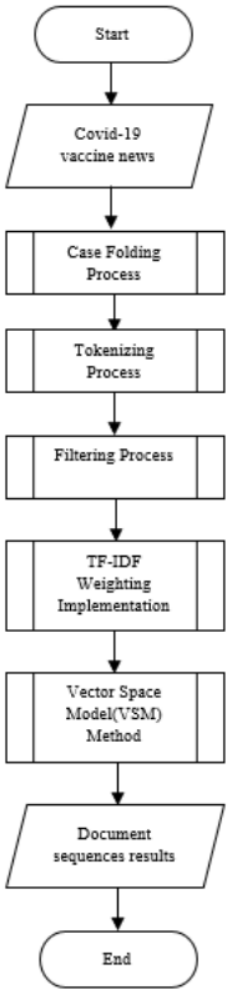

Figure 4. Schematic of the Covid 19 Vaccine Online News System

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

NLTK (Natural Language Toolkit)

The main thing in processing is by downloading the NLTK (Natural Language Toolkit) in the python library. NLTK is a publicly available (open source) module that includes a number of systems that can perform text processing which is very useful for doing work with human languages (Hammond, M. 2020).

***Case Folding* Process**

Case folding aims to change the letters in a document to be a small letter is the letter a to letter z. Characters other than those letters will be omitted and will be treated as delimiters . Furthermore, the process of deleting irrelevant numbers in the document with analyzed document Likewise with numbers, punctuation marks in sentences have no effect on text preprocessing. Remove punctuation

Tokenizing Process

Tokenizing is the process of separating the text into pieces called into token, further separating the sentence into words. Then s topword, the process of removing words that are not used in the text is carried out. Words dictionary save words that can be eliminated is required in this process (Herwijayanti, B. et all 2018). Then analyzed words, numbers, symbols, punctuation and other important entities.

***Filtering* (*Stopword Removal*)**

Filtering is taking the vocabulary that are considered important by the system which is the result of a token, carried out using means removing the less important or it is called an algorithm stopl or done by storing the important word this is called wordlist. Furthermore, the stopword will search for general words that appear in large numbers and are considered to have no meaning.

Stemming

The stemming is the stage of the process of eliminating the inflection of the word into its basic form, but the basic form does not have the same meaning as the root word.

RESULT

Data processing

Data processing is done through the stages of case folding, tokenizing, filtering and stemming for the 5 text about covid-19 in kompas.com news that will be used for further processing using TF-IDF and VSM.

Table 1. Data processing

Doc	Token
1	The vaccine distribution process involves the framework of the cold chain system, vaccine cold chain receives healthy facilities, Bambang hopes that community vaccines guarantee the quality of millions of Sinovac Indonesia vaccines, vaccines, clinical trials, Food Control Agency, BPOM, PT Bio Farma, Vaccination Program, Permit Bags for Emergency, Emergency Use Authorization, Eua, BPOM (Kompas Gramedia. 2021, Jan)
2	Spokesperson for healthy covid vaccination, Siti Nadia Tarmizi, while waiting for the distribution permit for the halal certification of the covid vaccine, to issue a ready-made order, such as national vaccination, Siti Nadia, is ready for health facilities. One of the simulations is giving public vaccines. One of the simulations is giving vaccines. The community is ready for orders. Information systems, information systems, Covid vaccine data systems, validated data, of course, targeting data integration. Ministry of Health BPS data Nadia Nadia shows technical regulations relating to the implementation of the Covid vaccination program are ready (Kompas Gramedia. 2020, Dec).
3	The receiving side of the education messages for the health worker category, check the status of receiving the vaccine, the careprotectiid site, the layout of the main page of the site, the main menu appears, writes, checks the status of a special ID for health workers, the vaccination program, selects check-in, the ID number, the ID card, enter the captcha code, ready, click the list. receive free covid vaccinations, safe and effective vaccines, protect the family of the covid virus (Kompas Gramedia. 2021, Jan).
4	Like the initial covid vaccine, send a notification short message service blast in December, sound, set target policies, such as covid vaccination, priority groups, people receive vaccines according to the law. Public notification of SMS must follow, like Covid vaccination. Beleid, signed by the Minister of Health,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	Cloud Agus Putranto, sells in December. Set up the list of Indonesian citizens for vaccinations (Kompas Gramedia. 2020, Dec).
5	Minister of Health Budi Gunadi Sadikin orders Minister of Health to send short messages to target communities such as covid vaccination Thursday, people receiving SMS are required to participate as vaccinations, unless people's behavior is full of criteria for receiving Covid vaccines according to vaccine indications, 10 en types of Covid vaccines are available, such as Indonesia's Covid vaccine type vaccination. pt bio farma astrazeneca sinopharm moderna novavax inc pfizer inc and biontech sinovac (Kompas Gramedia. 2020, Dec).

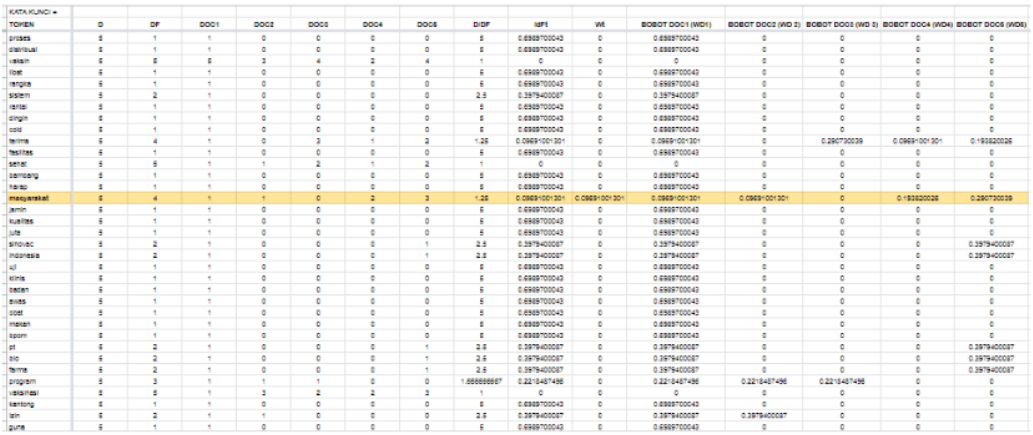

Figure 5 Calculation of TF-IDF

After calculating the TF-IDF as above, the calculation is carried out using keyword weighting as shown in the table below:

Table 2 . Keyword weighting results

Keyword Weighting					
	Doc1	**Doc2**	**Doc3**	**Doc4**	**Doc5**
Society	0,096910013	0,096910013	0	0,193820026	0,290730039
Virus	0	0	0,698970004	0	0
Covid	0	0,387640052	0	0,290730039	0,387640052
Total	**0,096910013**	**0,484550065**	**0,698970004**	**0,484550065**	**0,678370091**

Based on the results above obtained weighting value of a keyword similar between the document 2 and document 4. Therefore to know more rangking it will be calculated by using the Vector Space Model. The following is a calculation using the VSM method, starting with finding the values of |q|, d1, d2, d3, d4, and d5.

Table 3 . *Vector Space Model* Calculation

Vector Space Model					
|q| or Wq	DOC1j	DOC2j	DOC3j	DOC4j	DOC5j
0	0	0	0	0	0,158356251
0	0	0	0	0	0
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, Oct 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11179
e-ISSN : 2541-2019
p-ISSN : 2541-044X

0	0	0	0	0	1,954236268
0	0	0	0	0	0,488559067
0	0	0	0	0	1,954236268
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0	0	1,954236268
0	0	0	0	0	0,488559067
0	0	0	0	0	0,488559067
0	0	0	0,488559067	0	0
0	0	0	0,488559067	0	0
0	0	0	0,488559067	0	0
0	0	0	0,488559067	0	0
0	0	0	0,488559067	0	0
0	0	0	0,488559067	0	0
0,712279558	**3,854367341**	**5,810365714**	**6,026843148**	**3,850406492**	**4,306364259**

Then calculate the term value of each document as follows:

Table 4 . Term Calculation of Each Document

	∑t doc1	**∑t doc2**	**∑t doc3**	**∑t doc4**	**∑t doc5**
Society	0,009391551	0,009391551	0	0,018783101	0,028174652
Virus	0	0,037566202	0	0,028174652	0,037566202
Covid	0	0	0,488559067	0	0
Total	**0,009391551**	**0,046957753**	**0,488559067**	**0,046957753**	**0,065740854**

From the final calculation of the Vector Space Model it is found that :

$$\text{Doc 1 } \frac{\sqrt{\Sigma_t}}{\frac{q}{d_j}} = \frac{\sqrt{0,009391551}}{\frac{0,712279558}{3,854367341}} = 0,524410374$$

$$\text{Doc 2 } \frac{\sqrt{\Sigma_t}}{\frac{q}{d_j}} = \frac{\sqrt{0,46957753}}{\frac{0,712279558}{5,810365714}} = 1,767692186$$

$$\text{Doc 3 } \frac{\sqrt{\Sigma_t}}{\frac{q}{d_j}} = \frac{\sqrt{0,488559067}}{\frac{0,712279558}{6,026843148}} = 5,914226424$$

$$\text{Doc 4 } \frac{\sqrt{\Sigma_t}}{\frac{q}{d_j}} = \frac{\sqrt{0,046957753}}{\frac{0,712279558}{3,850406492}} = 1,171412232$$

$$\text{Doc 5 } \frac{\sqrt{\Sigma_t}}{\frac{q}{d_j}} = \frac{\sqrt{0,065740854}}{\frac{0,712279558}{4,306364259}} = 1,550165096$$

Table 5 . The final result of the Vector Space Model method

Rank	**Document**	**W**
1	3	5,914226424
2	2	1,767692186
3	5	1,550165096
4	4	1,171412232
5	1	0,524410374

DISCUSSIONS

The processed data related to Covid 19 originating from kompas.com news is carried out through several stages, the first step is Case Folding Process which changing the text in the document into lowcase and removing irrelevant numbers. The second step is Tokenizing Process that is separation of text into chunks, then the Filtering Process by removing words in documents that are considered unimportant using the stoplist algorithm. After that, stemming is done to eliminate word inflection, and finally the TF-IDF calculation is done by doing calculations from 5 documents. The results obtained in TF-IDF Document 1 are 0.096910013; Document 2 is 0.484550065;

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Document 3 is 0.698970004; Document 4 0.484550065; and Document 5 is 0.678370091, Documents 2 and documents 4 have the same weighting value, therefore the VSM method is needed to determine the order of the top documents.

After using the TF-IDF method, calculations using the Vector Space Model method is done, the news sourced from Kompas.com can be seen that the search results obtained the order of news documents from the top, namely news document 3 (DOC 3) with a weight of 5.914226424; news document 2 (DOC 2) with a weight of 1.767692186; news document 5 (DOC 5) with a weight of 1.50165096; news document 4 (DOC 4) with a weight of 1.171412232; and the last is news document 1 (DOC 1) with a weight of 0.5244103739.

The data processing is done through the several stages, there are case folding, tokenizing, filtering and stemming for 5 texts regarding covid-19 from kompas.com news which will be used for further processing using the TF-IDF and VSM methods. Based on the final results, it was found that news document 3 occupies the top rank with a weight of 5.914226424, the second rank is news document 2 with a weight of 1.767692186, the third rank is news document 5 with a weight of 1.550165096. The fourth rank is news document 4 with a weight of 1.171412232, and the lowest rank is news document 1 with a weight of 0.5244103739.

From this weighting, we can see that the higher the weight value obtained, the higher the level of similarity of the document to the query.

CONCLUSION

Based on calculations application from 5 online news documents using the TF-IDF method and the Vector Space Model, where the news is sourced from Kompas.com, it can be seen that the search results obtained the order of news documents from the top, namely News Document 3 (DOC 3) with with a weight of 5.914226424, news document 2 (DOC 2) with a weight of 1.767692186, news document 5 (DOC 5) with a weight of 1.550165096, news document 4 (DOC 4) with a weight of 1.171412232, and the last news document 1 (DOC 1) with a weight of 0.5244103739. By using the TF-IDF method and the Vector Space Model, the accuracy value in word weighting in a document can be known.

REFERENCES

F. Wiranto, “Development of a Time Frame Detection System for News Documents Based on a Vector Space Model,” Universitas Jember, 2019.

Rachman, Ff, 2020. An Analysis of the Pros and Cons of Indonesian Society Sentiments regarding the COVID-19 Vaccine on Social Media Twitter. Indonesian of Health Information Management Journal. Vol.8, No.2, December 2020, p.100-109

Abu El-Khair, I. (2017). TF*IDF. In Encyclopedia of Database Systems (pp. 1–2). Springers New York. https://doi.org/10.1007/978-1-4899-7993-3_956-2

Aziz, Abdul, 2015. Implementation of the Vector Space Model in Generating Automatic Frequently Asked Questions and Relevant Solutions for Customer Complaints. Scientific Journal of Informatics. p-ISSN 2407-7658

Kompas Gramedia, 2020 “A Step Towards Vaccines”, Kompas, 12 August 2020, p. 1

Baeza-Yates, Ricardo. Modern Information Retrieval. University of Pompeu Fabra. 1999

Manning, D. Christopher, Raghavan, P. & Schütze H. 2009. An Introduction to Information Retrieval. Cambridge University Press.

R. Melita Et Al., "Application of Term Frequency Inverse Document Frequency (TF-IDF) and Cosine Similarity Methods in Information Retrieval Systems to Know Web-Based Hadith Syarah (Case Study: Syarah Umdatil Ahkam)," J. Tek. Inform., vol. 11, no. 2, 2018.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. Studies in Computational Intelligence, 740, 373–397. https://doi.org/10.1007/978-3-319-67056-0_18

Ifa Musfiroh Nurjannah, Hamdani, “Application of the Term Frequency-Inverse Document Frequency (Tf- Idf) Algorithm for Text Mining” J. Inform. Mulawarman, vol. 8, no. 3, pp. 110–113, 2013.

Ah Dwijawisnu B, “Information Retrieval (IR) Design for Searching Main Idea of English Article Text with Vector Space Model Weighting,” J. Ilm. Technol. and Inf. ASIA, vol. 9, no. 1, 2015.

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

C Slamet Et All "Automated Text Summarization for Indonesian Article Using Vector Space Model" IOP Conference Series: Materials Science and Engineering. IOP Conf. Ser.: Mater. science. eng. 288 012037. 2018

F. Amin, "Search Engine Implementation Using Vector Space Model Method" Din. Tech. (Journal of Development of Technological Sciences , vol. V, no. 1, pp. 45–58, 2016.

Saptono, R., Prasetyo, H., & Irawan, A. (2018). Combination of Cosine Similarity Method and Conditional Probability for Plagiarism Detection in the Thesis Documents Vector Space Model. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(2-4), 139–143. Retrieved from https://jtec.utem.edu.my/jtec/article/view/4332

F. Amin, "Information Retrieval System by Ranking the Vector Space Model Method," vol. 18, no. 2, pp. 122–129, 2013.

Hammond, M. (2020). NLTK. In Python for Linguists (pp. 291–296). Cambridge University Press. https://doi.org/10.1017/9781108642408.013

Herwijayanti, B., Ratnawati, De, & Muflikhah, L. (2018). Online News Classification using TF-IDF Weighting and Cosine Similarity. Development of Information Technology and Computer Science, 2(1), 306–312.

Kompas Gramedia. (2021, Jan). Sinovac Covid 19 vaccine begins to be distributed to 34 provinces [Online] Avaiable : https://nasional. kompas.com/read/2021/01/03/14230441/vaccine-covid-19-sinovac-start-distributed-to-34-province

Kompas Gramedia. (2020, Dec) Update on the preparation for the covid 19 vaccination in Indonesia, where can the vaccine be obtained ? [Online]. Avaiable : https://www.kompas.com/ tren/read/2020/12/ 20/070 000 965 / update-preparation-vaccination-covid-19-in-Indonesia -in-where-vaccine -can ?page =all

Kompas Gramedia. (2021 , Jan). 40 , 2 Million People Will Receive Vaccine Covid-19 Stage One, this breakdown [Online]. Avaiable : https:// www.kompas.com/tren/read/2021/01/02/ 205404 065/402-million-people-will-receive-vaccinecovid-19-the-first-step-this-details?page=all

Kompas Gramedia. (2020 , Dec) Starting today the government will send sms to recipients of the covid 19 vaccine [Online]. Avaiable : https:// nasional.kompas.com/read/2020/12/31/07443471/mulai-hari- this-government-will-send-sms-to-vaccine-covid-19-recipients?page=all

Kompas Gramedia. (2020, Dec) Residents who receive sms from the Ministry of Health must have a covid-19 vaccine [Online] Avaiable : https://nasional.kompas.com/read/2020/12/31/11582611/warga-yang-received -sms-from-kemenkes-mandatory-vaccine-covid-19

*name of corresponding author

TF-IDF Method and Vector Space Model Regarding the Covid-19 Vaccine on Online News

ORIGINALITY REPORT

17%
SIMILARITY INDEX

PRIMARY SOURCES

1	cso.kmi.open.ac.uk Internet	116 words — 2%
2	ebin.pub Internet	99 words — 2%
3	www.ejurnal.stmik-budidarma.ac.id Internet	90 words — 2%
4	journals.codesria.org Internet	80 words — 2%
5	Septya Egho Pratama, Wahyudin Darmalaksana, Dian Sa'adillah Maylawati, Hamdan Sugilar, Teddy Mantoro, Muhammad Ali Ramdhani. "Weighted inverse document frequency and vector space model for hadith search engine", Indonesian Journal of Electrical Engineering and Computer Science, 2020 Crossref	67 words — 1%
6	jist.publikasiindonesia.id Internet	33 words — 1%
7	Ferry Wiranto, Achmad Maududie, Tio Dharmawan. "Time Frame Detection Based on Online News Documents Using Vector Space Model", 2019 International	28 words — 1%

Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), 2019
Crossref

8	Lecture Notes in Computer Science, 2015. Crossref	28 words — 1%
9	"The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)", Springer Science and Business Media LLC, 2020 Crossref	24 words — < 1%
10	ejournal.nusamandiri.ac.id Internet	22 words — < 1%
11	ejurnal.ung.ac.id Internet	22 words — < 1%
12	Utomo Pujianto, Muhammad Fahmi Hidayat, Harits Ar Rosyid. "Text Difficulty Classification Based on Lexile Levels Using K-Means Clustering and Multinomial Naive Bayes", 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019 Crossref	21 words — < 1%
13	journal.utem.edu.my Internet	20 words — < 1%
14	what-when-how.com Internet	19 words — < 1%
15	Submitted to Telkom University Your Indexed Documents	16 words — < 1%
16	en.wikipedia.org Internet	15 words — < 1%

17	bib.irb.hr Internet	14 words — < 1%
18	Fahrudin Mukti Wibowo, Iqsyahiro Kresna A, Aditya Wijayanto. "Model Komunikasi Alternatif dengan Teknologi MANET (Mobile Ad-Hoc Network) untuk Daerah Rural", JRST (Jurnal Riset Sains dan Teknologi), 2021 Crossref	11 words — < 1%
19	Khadjeh Nassirtoussi, Arman, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment", Expert Systems with Applications, 2015. Crossref	11 words — < 1%
20	Ronan Cummins, Colm O'Riordan. "Evolving local and global weighting schemes in information retrieval", Information Retrieval, 2006 Crossref	10 words — < 1%
21	timesofindia.indiatimes.com Internet	10 words — < 1%
22	Astari Diah Ningsih. "Pengaruh Tunjangan Kinerja dan Disiplin Kerja Terhadap Kinerja Pegawai Direktorat Kepolisian Perairan dan Udara Polda Sumatera Selatan", Tanah Pilih, 2021 Crossref	9 words — < 1%
23	Submitted to Telkom University Your Indexed Documents	9 words — < 1%
24	ijrset.in Internet	9 words — < 1%
25	www.ijraset.com Internet	

9 words — < 1%

26 "Proceedings of International Conference on Intelligent Computing, Information and Control Systems", Springer Science and Business Media LLC, 2021 — 8 words — < 1%
Crossref

27 climbtheladder.com — 8 words — < 1%
Internet

28 Deny Haryadi, Dewi Marini Umi Atmaja, Arif Rahman Hakim, Wina Witanti. "Classification of Drug Effectiveness Based on Patient's Condition Using Text Mining With K-Nearest Neighbor", 2022 International Conference on ICT for Smart Society (ICISS), 2022 — 7 words — < 1%
Crossref

29 N. A. V. O. Kaushalye, S. Koswatte. "Crowdsourced Data Relevance Analysis for Crowd-assisted Flood Disaster Management", Journal of Geospatial Surveying, 2021 — 7 words — < 1%
Crossref

30 Nan Sun, Dongsheng Liu, Anding Zhu, Yahui Chen, Yufei Yuan. "Do Airbnb's "Superhosts" deserve the badge? An empirical study from China", Asia Pacific Journal of Tourism Research, 2019 — 6 words — < 1%
Crossref

31 www.msocialsciences.com — 6 words — < 1%
Internet

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF