INTENSIF, Vol.4 No.1 February 2020
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

# Implementation of TF-IDF Algorithm to detect Human Eye Factors Affecting the Health Service System

*Implementasi Algoritma TF-IDF untuk Mendeteksi Faktor Mata Manusia yang Mempengaruhi terhadap Sistem Layanan Kesehatan*

**Received:**
12 November 2019
**Revised:**
29 December 2019
**Accepted:**
7 January 2020

**[1]Made Sudarma, [2*]Juli Sulaksono**
*[1]Universitas Udayana, [2]Universitas Nusantara PGRI Kediri*
*[1]Bali, Indonesia, [2]Kediri, Indonesia*
*E-mail: [1]msudarma@unud.ac.id, [2]jsulaksono@unpkediri.ac.id*
*Corresponding Author

**Abstract** —Elderly is someone whose age is around 60-74 years, at that age, one's health tends to decrease, and it has an impact on reduced perception, cognition, and psychometry. One result of cognitive decline is a decrease in memory. Programs have been provided by the Indonesian government, such as submitting information, producing brochures, and making announcements on the health services website. But this counseling is not optimal because the elderly tend to be lazy to read this because the eyes have begun to look away from other than that the eye health of the elderly has already started to decrease. So that the health information provided by the health department can be optimized, we try to make a model that is used to summarize an article so that the article is easily understood by the elderly. To summarize the article, this study uses the term frequency-inverse document frequency (TF-IDF) algorithm. TF-IDF It is an algorithm used to summarize sentences so that it is easier to understand and understand. By using the TF-IDF algorithm, it is hoped that the elderly will more easily read health articles. User Experience Questionnaire after the application of writing software summary is higher than before the application of writing software summary that is 25.27> 19.30.

**Keywords**—Elderly, Information, Summary, TF-IDF

***Abstrak***— *Lansia adalah seseorang yang usianya berkisar 60-74 tahun, pada usia itu kesehatan seseorang cenderung menurun berdampak pada penurunan persepsi, kognitf dan psikometri. Salah satu akibat penurunan kognitif adalah penurunan memori. Program telah disediakan oleh pemerintah Indonesia, seperti memberikan informasi, memberikan brosur, dan memberikan pengumuman di situs web layanan kesehatan. Tetapi konseling ini tidak optimal karena para lansia, cenderung malas membaca ini karena mata sudah mulai berpandangan jauh selain itu kesehatan mata lansia sudah mulai berkurang. Agar informasi kesehatan yang diberikan oleh dinas kesehatan dapat optimal kami mencoba membuat model yang digunakan untuk meringkas sebuah artikel sehingga artikel tersebut mudah dipahami oleh para lansia. Untuk meringkas artikel, penelitian ini menggunakan algoritma term frequency–inverse document frequency (TF-IDF). TF-IDF Merupakan algoritma yang digunakan urntuk meringkas kalimat sehingga lebih mudah dipahami dan dimengerti Dengan menggunakan algoritma TF-IDF diharapkan lansia akan lebih mudah mudah membaca artikel kesehatan. User Experience Quetionaire sesudah penerapan software summary tulisan lebih besar daripada sebelum penerapan software summary tulisan yaitu 25.27 > 19.30.*

***Kata Kunci***—*Lansia, Informasi, Ringkasan, TF-IDF*

INTENSIF, Vol.4 No.1 February 2020
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

## I. INTRODUCTION

Elderly is someone whose age ranges between 60-74 years [1]; at this age, most elderly experience health problems. Health problems experienced by the elderly are hearing loss, reduced vision, and have started to be senile [1]. The elderly show age-related decline in understanding complex sentences; this is associated with a decrease in cognitive abilities [2], [3] At that age, the health of the elderly has begun to decline.

With the declining health of the elderly, the elderly often do health checks at hospitals, health centers, and clinics. Also, the Indonesian government provides counseling to the elderly to increase the elderly's knowledge about health. Information is conveyed through brochures, notice boards, and health service websites [4]. With the reduced ability to see the elderly, the Indonesian government's efforts are less useful because the elderly are lazy to read. To overcome this problem, we try to examine an application that is used to summarize the information provided by the government [5]. With this tool, it is hoped that the elderly can easily understand the information provided by the Indonesian government.

Text Mining is the discovery of information about data sets [6]. The data collected can be in the form of image data, video data, and text data. The principle for summarizing texts is to mark the passages that appear most often. Easy to summarize text into short stories application to convert data by removing unnecessary words [7].

Some studies use text mining to obtain comparisons of original data with modified data or similar data [8] [9]. Other reviews of text mining use the term frequency-inverse document frequency (TF-IDF) algorithm to provide analysis related to acupoint characteristics and identification of unknown patterns from classical medical texts [10]. The TF-IDF method can also provide data classification results as in other data mining methods such as the c.45 algorithm [11], as in research conducted by Herwijayanti regarding online news classification [12].

In this text, mining research is used to reduce unnecessary words so that the text, the user can understand with secure information on the text [13], [7]. The workings of the system to be built are the elderly photographing announcements available at hospitals, puskesmas, and clinics. After the image is stored on an old smartphone, it will enter the image into the Vision API. Fire Vision is Google's technology for converting images into text.

The extracted text will be changed by the TF-IDF method. TF-IDF (Term Frequency - Inverse Document Frequency) algorithm is an algorithm that can be used to analyze the relationship between a phrase/sentence and a collection of documents. TF-IDF in this study was used to summarize information obtained from the brochure. TF-IDF is a method that creates the highest

INTENSIF, Vol.4 No.1 February 2020
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

weighting value [14] [15]. By utilizing the TF-IDF method, the information collected by the elderly from the brochure can be summarized so that it is easily understood [16].

## II. RESEARCH METHOD

The following figure 1 is step by step used by the system to summarize information. Details can be seen in the flowchart below.

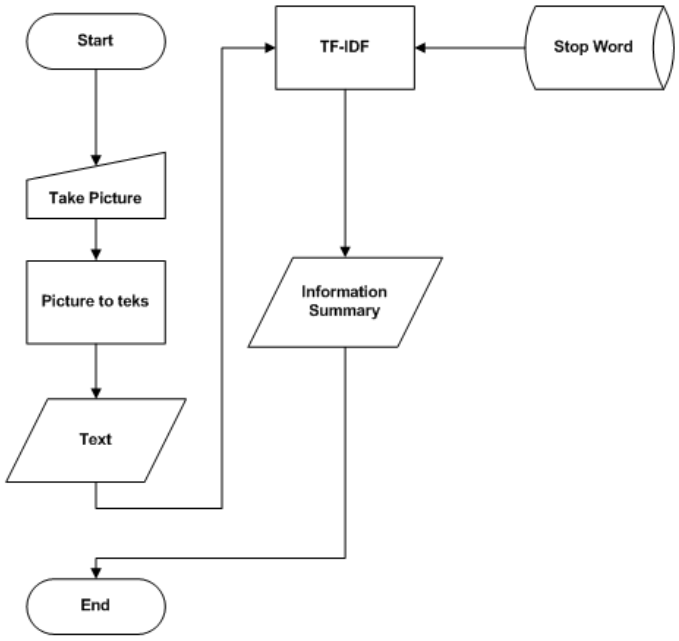


**Figure 1.** FLOWCHART SYSTEM

2.1 Take Picture

The picture is taken from a poster available at a health center, hospital, or clinic. Usually, every time there is counseling by the health department, always put up posters. Pictures were taken using a smartphone camera. After that, the image is converted into text.

2.2 Convert images into text

To convert images that contain information about health into text, Google's API, Google Cloud Vision, is used. With this technology, we can convert images into text.

2.3 Text Processing

Text output is saved in .txt extension files. This file is the original file of the poster, with the amount of text in the poster can be ascertained if the elderly are reluctant to read. Before processing with TF-IDF text processing needs to be done, here are some stages of text processing.

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

a. Tokenization

Tokenization is a process to separate each word string in a sentence, and also includes a process to delete duplicate words, numbers, punctuation marks, characters other than letters of the alphabet & scientific symbols, and change any existing capital letters to lowercase/necessary letters, this process more precisely referred to as the process of uniforming texts, so that they have the same magnitude.

b. Stopword

Stopword is a collection of words that do not have meaning, the function of removing words that do not have the sense in this section to summarize the words that are on the poster one of the words that are considered to have no words in this study are (in order, in, to).

c. Index

This stage is used to mark the remaining words. The remaining words are then weighted so that a summary of the sentence will appear

2.4 TF-IDF

TF-IDF is a weighting/weighting process in which the weight/weight calculation will be calculated for each index term generated at the text preprocessing stage [17]. TF-IDF is an algorithm used to give relationship weight to a word. TF-IDF is a method used to measure how important a word is in a document. The frequency with which a word appears will be used to calculate how important the word is. The weight of a word will be considered large if the analyzed word appears more frequently and allows the word not to be deleted. Whereas words that have a small frequency of occurrence are very likely to be deleted because they are considered not influential. In the TF¬-IDF algorithm, the formula is used to calculate the weight (W) of each document against keywords with the formula that is [18] :

$$Wdt = tfdt * Idft \quad (1)$$

Information:

Wdt= document weight ked to ket

tfdt = the number of words searched for in a document

Idft = Inverse Document Frequency (log (N/df) )

N = total document

df = many documents containing the search term.

INTENSIF, Vol.4 No.1 February 2020
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

## III. RESULT AND DISCUSSION

### A. Sample

Based on calculations using the Slovin formula, the number of samples is 27 people. By utilizing TF-IFD, texts that are initially long and cannot be well understood by the elderly can be easily understood by the elderly. The results of the sampling test were re-tested to 5 older people; from the results of the trial, the elderly still found it difficult. Because they experience visual impairment. 3 The elderly recommends that the summary text be converted into sound. in this case using e=10%(0.1). Berikut ini adalah perhitungannya.

$$n = \frac{N}{1+N(e)^2} \quad (1)$$

$$n = \frac{30}{1+30(0.1)^2}$$

$$n = 23.07$$

### B. Data Normality Testing

In this case, the respondent studied was a member of the elderly posyandu, known that the number of participants was 27 because there were too many, so a sample was needed to represent the population.

### C. Data Normality Testing

The normality test aims to determine whether the data used in this study has a normal distribution or not. Normality testing is carried out by the Kolmogorov Smirnov One Sample test. With the hypothesis tested as follows:

H0: Data is normally distributed

Ha: Data is not normally distributed

The criterion is if the significance value is more significant than 0.05, then H0 is accepted, and Ha is rejected. If the significance value is less than 0.05, then H0 is rejected, and Ha is accepted. Normality testing is performed on each User Experience Questionnaire data before and after the application of the written software summary. The normality test results can be seen in Table 1 as follows:

**Table 1.** DATA NORMALITY TEST RESULT

| Rasio | Period | Kolmogorov-Smirnov Z | Significant Value |
|---|---|---|---|
| *User Experience Questionnaire* | Before applying the software | 0.145 | 0.106 |
| | After applying the software | 0.126 | 0.200 |

INTENSIF, Vol.4 No.1 February 2020
**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

From the table 1, it can be seen that the significance values both in the period before and after the application of the software summary are all higher than 0.05, then H0 is accepted, and H1 is rejected, so it can be concluded that the data are typically distributed. Thus to perform different tests on User Experience Querytaire data before and after the application of written software summary, the parametric statistical method is used, namely paired sm-test t-test.

**E. Hypothesis test**

The purpose of this study is to determine whether the subject's understanding after the application of the written software summary is more significant than before the application of the written software summary. For achieving the research objectives and test the proposed research hypotheses, paired sample t-tests will be conducted on the User Experience Querytaire data. To process the data used computer aids with the SPSS 23.0 program with the following hypothesis:

H0: $\mu \leq 0$: The value of the User Experience Questionnaire after the application of the written software summary is equal to or smaller than before the expected written software summary.
Ha:$\mu > 0$: The value of the User Experience Questionnaire after the application of the written software summary is more significant than before the expected writing software summary.

If the paired sample t-test produces significance $\geq$ 0,05, then Ha is rejected, and H0 is accepted. If the resulting significance value < 0,05, then Ha is approved, and H0 is rejected. The following will be presented with the results of paired sample t-tests in the User Experience Questionnaire data. Table 2 is the results of the paired sample t-test of User Experience Questionnaire data before and after the application of the written software summary:

**Table 2.** PAIRED SAMPLE T-TEST OF USER EXPERIENCE RESULTS

| | *User Experience Questionnaire Before and After Application of Writing Summary Software* |
|---|---|
| t count | -8.306 |
| Significant Value | 0.000 |
| Statistic Descriptive : | |
| Mean *User Experience Quetionaire* Before Application = 19.30 | |
| Mean *User Experience Quetionaire* After Application = 25.27 | |

Tabel 2 is the Result of the Paired Sample t-test of the User Experience Questionnaire Before and after Application of Software Summary Writing Results of Data Normality Test. Based on table 2, it is known that the paired sample t-test produces a significance value of 0,000, where the amount is smaller than 0.05. Thus H0 is rejected, and it is concluded that the average cost of the User Experience Questionnaire after the application of the written software summary is more significant than before the implementation of the written software summary.

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

Judging from the mean (average) value, the User Experience Questionnaire after the application of the written software summary is more significant than before the application of the written software summary, namely 25.27> 19.30. Based on the results of data analysis using paired sample t-test, it can be seen that with the hope that the summary software is beneficial to increase understanding of an object written.

## IV. CONCLUSSION

The TF-IDF algorithm is useful for encapsulating text from health information. The problem faced in this research is to make a good flow system so that it is easy to use by the elderly. Twenty-seven sampling test results, three elderly suggest output is not text, but sound. It is due to the vision of the elderly have started to run away. Judging from the mean (average) value, the User Experience Questionnaire after the application of the written software summary is more significant than before the application of the written software summary, namely 25.27> 19.30. Based on the results of data analysis using paired sample t-test, it can be seen that the application of a written software summary is beneficial in increasing the understanding of an object written.

## REFERENCE

[1] D. Limawan, Y. M. Mewo, and S. H. M. Kaligis, "Gambaran Kadar kalsium serum pada usia 60-74 Tahun," *J. e-Biomedik*, vol. 3, no. 1, 2015.

[2] D. Caplan, G. DeDe, G. Waters, J. Michaud, and Y. Tripodis, "Effects of age, speed of processing, and working memory on comprehension of sentences with relative clauses.," *Psychol. Aging*, vol. 26, no. 2, p. 439, 2011.

[3] J. Yoon, L. Campanelli, M. Goral, K. Marton, N. Eichorn, and L. K. Obler, "The effect of plausibility on sentence comprehension among older adults and its relation to cognitive functions," *Exp. Aging, Res.*, vol. 41, no. 3, pp. 272–302, 2015.

[4] S. M. Carmen, "Importance of Counselling for Elderly Before Institutionalization," *Procedia-Social Behav. Sci.*, vol. 84, pp. 1630–1633, 2013.

[5] D. A. Kurniawati and A. Santoso, "Peningkatan Mutu Pelayanan Kesehatan Usia Lanjut Melalui Peningkatan Kinerja Kader Posyandu Lansia," in *Prosiding Seminar Nasional Unimus*, 2018, vol. 1.

[6] O. S. Kwon, J. Kim, K.-H. Choi, Y. Ryu and J.-E. Park, "Trends in deqi research: a text mining and network analysis," *Integr. Med. Res.*, vol. 7, no. 3, pp. 231–237, Sep. 2018.

[7] M. Sudarma and N. P. Sutramiani, "The Thinning Zhang-Suen Application Method in the Image of Balinese Scripts on the Papyrus," *Int. J. Comput. Appl.*, vol. 91, no. 1, pp. 9–13, 2014.

[8] S. Sucipto, A. G. Tammam, and R. Indriati, "Hoax Detection at Social Media With Text Mining Clarification System-Based," *JIPI (Jurnal Ilm. Penelitian. Dan Pembelajaran Inform.*, vol. 3, no. 2, pp. 94–100, 2018.

[9] I. Y. Anggraini, S. Sucipto, and R. Indriati, "Cyberbullying Detection Modelling at Twitter Social Networking," *JUITA J. Inform.*, vol. 6, no. 2, p. 113, 2018.

[10] T. Lee *et al.*, "Data mining of acupoint characteristics from the classical medical text: Donguibogam of Korean medicine," *Evidence-based Complement. Altern. Med.*, vol. 2014, Dec. 2014.

**ISSN: 2580-409X (Print) / 2549-6824 (Online)**
**DOI:** https://doi.org/10.29407/intensif.v4i1.13858

[11] Sucipto, Kusrini, and E. L. Taufiq, "Classification method of multi-class on C4.5 algorithm for fish diseases," in *Proceeding - 2016 2nd International Conference on Science in Information Technology, ICSITech 2016: Information Science for Green Society and Environment*, 2016, pp. 5–9.

[12] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, "Klasifikasi Berita Online Dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," *J. Pengemb. Teknol. Inf. dan Ilmu Komputer. e-ISSN*, vol. 2548, p. 964X, 2018.

[13] S. S. Bhanuse, S. D. Kamble, and S. M. Kakde, "Text mining using metadata for generation of side information," *Procedia Comput. Sci.*, vol. 78, pp. 807–814, 2016.

[14] H. Gupta, A. Kottwani, S. Gogia, and S. Chaudhari, "Text analysis and information retrieval of text data," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2016, pp. 788–792.

[15] S. Jabri, A. Dahbi, T. Gadi, and A. Bassir, "Ranking of text documents using TF-IDF weighting and association rules mining," in *2018 4th International Conference on Optimization and Applications (ICOA)*, 2018, pp. 1–6.

[16] S. Kalra, L. Li, and H. R. Tizhoosh, "Automatic Classification of Pathology Reports using TF-IDF Features," *arXiv Prepr. arXiv1903.07406*, 2019.

[17] W. Dai, "Improvement and Implementation of Feature Weighting Algorithm TF-IDF in Text Classification," in *2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018)*, 2018.

[18] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation," *arXiv Prepr. arXiv1806.06407*, 2018.