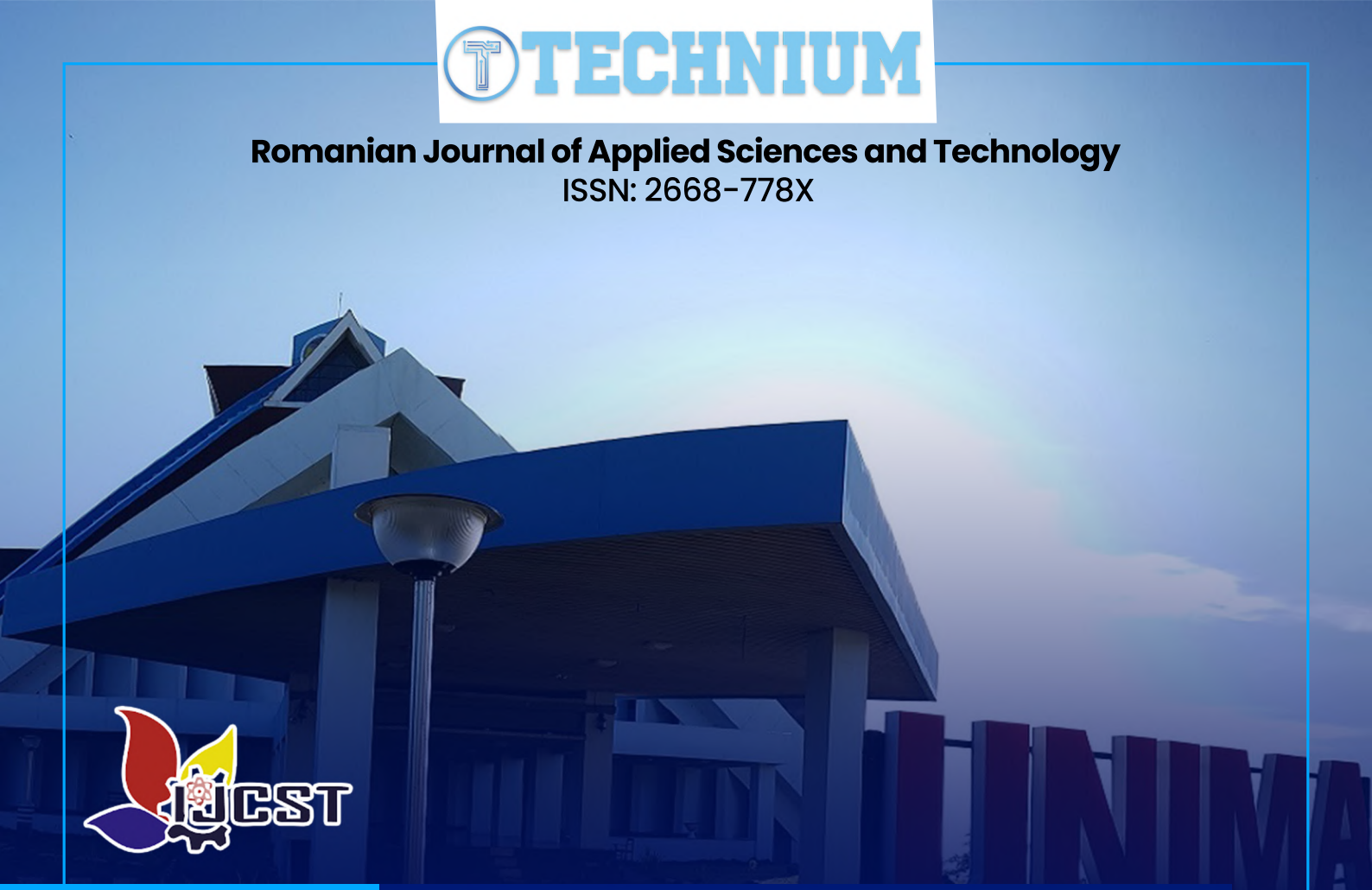
TECHNIUM

Romanian Journal of Applied Sciences and Technology
ISSN: 2668-778X



vol. 17 / 2023

The 7th International Conference on Science Technology

organized by
Faculty of Social Science and
Law Universitas Negeri Manado and
Consortium of International Conference
on Science and Technology

The Innovation Breakthrough in Digital and Disruptive Era

PLUS COMMUNICATION

International Communication & PR

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 17, 457-462, November, 2023
ISSN: 2668-778X
www.techniumscience.com

Medical Information Retrieval with Weighting Critical Score for Acute Respiratory Infection (ARI) Desease Detection

Fika Hastarita Rachman[1*], *Rike Ayu Arista*[1], *Ika Oktavia Suzanti*[1], *Yonathan Ferry Hendrawan*[1], and *Aryono Yerey Wibowo*[2]

[1]Informatics Departement, University of Trunojoyo Madura, Indonesia
[2] Assyafi'i Sentosa Lengkong Clinic, Nganjuk, Indonesia

Abstract. Medical Information Retrieval (Med-IR)is part of computer science that discusses the search for a medical document. Medical Information Retrieval is needed by patients to know the initial prediction of the symptoms they are experiencing. ARI (Acute Respiratory Infection) is a disease that almost everyone has experienced which can cause death. This study uses a dataset of ARI sufferers and user queries that contain symptoms in text form. Furthermore, the query data is processed with the Med-IR application using Bi-Gram, TF-IDF as the feature extraction and Cosine Similarity as the similarity method, so that a return document is produced which is expected to be used as an early prediction of ARI in patients. The research also uses a critical disease wighting process, so that the results of the Med-IR are complemented by predictions of the severity level of the disease. From the results of research conducted at the Assyafi'u Sentosa Lengkong Clinic, Nganjuk, the best results were obtained for precision values of 85.5% and 52.9% for recall values . The evaluation of disease severity with Mean Absolute Percentage Error (MAPE) getting a low score of 2,529%. Keyword : Medical Information Retrieval, ARI, Weighting critical disease, Bi-Gram, TF-IDF, Cosine Similarity.

* Corresponding author: fika.rachman@trunojoyo.ac.id

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 17, 457-462, November, 2023
ISSN: 2668-778X
www.techniumscience.com

1 Introduction

There are obstacles that prevent people from going to doctors or health services, even to the point of trivializing a disease. Especially if you only experience coughs and colds, this disease is a disease that almost all humans have experienced, even this disease does not look at age, from toddlers to the elderly. This disease is often underestimated by the community because of the lack of knowledge that coughs and colds are Acute Respiratory Infection (ARI).

ARI is an infectious disease that can cause death, so this disease cannot be underestimated [1]. Several diseases are classified as ARI, namely Upper Respiratory Infections and Lower Respiratory Infections [2]. The types of diseases that are often suffered by tropical people are 5 diseases: Common Cold (CC), Pharyngitis (P), Laryngitis (L) including Upper Respiratory Tract Infection and Bronchitis (Br), and Pneumonia (Pn) which includes Lower Respiratory Tract Infection [3].

Table 1. Types of ARI

Disease name	Detail
Common Cold	A respiratory disease that spreads easily and attacks the nose, the main cause of this disease is a virus, therefore it becomes easily transmitted, more than 200 viruses can cause the common cold.
Pharyngitis	Inflammation of the esophagus, the main cause of which is bacteria, viruses, and can also be caused by fungi. This disease occurs in the area at the back of the throat.
Laryngitis	Inflammation that occurs in the larynx. If it is less than 3 weeks it is called acute inflammation, whereas if it is more than 3 weeks it is said to be chronic.
Bronchitis	Bronchial inflammation of the lungs. If this disease is examined using a stethoscope, crackles will be heard (abnormal breathing sounds). To diagnose it based on some of the symptoms he experienced and especially from the presence of mucus.

To diagnose the symptoms caused by ARI, it requires the help of a knowledgeable doctor. Symptoms of ARI vary greatly depending on the cause, so it will be difficult for the general public to make an accurate diagnosis like a diagnosis from an expert or doctor [4]. Medical Information Retrieval that can return medical documents that match the symptoms experienced by patients is needed in the early prediction of ARI.

Med-IR for text documents has been done [5][6]. Research [7] also uses the same object, ARI, but uses a classification method to detect ARI disease. Research [8] compare Cosine Similarity, Jaccard Similarity and Dice Coefficient for automatic assessment of short answers with questions and answers in Indonesian. This study shows that the cosine similarity method is able to obtain the highest correlation. So in this study using the cosine similarity method for Med-IR.

In real conditions, based on symptom data obtained from the inpatient clinic doctor Assyafi'u Sentosa, Lengkong generally consists of two words, for example coughing up phlegm, dry coughing, coughing up blood. Previous research did not pay attention to this problem, so the TF-IDF method used unigram conceptin the cosine similarity method still uses the unigram concept. Based on these problems, this study intends to develop an innovative Med-IR ARI medical document that can provide predictions of ARI disease and the severity of the disease.

2 Research Method

Figure 1 shows the system architecture for making an expert system with input in the form of user symptom text queries using the Bi-Gram, TF-IDF and cosine similarity methods.

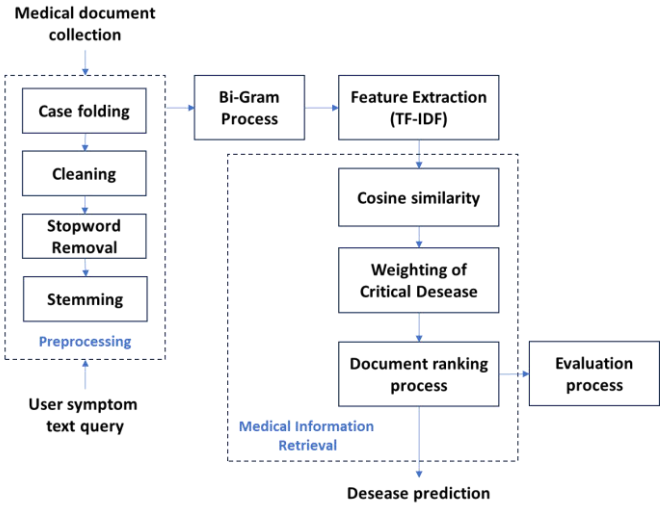


Fig. 1. Proposed Method

2.1 Dataset

There are two data that are input into the system in the early stages, namely expert documents which are documents in the form of symptoms and diagnoses from doctors, while user essay queries are in the form of symptoms that are input by the user.

The dataset used for this information retrieval system uses data from the medical records of ARI from the Assyafi'u Sentosa Lengkong clinic for 100 patient data along with the severity of their disease. Data collection begins September 2020 – September 2021. There are 5 types of diseases that are classified as ARI and 24 symptoms of ARI disease along with the weight of symptoms for each disease. The types of diseases used are Common Cold (CC), Pharyngitis (P), Latyngitis (L), Bronchitis (Br) and Pneumonia (Pn). Symptoms of ISPA disease used in the extraction process [9] and approved by experts from the Assyafi'u

TECHNIUM
Technium: Romanian Journal of Applied Sciences and Technology
Vol. 17, 457-462, November, 2023
ISSN: 2668-778X
www.techniumscience.com

Sentosa Lengkong Nganjuk Inpatient Clinic doctor as shown in Table 1.

The output of this study is returning medical documents that have been suffered by patients to users and predicting the diseases suffered by users based on the order of document ranking results. The assumption is that the top ranking is similar to the Query User document. In addition, the severity of the disease is also detected by adding the weight of each symptom from the process of weighting of critical disease.

Table 1. Symptoms of ARI

Code	Symptom
G01	High fever
G02	Cough with phlegm
G03	Dry cough
G04	Nasal congestion
G05	Headache
G06	Sore throat
G07	Swallow pain
G08	Weak body
G09	Out of breath
G10	Sneezes
G11	Breathing frequency
G12	quick breath
G13	Breath sound
G14	Rough breath
G15	Appetite
G16	Eat less
G17	Hoarseness
G18	Chest pain
G19	The sense of smell
G20	Reduced smell
G21	Bleeding cough
G22	Nausea or vomiting
G23	Runny nose
G24	Watery eyes

Sample Indonesian Medical document
Expert Document (ED)

Saya mengalami batuk kering seminggu ini, sehingga membuat saya nyeri dada hingga sesak nafas. Badan saya lemas karena sering mual dan muntah ketika makan. Saya juga mengalami demam tinggi dan sakit kepala.
(I have had a dry cough this week, so it makes me have chest pain and shortness of breath. My body is weak because I often feel nauseous and vomit when eating. I also have a high fever and headache.)

User symptom text query (USQ)

Saya mengalami sakit kepala dan demam tinggi, saya juga mengalami batuk kering hingga saya sakit tenggorokan, badan saya lemas karena sering mual dan muntah setelah saya makan. Saya juga terkadang merasakan nyeri di dada dan frekuensi nafas saya cepat dan pernah mengalami sesak nafas.
(I have a headache and high fever, I also have a dry cough that causes a sore throat, my body is weak because I often have nausea and vomiting after I eat. I also sometimes feel pain in my chest and my breathing rate is fast and I have experienced shortness of breath.)

2.2 Preprocessing

The preprocessing stage aims to make the data more structured to facilitate further processing. In this research, the preprocessing stage uses a literary library. The case folding stage is the first stage of text preprocessing. This step is used to change the text to lowercase. The next stage is cleaning, which is used to remove punctuation marks, numbers and empty characters which are likely to become obstacles in the next process.

The stopword removal stage is used to remove words that have no meaning to be deleted. Examples are conjunctions “yang”, “dengan” ,”dan”, “dari”. The final stage is stemming. Stemming is used to change the form of a word into a basic word. In this stage a literary library is used which basically applies the Nazief Adriani Stemming Algorithm method. The final results of the preprocessing stages are shown in Table 2.

Table 2. Example of preprocessing result document

Document	Result of Stemming in Indonesian
USQ	alami sakit kepala demam tinggi juga alami batuk kering hingga sakit tenggorok badan lemas sering mual muntah makan juga terkadang rasa nyeri dada frekuensi nafas saya cepat pernah alami sesak nafas
ED	alami batuk kering minggu sehingga buat nyeri dada hingga sesak nafas badan lemas sering mual muntah makan juga alami demam tinggi sakit kepala

2.3 Bi-Gram Process

After the next preprocessing is the process of forming a bi-gram matrix document using the bigram concept. Bigram is part of the N-gram.

N-grams are a set of words where each has a length of n words. Based on the number of gram substring pieces, n-grams consist of uni-grams, bi-grams, tri-grams, squad-grams and so on the number of n in n-grams [10]. N-grams are often used in word processing (text mining) and also often used in language processing or often called natural language processing [11]. The use of n-grams also greatly affects the level of accuracy or the level of similarity of the documents being compared. An example of the results of the bi-gram process for ED is shown in table 3.

Table 3. Example of Bi-Gram Process

alami sakit	sakit kepala	kepala demam	demam tinggi
tinggi juga	juga alami	alami batuk	batuk kering
kering hingga	hingga sakit	sakit tenggorok	lemas sering

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 17, 457-462, November, 2023
ISSN: 2668-778X
www.techniumscience.com

2.4 Feature Extraction using TF-IDF

Term frequency (TF) is a term weighting method whose function is used to determine the weight of a word on the basis of the number of occurrences of a word in a document or term frequency [12]. So to determine the weight of a word, that is through the number of occurrences of a word. So it can be concluded that the higher the number of word occurrences, the higher the suitability value will be [13][14].

$$W_{TF}(t_i, d_j) = f(t_i, d_j) \quad (1)$$

$W_{TF}(t_i, d_j)$: tf term- value in document-j
$f(t_i, d_j)$: number of occurrences oterm-i in document-j

Furthermore, the Inverse document Frequency (IDF) weighting process has the aim of knowing how important the influence of a term is on a document with other documents. In a document can contain terms that are very valuable but rarely appear.

$$W_{IDF} = 1 + log\frac{D}{d(t_i)} \quad (2)$$

$W_{IDF}(t_i, d_j)$: IDF value of term-i in document-j
d_{t_i} : number of documents containing term-i
D : number of documents

After searching for the weighting values at the TF and IDF stages, the next step is the TF-IDF stage. The following is the formula for expressing the weight of the processed document to the key document.

$$W_{tf-idf}(t_i, d_j) = W_{TF}(t_i, d_j)\ x\ W_{IDF}(t_i, d_j) \quad (3)$$

$W_{tf-idf}(t_i, d_j)$: tf-idf term-i value in document-j
$W_{IDF}(t_i, d_j)$: term-i IDF value in document-j
$W_{TF}(t_i, d_j)$: term-i tf value in j-document

2.5 Cosine Similarity

The cosine similarity method is a method used to determine the similarity value between two objects. The similarity value of the two documents resulting from the cosine similarity method is obtained from two vectors that compare two text documents where the values used range from 0 to 1, then the documents are considered similar when the cosine value is 1 [15].

2.6 Weighting of Critical Desease

This stage is used to determine the severity of the patient's disease. The simplicity of this process is to add up the weights contained in the symptom words contained in the USQ. The weight of each symptom for each disease is shown in table 4. This weight is used to determine the severity of the disease. The severity category is from dr. Aryono Yerey Wibowo, MM doctor of the Assyafi'u Sentosa Lengkong Nganjuk Inpatient Clinic.
<= 30 = Mild
>30 - 60 = Moderate
>60 - 100 = Weight

This happens because there may be symptoms that are shared by different diseases, but there are symptoms that are very visible and are the main signs in detecting certain diseases. For example in Table 4, G07 is owned by P and L, but for each disease the weight for G07 is different because the probability of G07 appearing in P disease is more decisive than L disease.

Table 4. Weight Symptoms of each disease

Symptom Code	Desease Code				
	CC	*P*	*L*	*Br*	*Pn*
G01	10	5	5	5	5
G02	10	10	10	5	5
G03	10	10	10	10	10
G04	10		5	10	
G05	10	5	5		5
G06		20	20		
G07		20	15		
G08				15	10
G09				10	10
G10	10	5			
G11					5
G12					5
G13				10	2,5
G14				10	2,5
G15		5	5		
G16		5	5		
G17	10	10	10		
G18					15
G19	5				5
G20	5				5
G21				10	5
G22		5	10		10
G23	10			15	
G24	10				
Total weight	100	100	100	100	100

This weighting is used to determine the value of each symptom. To get the weight of each symptom requires the help of a doctor as an expert, both general symptoms and specific symptoms.

2.7 Evaluation System

In making a system application, an evaluation or process of testing the system is needed, for information retrieval systems are tested using precision and recall. Precision is used to measure the system's ability to return only relevant documents, while recall is used to measure the system's ability to return all relevant documents [19].

$$Precision = \frac{Dr}{Dt} \quad (4)$$

$$Recall = \frac{Dr}{Nr} \quad (5)$$

Dr : Number of relevant documents obtained
Dt : The total number of documents obtained
Nr : The total number of relevant documents in the collection

Meanwhile, the evaluation on severity weighting uses the MAPE formula. MAPE is used to determine

the error level between the severity of the disease from the doctor and the system results. If the lower the MAPE number, the lower the error rate will be.

$$MAPE = \sum_{k=1}^{m} 1/n \left| \frac{\text{actual} - \text{prediction}}{\text{actual}} \right| * 100\% \quad (6)$$

3 Results and Analysis

Testing is done by dividing the dataset into training and testing sections with 3 test scenario models. Testing1 the percentage of the training: testing part is 90%:10%, Testing2 the percentage of the training:testing part is 80%:20%, and Testing3 the percentage of the training:testing part is 70%:30 %. The test results are shown in table 5.

Table 5. System test results

Testing	Mean of MAPE	Mean of Precision	Mean of Recall
Testing 1	4.3452%	0,855	0,529
Testing 2	2.5297%	0,773	0,547
Testing 3	3.1150%	0,812	0,517

In this system analysis, precision and recall calculations for scenario 1, 2, 3 trials using a threshold of 0.5 get different results. the common cold has a total of 37 out of 100 data while in laryngitis there are only 2 out of 100 data. Scenario 1 has the highest average value where the precision value is 0.855 which means that this system can return 85.55% of relevant documents while the recall value is 0.529 which means that the information retrieval system is able to obtain 52.9% of relevant documents from all relevant documents. in the document used.

Meanwhile, the value of similarity in disease outcomes between system results and doctor results, and the error rate is the difference in weight between system results and doctor results. Then what affects the accuracy and error rate in this system is the error to type in symptoms and other names for symptoms, for example, high fever has the same word meaning as high fever. Where if high fever contains weight in the system while high heat has no weight so it is not detected. Then the number of data divisions does not affect the results of accuracy and error rate.

4 Conclusion

Based on precision and recall testing using a threshold of 0.5 on an information retrieval system using the bi-gram and cosine similarity methods, the best average value is in the test scenario with 90% training data distribution and 10% testing data with the average value is 0.855 for precision and 0.529 for recall. For the error rate using the MAPE calculation the best value is in the test scenario with the distribution of 80% training data and 20% testing data with the smallest error rate of 2.5297%.

So it can be concluded that this medical information retrieval system can provide good search results with a precision value of 85.5% and a recall value of 52.9% and has a fairly low error rate on system weighting, namely 2.5297%.

The development of further research based on the analysis of the trial results can be carried out by adding the process of synonym recognition in order to find out the similarity of meaning for different terms. So it is expected to increase the precision value

ACKNOWLEDGEMENTS

We also want to thank the **Assyafi'u Sentosa Clinic, Lengkong, Nganjuk** for allowing me to collaborate and provide research datasets. We also thank **dr. Aryono Yerey Wibowo, MM.,** Doctor who is an expert in this research. Project completion would not have been possible without their help and insight.

References

[1] S. Fatmawati, M. Awal, and M. Rifai, "Risks Affecting the Incidence of Acute Respiratory Infections in Toddlers," *Jurnal Ilmiah Kesehatan Sandi Husada*, vol. 10, no. 2, 2021.

[2] C. Dasaraju, Purushothama V Liu, "Infections of the Respiratory System," in *Medical Microbiology. 4th edition.*, 1996.

[3] N. Medicinewise, "Respiratory tract infections (RTIs) – nose, throat and lungs," 2023. [Online]. Available: https://www.nps.org.au/consumers/respiratory-tract-infections-rtis-nose-throat-and-lungs. [Accessed: 20-Aug-2023].

[4] R. Tullah, F. H. Saputri, and R. Prasetya, "EXPERT SYSTEM FOR DETECTING WEB-BASED RESPIRATORY TRACT INFECTIONS AT THE KALIDERES DISTRICT HEALTH CENTER," *Jurnal Sisfotek Global*, vol. 12, no. 2, 2022.

[5] A. J. P. L, S. Sengan, K. G. K, V. J, J. Gopal, and S. V Priya Velayutham, "Medical information retrieval systems for e-Health care records using fuzzy based machine learning model," *Microprocessors and Microsystems*, 2020.

[6] B. Sugara, Dody, and Donny, "Sistem Temu Kembali Informasi Pada Gejala Autisme Dengan Metode Vector Space Model," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, 2019.

[7] Z. E. Fitri, L. N. Sahenda, P. S. D. Puspitasari, P. Destarianto, D. L. Rukmi, and A. M. N. Imron, "The The Classification of Acute Respiratory Infection (ARI) Bacteria Based on K-Nearest Neighbor," *Lontar Komputer*, vol. 12, no. 2, 2021.

[8] T. Wahyuningsih, Henderi, and Winarno, "TEXT MINING AN AUTOMATIC SHORT ANSWER GRADING (ASAG), COMPARISON OF THREE METHODS OF

COSINE SIMILARITY, JACCARD SIMILARITY AND DICE'S COEFFICIENT," *Journal of Applied Data Sciences*, vol. 2, no. 2, 2021.

[9] R. T. Feraldy, F. Iskandar, and H. E. T. Esti, "Sistem Pakar Diagnosa Penyakit ISPA Berbasis Web Dengan Metode Forward Chaining," *JOINTECS*, vol. 5, no. 2, 2020.

[10] R. S. Citawan, V. C. Mawardi, and B. Mulyawan, "Automatic Essay Scoring in E-learning System Using LSA Method with N-Gram Feature for Bahasa Indonesia," in *ICESTI*, 2017, vol. 01037, pp. 1–17.

[11] T. Georgieva-Trifonova and M. Duraku, "Research on N-grams feature selection methods for text classification," in *IOP Conference Series: Materials Science and Engineering*, 2021.

[12] I. Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF And Logistic Regresion," in *Information Technology International Seminar (ITIS)*, 2020, pp. 238–242.

[13] S. W. Kim and J. M. Gil, "Research paper classification systems based on TF - IDF and LDA schemes," *Human-centric Computing and Information Sciences*, pp. 9–30, 2019.

[14] F. H. Rachman and F. Solihin, "APLIKASI TEXT TO SPEECH DALAM SISTEM PENERJEMAH BAHASA INDONESIA-MADURA MENGGUNAKAN METODE FSA (FINITE STATE AUTOMATA)," *Jurnal Sarjana Teknik Informatika*, pp. 1–10.

[15] A. Jauhari, I. O. Suzanti, N. Pangestika, W. Diantisari, and Y. D. Pramudita, "Enhanced Confix Stripping Stemmer And Cosine Similarity For Search Engine in The Holy Qur ' an Translation," in *Information Technology International Seminar (ITIS)*, 2020, pp. 207–212.