IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.17, No.1, January 2023, pp. 45~54
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.80077

Information Retrieval for Early Detection of Disease Using Semantic Similarity

Aszani[1], Hayyu Ilham Wicaksono*[2], Uffi Nadzima[3], Lukman Heryawan[4]
[1,2,3,4]Department of Computer Science and Electronics, FMIPA UGM, Yogyakarta, Indonesia
e-mail: [1]aszani@mail.ugm.ac.id,[*2]**hayyuilhamwicaksono@mail.ugm.ac.id**,
[3]uffinadzima@mail.ugm.ac.id, [4]lukmanh@ugm.ac.id

Abstrak

Pertumbuhan rekam medis yang terus meningkat perlu dimanfaatkan dengan baik untuk meningkatkan kinerja dokter dalam mendiagnosis suatu penyakit. Sebuah metode temu kembali informasi diusulkan untuk memberikan rekomendasi diagnosis berdasarkan gejala yang diuji dengan dataset rekam medis. Peneliti mengeksplorasi teknologi temu kembali informasi dengan menerapkan metode TF-IDF dan cosine similarity. Tantangan pada penelitian ini adalah gejala pada dataset rekam medis merupakan data kotor yang diperoleh dari pasien yang kurang akrab dengan istilah-istilah biologis. Oleh karena itu dilakukan pencocokan gejala pada data rekam medis dengan istilah gejala yang digunakan dalam sistem dan dari hasilnya tersebut dilakukan augmentasi data untuk meningkatkan jumlah data hingga sekitar 3 kali lebih bamyak. Pada TF-IDF hasil akurasi tertinggi dengan $k \in (1,10)$ *hanya* $\approx 16\%$ *sedangkan setelah dilakukan augmentasi pada data ujinya, akurasi menjadi* $\approx 20\%$*. Hasil akurasi tertinggi dengan nilai k yang sama dengan metode cosine similarity* $\approx 9{,}5\%$ *dan dengan data uji yang telah diaugmentasi akurasinya meningkat menjadi* $\approx 16\%$*. Dari penelitian ini disimpulkan bahwa pada sistem dengan masukkan gejala-gejala dengan jumlah yang cukup dan relevan akan memberikan prediksi penyakit yang lebih akurat. Hasil prediksi menggunakan metode TF-IDF dengan* $k \in (1,10)$ *lebih akurat dibanding dengan prediksi menggunakan metode cosine similarity.*

Kata kunci*—Augmentasi Data, Cosine Similarity, Disease Detection, Information Retrieval, TF-IDF*

Abstract

The growth of medical records continues to increase and needs to be used to improve doctors' performance in diagnosing a disease. A retrieval method returns proposed information to provide diagnostic recommendations based on symptoms from medical record datasets by applying the TF-IDF and cosine similarity methods. The challenge in this study was that the symptoms in the medical record dataset were dirty data obtained from patients who were not familiar with biological terms. Therefore, the symptoms were matched in the medical record data with the symptom terms used in the system and from the results, data augmentation was carried out to increase the amount of data up to about 3 times more. In the TF-IDF the highest accuracy with $k \in (1,10)$ *is only* $\approx 16\%$*, while after augmentation of the test data, the accuracy becomes* $\approx 20\%$*. The highest accuracy results with the same k value using the cosine similarity method is* $\approx 9.5\%$ *and with the augmented test data accuracy increasing to* $\approx 16\%$*. From this study it was concluded that a system with sufficient and relevant input of symptoms would provide a more accurate disease prediction. Prediction results using the TF-IDF method with* $k \in (1,10)$ *are more accurate than predictions using the cosine similarity method.*

Keywords*—Cosine Similarity, Data Augmentation, Disease Detection, Information Retrieval, TF-IDF*

Received December 12th,2022; Revised December 22th, 2022; Accepted January 31th, 2023

1. INTRODUCTION

Disease diagnosis technology is used to find appropriate and relevant disease information quickly. The search was obtained from a massive collection of medical and clinical record data [1]. The medical record is a file containing notes and documents about patient identity, examination, treatment, actions and other services for patients in health facilities that are carried out manually or electronically [2]. The growth in the number of medical record data that continues to grow can be used as information material and a decision-support tool in making a diagnosis. The correct early-stage diagnosis of a disease will positively influence the prevention of more dangerous risks. Classifying a disease symptom into an ICD code to help provide recommendations for disease diagnosis is also very laborious and time-consuming. An experienced professional takes about 20 minutes per case [3]. Therefore it is necessary to have a system that can help improve the accuracy of the diagnosis and reduce the time needed to determine the patient's diagnosis [4]. This system is expected to facilitate and speed up patients' health care and treatment process.

A correct diagnosis at an early stage is critical to saving lives significantly. In previous research, the design and development of a decision support system for early diagnosis of congenital heart disease using the GUI Matlab has been proposed with the implementation of a Backpropagation Neural Network. The specific dataset used in that study is signs, symptoms and results of physical evaluation of heart disease in patients. The system can reduce delays and inaccuracies in diagnosis [4]. The information retrieval model is applied to search engines in many types of research. The results show that using the information retrieval model is optimal for retrieving relevant information, updating data, and adapting to user needs [5]. Other related research for this research uses the TF-IDF Algorithm and Cosine Similarity in a seminar recommendation system. with an approach through text or words from a combination of titles and descriptions at a seminar which functions to find valuable and helpful information in a text [6].

In this study, information retrieval technology is proposed to diagnose the early stage of a disease based on the symptoms. The researcher implemented the TF-IDF algorithm and cosine similarity methods to predict the patient's disease. The output of this system is recommendations for several diseases relevant to the symptoms inputted. The prediction results are in the form of diseases with the highest match score based on the symptoms query. The disease detail describes based on the Wikipedia infobox and ICD 10 code for each disease. The goal is that the system will not only provide predictions for the diagnosis of certain diseases but also provide knowledge regarding the disease experienced by the patient. Therefore, the developed system can contain the principles of a retrieval system, namely providing relevant information, updating data, continuously adapting to user needs, improving device performance, and solving problems encountered.

We used the top-k-accuracy method to evaluate prediction system by including all available disease symptoms in the rekam medis dataset. Evaluation is used to get disease prediction accuracy scores and whether it matches the diagnosis of each symptom in the test data. To increase the data test, researchers perform data test augmentation to identify possible variations of similar symptoms more optimally. The problem of information retrieval is determining which information has the best relevance according to a particular user's request [5]. Therefore, the evaluation results of the designed system need to be analyzed to determine the accuracy predictions from the designed system, the importance of queries with relevant symptoms and which techniques are superior for information retrieval systems for disease diagnosis can be helpful in further research and collection of symptoms in medical record data that is more systematic and synchronous.

This paper is divided into four sections. This section, Introduction, gives the background of the problem and general information about this paper. Section 2 gives information about the dataset and some knowledge for this research. Section 3 presents the

implementation of the system and the evaluation process of the methods used. The last section, Section 4, presents the conclusion and future works.

2. METHODS

2.1 Dataset

The dataset used in this study was taken from several different sources:

a) Dataset obtained from GitHub repository Disease Detection by Symptoms with treatment recommendations [7]. The dataset consisted of 261 diseases with 489 symptom features in English, rated 0 and 1. The researcher scraped through the National Health Portal of India website. The dataset is translated into Indonesian to be implemented in the system designed in this study.
b) The ICD 10 dataset is combined with the data in the previous point as details for disease codes according to international standards.
c) The dataset from the Rekmed website is in the form of a patient's medical record containing the symptoms experienced by the patient and the diagnosis given. This data is used as test data for system evaluation, namely the symptoms experienced by patients as input to the system and disease diagnoses given as output.
d) Additional information regarding the output of the predicted disease diagnosis details retrieved and translated from the Infobox Wikipedia.

2.2 Literature Review

An information retrieval system is a search for material (usually documents) from something unstructured (usually text) and fills in information from extensive collections stored on a computer [8]. The goal of information retrieval is to provide the best possible information according to the user's needs from the stored database. Search problem information is determining which information is best relevant to the request of specific users. The information retrieval system has two main stages: index construction/indexing and query processing to retrieve the appropriate documents/information.

2.2.1 TF-IDF

TF-IDF is a technique for measuring a word in a document. At this stage, calculating each word's weight indicate its importance in the document and corpus. In this study, symptoms are terms, and disease is a document. The TF-IDF (Term Frequency-Inverse Document Frequency) method is used to weigh document terms. This method is used for information retrieval, a statistical weighting technique on the text. The weight on the TF-IDF is the frequency of occurrence of a word/term i in each document j which is indicated as Term Frequency denoted by $tf_{i,j}$ and the total occurrence of a word/term in all documents is denoted by df_i. The Inverse Document Frequency (IDF) equation is shown in equation (1) [9].

$$idf_i = \log\left(\frac{N}{df_i}\right) \tag{1}$$

The notation N is the total number of documents weighted. Furthermore, the $tf_{i,j}$ and idf_i values obtained are used to calculate the weight of words or terms in each document as shown in equation (2).

$$w_{i,j} = tf_{i,j} \times idf_i \tag{2}$$

2.2.2 Cosine Similarity

According to Kocher and Savoy [10] cosine similarity is a measure of the degree of similarity between two vectors and in the inner product, this measurement is most popularly used. In its use in this study, cosine similarity measurements are used to indicate the degree of similarity between two sentences. The measurement results produce a value between 0 and 1 where if the value is 0 then there is no similarity between the two sentences and vice versa. If the resulting value is 1, the sentence is identified as identical. Equation (3) shows the cosine similarity measurement to compare the two sentences K_1 and K_2. With K_{1i} and K_{2i} being vector components in sentences K_1 and K_2.

$$sim(K_1, K_2) = \frac{K_1 \cdot K_2}{||K_1|| ||K_2||} = \frac{\sum_{i=1}^{n} K_{1i} K_{2i}}{\sqrt{\sum_{i=1}^{n} (K_{1i})^2} \sqrt{\sum_{i=1}^{n} (K_{2i})^2}} \quad (3)$$

2.2.3 Evaluation

Several evaluation techniques are suitable for measuring the predicted score of a recommendation system without a data training process. The top-k-accuracy score is one of the evaluation models to generalize a prediction from the score's accuracy [11]. Prediction of information using Information Retrieval is generally a calculation of the many relevant and irrelevant items, which will lead to specific information recommendations. Evaluation with a top-k-accuracy score can be used in binary and multiclass classification cases. To get the top-k-accuracy score can be defined as shown in equation (4) if f_{ij} is the class of predictions for *i* samples to *j*, the highest score predictions, where *y* is the correct value.

$$\text{top} - k \text{ accuracy}(y, \hat{f}) = \frac{\sum_{i=0}^{n_{\text{samples}}-1} \sum_{j=1}^{k} 1(\hat{f}_{i,j} = y_i)}{n_{\text{samples}}} \quad (4)$$

2.2.4 Augmented Data

Data augmentation is carried out to overcome the problem of small datasets. Previous research applied text data augmentation to improve textual data. This approach creates new questions by using cosine similarity to find similar words and replace them in the same order. This method will produce various vocabularies in the dataset that will be created [12].

```
DEFINE FUNCTION rSubset(arr, r):

  RETURN list(combinations(arr, r))

SET df TO pd.DataFrame(columns TO ['Diagnosis', 'Gejala'])

FOR key IN dataset.keys():

  SET df TO df.append({'Diagnosis' : key, 'Gejala' : dataset[key]}, ignore_index TO True)

  IF len(dataset[key])>5:
      SET listcomb TO rSubset(dataset[key],len(dataset[key])-1)

      FOR j IN range(len(listcomb)):

        SET X TO list(listcomb[j])

        SET df TO df.append({'Diagnosis' : key, 'Gejala' : X}, ignore_index TO True)
  return go(f, seed, [])
}
```

Figure 1 Augmented Data

In this study, a new combination of symptoms was created for the test data by using disease diagnoses in Rekmed data that had more than five symptoms. So that an increase in test data is obtained by almost 3 times. The combination algorithm we apply to increase the dataset can be seen in Figure 1.

2.3 System Design

In this section, we analyze the steps needed on the prediction system. The system proposed in this study has the design as shown in Figure 2. Based on the design, the system implementation begins with the user entering symptoms. The system will carry out preprocessing stages for these symptoms, including case folding, tokenization, and stemming. The system performs query expansion with the synonym of entering symptoms. The synonyms for these symptoms will be adjusted to the symptoms contained in the database. The system provides related symptoms to the doctor so that the doctor can select the symptoms that relevant to the patient's symptoms to produce a list of final symptoms.

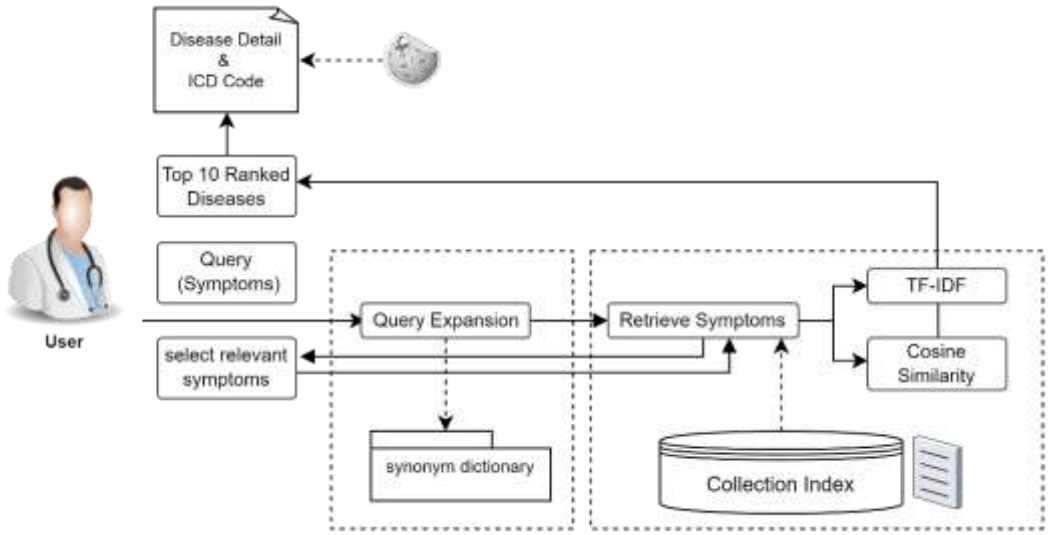

Figure 2 Architecture System Overview

The relevant symptoms are converted into vector form. Symptom vectors are used to measure the level of similarity with the data using two models, namely the TF-IDF weighting and cosine similarity. The measurement results will display a list of disease predictions whose symptoms are similar to the list of symptoms in the database. The system will display as many as k disease predictions with a percentage of similarity levels. Based on the diagnosis from the system, doctors can determine or validate the most suitable disease by entering symptoms. The system will display detailed disease information and the ICD-10 code of the disease for the details to support knowledge of the predicted disease.

3. RESULTS AND DISCUSSION

3.1 Implementation

The system made into components based on system design. The system is built into components so that every component can build independently in other words system can be built with certain techniques or methods without affecting the other system components. The design in Figure 2 is implemented in the Python programming language. The following subsections discuss each step of system in detail.

3.1.1 Query Processing

The user's query contains the disease's symptoms in single line, and each symptom is separated by a comma(,). Preprocessing query involved such as tokenization to get each symptom, case folding to convert the symptoms into lowercase, stopwords removal and stemming using Sastrawi.

3.1.2 Query Expansion

This process uses the input from the query preprocessing results. Each symptom will be expanded by using Indonesian synonyms taken from an Indonesian language json thesaurus (https://github.com/victoriasovereigne/tesaurus) to get list of matching symptoms. The list has been indexed so that users can select the symptoms according to the index. The example is shown in Figure 3.

```
Masukan gejala dipisahkan dengan koma(,):
'gatal', 'rasa mau mati', 'kulit kuning', 'mati rasa', 'penambahan berat badan',
'warna kekuningan kulit pada mata bagian putih'

Gejala pencocokan teratas dari pencarian Anda!
0 : titik atau spot merah pada mata putih
1 : gatal
2 : mati rasa
3 : penambahan berat badan
4 : kulit kuning
5 : warna kulit kebiruan
6 : kulit pucat
7 : warna pucat
8 : rasa mau mati
9 : belang kulit putih
10 : warna kekuningan kulit pada mata bagian putih

Silakan pilih gejala yang relevan. Masukkan indeks (dipisahkan dengan spasi):
1 2 3 4 8 10
Daftar akhir dari gejala yang diberikan untuk prediksi adalah :
gatal
mati rasa
penambahan berat badan
kulit kuning
rasa mau mati
warna kekuningan kulit pada mata bagian putih
```

Figure 3 Implementation of entry selection symptoms

3.1.3 Symptoms Selection

In addition to expanding the matched symptom query, the system will display symptoms that may be related to the symptoms entered by the user. The user can select one or more symptoms from the list of matching symptoms, continue to next list of other matching symptoms if any, or stop to add more matching symptoms. Figure 3 is an example of the implementation of entering symptoms, a combination of expanding symptoms that the system has processed, and selecting matching symptoms.

3.1.4 Disease Prediction

At this stage, the TF-IDF and cosine similarity as retrieval models are applied to return output that matches the given symptoms.

a) Prediction Model with TF-IDF and Cosine Similarity
   The diagnosis is obtained from the disease prediction that matches the symptoms given. The model used to predict TF-IDF and cosine similarity. Based on Figure 4 and Figure

5, the two models developed will display the top 10 predictive diagnoses similar to the symptoms in the dataset. The TF-IDF model tends to have a higher predictive score than the cosine similarity model

b) Disease Details

In this section, the user can select one of the diseases from the prediction results to find more detailed information from the Infobox Wikipedia and ICD-10 codes of selected diseases.

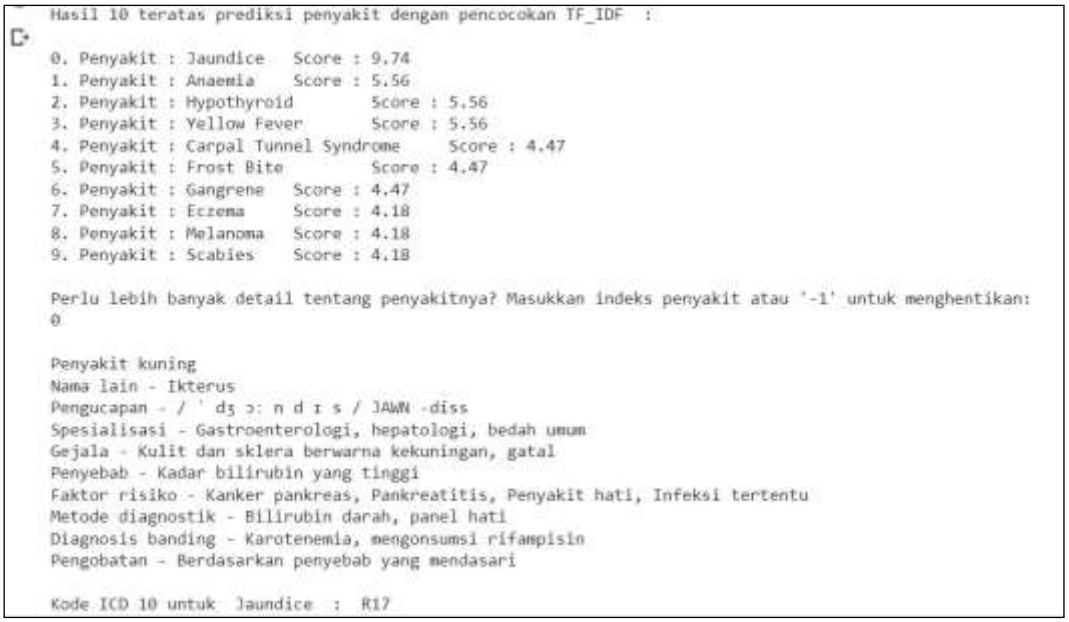

Figure 4 The results of the diagnosis using the TF-IDF model

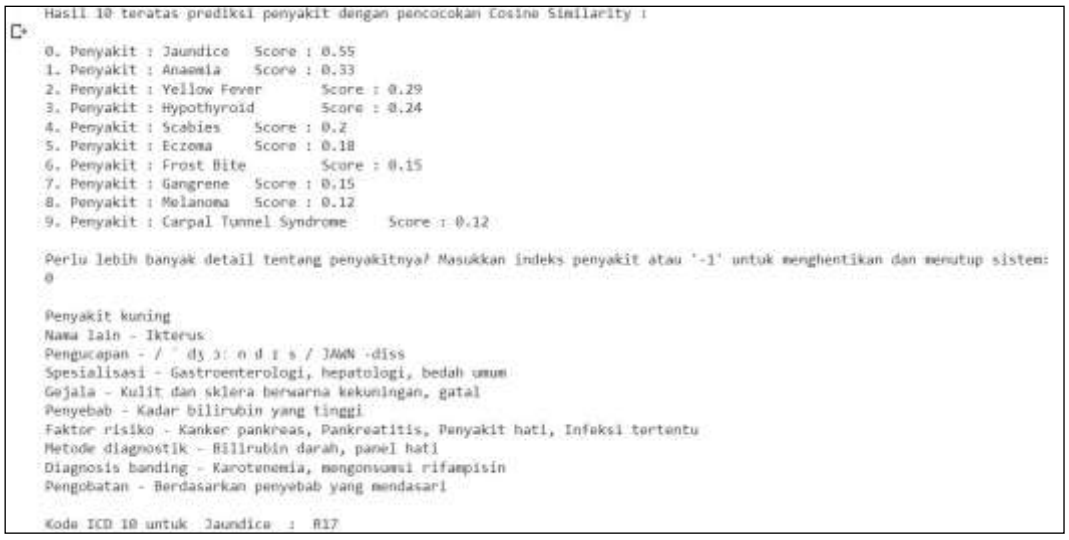

Figure 5 The results of the diagnosis using cosine similarity model

3.2 Evaluation

The prediction system is evaluated based on the accuracy of the disease prediction results compared to the actual diagnosis. This evaluation process determines whether the system runs according to the analysis results and the expected goals. System prediction accuracy is evaluated using test data in the data Rekmed. The result is shown in Table 1.

Table 1 Comparison Evaluation of TF-IDF and Cosine Similarity

k	TF-IDF (105 Data)	Cosine Similarity (105 Data)	TF-IDF (295 Data)	Cosine Similarity (295 Data)
1	0.057143	0.009524	0.061017	0.00339
2	0.085714	0.019048	0.091525	0.00678
3	0.12381	0.028571	0.128814	0.010169
4	0.12381	0.038095	0.142373	0.013559
5	0.152381	0.047619	0.189831	0.040678
6	0.161905	0.057143	0.19661	0.044068
7	0.161905	0.066667	0.19661	0.088136
8	0.161905	0.07619	0.19661	0.111864
9	0.161905	0.085714	0.20339	0.155932
10	0.161905	0.095238	0.20339	0.159322

Based on the results of system evaluation using the top-k-accuracy score, as shown in Table 1, is obtained by comparing the accuracy of the two retrieval methods on the both of test data we used in this research, then the results in the table were visualized in graphical form as shown in Figure 6. In TF-IDF, the accuracy of disease prediction with k predictions with the highest value, namely at $k = 10$, is only 16%. whereas after the augmentation of the test data, the prediction accuracy of $k = 10$ increases. With an increase in $k = 1$ around 1% and $k = 10$ about 4% to 20%. The accuracy of the cosine similarity method is lower than that of the TF-IDF method, with the highest prediction rate at $k = 10$, namely 9.5%. After augmented test data, the accuracy decreases for $k = 1$ to $k = 6$. The accuracy value increases from $k = 7$ to $k = 10$, with the highest accuracy, close to 16%. The increase in accuracy results on the augmented test data is analyzed that entering relevant symptoms will undoubtedly increase the accuracy of disease prediction. In the test data that we use in this study, although the data has been augmented to be almost three times greater than the previous test data, the increase in accuracy results is a manageable size.

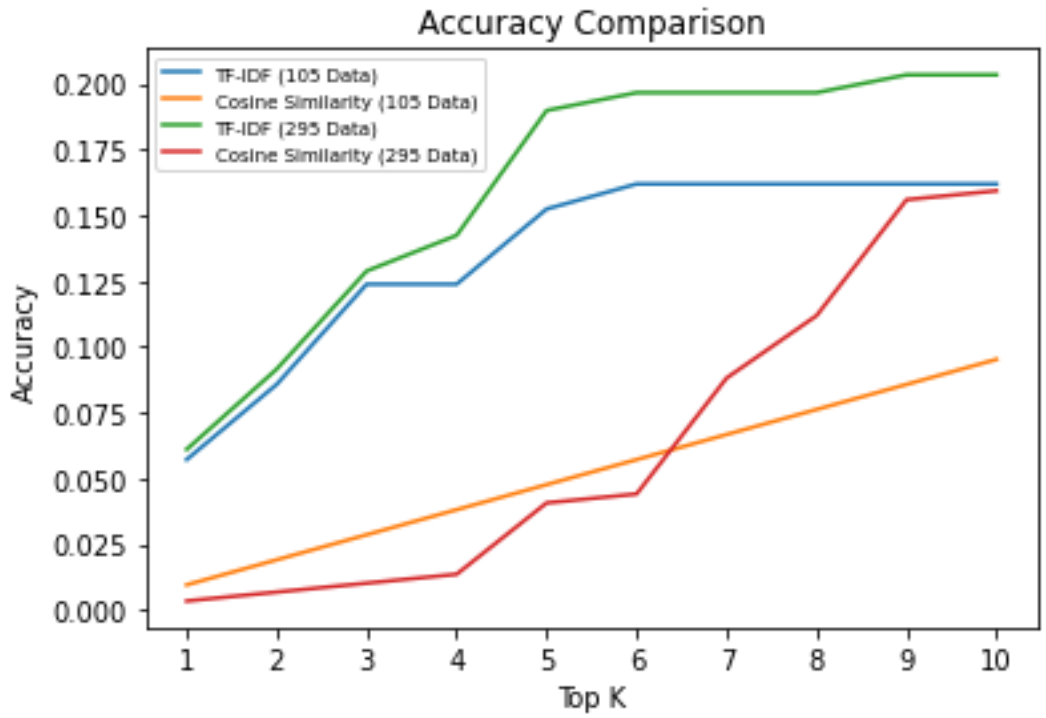

Figure 6 Comparison accuracy using top-k-accuracy score

The highest accuracy score is the prediction using TF-IDF with 295 test data. Compared based on the method, TF-IDF always gives more accurate prediction results than the results with cosine similarity, while comparisons based on test data on each method of augmented test data are more accurate. Even though there is an increase in the accuracy value of the augmented test data, the increase is not too significant. This is analyzed because the test data is augmented by combining the symptoms in the diagnosis data which has quite a large list of symptoms, while the data should be augmentable by expanding it by adding symptoms that are relevant to diagnosis so that better augmentation results can be obtained.

4. CONCLUSIONS

It is considered that including symptoms that are irrelevant to the disease will not affect the prediction results. From this study, test data that had been augmented generally resulted in better prediction accuracy. The level of accuracy of prediction results using the TF-IDF method is greater than the prediction using the cosine similarity method. For further research, a different augmentation method can be used, or the addition of disease diagnosis data with more relevant symptoms to provide more accurate disease predictions.

REFERENCES

[1] M. Mustakim and R. Wardoyo, "Survey Model-Model Pencarian Informasi Rekam," *JISKA J. Inform. Sunan Kalijaga*, vol. 3, no. 3, pp. 132–144, 2019, [Online]. Available: https://doi.org/10.14421/jiska.2019.33-01

[2] R. Silalahi and E. J. Sinaga, "Perencanaan Implementasi Rekam Medis Elektronik Dalam Pengelolaan Unit Rekam Medis Klinik Pratama Romana," *J. Manaj. Inf. Kesehat. Indones.*, vol. 7, no. 1, p. 22, 2019, doi: 10.33560/jmiki.v7i1.219.

[3] C. of Australia, "MBS Telehealth Services from 1 July 2022," 2022. http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/Factsheet-telehealth-1July22 (accessed Oct. 25, 2022).

[4] V. K and J. Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks," *Int. J. Comput. Appl.*, vol. 19, no. 6, pp. 6–12, 2011, doi: 10.5120/2368-3115.

[5] A. M. Nuraini Ahmad, Arienda Addis Prasetyo, "Penerapan Information Retrieval Pada Search Engine," *J. Inov. Has. Penelit. dan Pengemb.*, vol. 1, no. 31, pp. 15–23, 2021, [Online]. Available: https://jurnalp4i.com/index.php/knowledge/article/view/771

[6] M. Yusuf and A. Cherid, "Implementasi Algoritma Cosine Similarity Dan Metode TF-IDF Berbasis PHP Untuk Menghasilkan Rekomendasi Seminar," *J. Ilm. Fak. Ilmu Komput.*, vol. 9, no. 1, pp. 8–16, 2020, [Online]. Available: https://publikasi.mercubuana.ac.id/index.php/fasilkom/article/view/8830

[7] Rahul Maheshwari, "Disease Detection based on Symptoms with treatment recommendation." https://rahul-maheshmaheshwari.medium.com/disease-detection-based-on-symptoms-with-treatment-recommendation-with-scrapped-data-set-54e6be60a3d1 (accessed Oct. 25, 2022).

[8] Christopher D. Manning, Prabhakar Raghavan and H. Schütze, "Introduction to Modern Information Retrieval (2nd edition)," *Libr. Rev.*, vol. 53, no. 9, pp. 462–463, 2004, doi: 10.1108/00242530410565256.

[9] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," *Proc. 2016 4th Int. Conf. Cyber IT Serv. Manag. CITSM 2016*, 2016, doi: 10.1109/CITSM.2016.7577578.

[10] K. Park, J. S. Hong, and W. Kim, "A Methodology Combining Cosine Similarity with Classifier for Text Classification," *Appl. Artif. Intell.*, vol. 34, no. 5, pp. 396–411, 2020,

doi: 10.1080/08839514.2020.1723868.
[11] scikit-learn developer, "Metrics and scoring: quantifying the quality of predictions," 2022. https://scikit-learn.org/stable/modules/model_evaluation.html (accessed Nov. 29, 2022).
[12] T. Phreeraphattanakarn and B. Kijsirikul, "Text data-augmentation using Text Similarity with Manhattan Siamese long short-term memory for Thai language," *J. Phys. Conf. Ser.*, vol. 1780, no. 1, 2021, doi: 10.1088/1742-6596/1780/1/012018.