Published online on: **https://journal.ilmudata.co.id/index.php/RIGGS**

Journal of Artificial Intelligence and Digital Business
(RIGGS)

Vol. 1 No. 1 (2022) 13 - 18 e-ISSN: 2963-914X

Dice Similarity and TF-IDF for New Student Admissions Chatbot

Muhammad Riko Anshori Prasetya[1], Arif Mudi Priyatno[2]
[1]Information System, Sains and Technology, Sari Mulia University
[2]Department Digital Bussiness, Faculty of Economics and Business, Pahlawan Tuanku Tambusai University
[1]riko.anshori@gmail.com, [2]arifmudi@universitaspahlawan.ac.id

Abstract

CS is one of the most important functions of any client-related organization, whether a business or a school (customer service). Notably from the committee responsible for student selection, CS, on the other hand, has a very limited capacity to be handled by humans, which can reduce university satisfaction. Therefore, we require technological assistance, which in this case takes the form of an AI-based chatbot. The objective of this study is to design and develop a chatbot system utilizing NLP (natural language processing) to aid the CS of the new student admissions committee at Pahlawan Tuanku Tambusai University in answering questions from prospective new students. The employed method is dice similarity weighted by TFIDF. The results of the conducted tests indicated that the recall rate was 100 percent and the precision reached 76.92 percent. The evaluation results indicate that the chatbot can effectively respond to questions from prospective students.

Keywords: costumer service, NLP, chatbot, dice similarity, TF-IDF

1. Introduction

As a result of the rapid advancement of technology and information in this age of globalization, numerous fields of labor, including education, are required to adapt to the changes brought about by these developments. Technology is being used by an increasing number of educational institutions across the globe in order to enhance the standard of the services they provide, with the true objective of drawing in more students. In addition to the quality of the education that is offered, a university needs to make it easier for prospective students to enroll at the college by providing information services that are both prompt and accurate. This is necessary in order to win the satisfaction of prospective students [1].

There are a multitude of precautions in place to expedite the delivery of information to prospective new students, including telephone service and live chat support from customer service. The next step in disseminating this information is crucial so that the message reaches prospective students as soon as possible, allowing them to immediately prepare everything needed to enroll in college. This service requires customer service to be capable of responding to questions from prospective students within 24 hours[2]. However, the capacity of customer service itself is also limited, and representatives of the department require sufficient time to respond to inquiries from prospective students [1]. In addition, an increase in prospective students during the admissions process will almost certainly lead to an increase in the number of questions asked, as well as an increase in the amount of time spent answering those questions; this, in turn, may lead to a decrease in overall satisfaction with the university[3].

The use of an artificial intelligence program known as a chatbot is one of the steps that can be taken to make it simpler for a customer service representative to respond to these questions. The Chatbot itself is a piece of software that utilizes Natural Language Processing (NLP), which is a subset of Artificial Intelligence (AI). The model for the Chatbot was derived from Human Computer Interaction (HCI), which allows computers to communicate with humans through text [4]. Because questions will be answered by the chatbot around the clock, prospective students who are looking for information will have a much simpler time finding what they need thanks to the creation of this chatbot [3].

For this study purpose, a chatbot was developed that can respond to messages from prospective new students using human everyday language. In this study, a chatbot will be developed that will act as an agent to assist a customer service assistant in answering questions from

Accepted: 31-07-2022 | Received in revised: 31-07-2022 | Published: 31-07-2022

prospective students 24 hours a day, seven days a week. The chatbot will respond to messages from prospective students using NLP, which will include preprocessing and a weighting process called TFIDF to generate a value that will be used in the dice similarity process. Dice Similarity is useful so that the chatbot can find answers from user input by comparing the answers with all documents stored in the knowledge base, so that the chatbot application itself can assist customer service in automatically and flexibly serving prospective students.

2. Research Methods

Figure 1 displays the multiple plots utilized in this study.

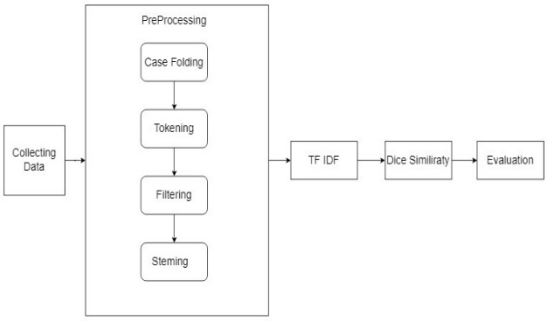

Figure 1. Resarch Step

2.1. Collecting Data

For the purpose of this study, the data collected are comprised of information taken from a number of questions and responses that have been sent to customer service by Pahlawan Tuanku Tambusai University since 2017. There are a total of 42 questions that were compiled after being asked repeatedly by prospective students to the Computer Science program at Pahlawan Tuanku Tambusai University. The questions that prospective students have asked and the responses that they have received are presented in Table 1.

Table 1. Prospective Students' Questions

Question	Answer
kak apa akreditasi program studi S1 Gizi ?	akreditasi program studi S1 Gizi yaitu B kak
kak apa akreditasi program studi S1 KESEHATAN MASYARAKAT ?	akreditasi program studi S1 KESEHATAN MASYARAKAT yaitu B
kak apa akreditasi program studi Diploma III Kebidanan ?	akreditasi program studi Diploma III Kebidanan yaitu B kak
kak apa akreditasi program studi Diploma III Keperawatan ?	akreditasi program studi Diploma III Keperawatan yaitu B kak

2.2. Preprocessing

During the text mining process, raw data is meaningless and of no use. It is necessary to process the raw data before it can be read by a computer [5]. Preprocessing is the name given to the method that is used to process the raw data itself [6]. The preprocessing stage of this research makes use of the following four processes:

1. Case Folding

The process of changing all of the uppercase letters and symbols in the message into lowercase letters and symbols is referred to as *case folding* [7]. Table 2 displays an example of the use of case folding that was performed for the purpose of this research.

Table 2. Case Folding Process

Pre Process	After Process
kak apa akreditasi program studi S1 Gizi ?	kak apa akreditasi program studi s1 gizi

2. Tokening

The process of separating a piece of text into its component sentences is known as tokening [5]. Table 3 illustrates an application of tokening that was performed for the purpose of this research.

Table 3. Tokening Process

Pre Process	After Process
kak apa akreditasi program studi s1 gizi	‘kak’, ‘apa’, ‘akreditasi’, ‘program’, ‘studi’, ‘s1’, ‘gizi’

3. Filtering (Stopword Removal)

A filtering procedure is used to eliminate words that were deemed unimportant during the preceding procedure[6]. Table 4 contains examples of the use of filtering in this study.

Table 4. Tokening

Pre Process	After Process
‘kak’, ‘apa’, ‘akreditasi’, ‘program’, ‘studi’, ‘s1’, ‘gizi’	‘akreditasi’, ‘program’, ‘studi’, ‘s1’, ‘gizi’

4. Stemming

Stemming is the final step of the preprocessing that is being done for this research. The act of "stemming" in and of itself refers to the process of removing an affix from the text. The use of stemming in this study can be seen in Table 5.

Table 5. Stemming

Pre Process	After Process
‘akreditasi’, ‘program’, ‘studi’, ‘s1’, ‘gizi’	‘akreditasi’, ‘program’, ‘studi’, ‘s1’, ‘gizi’

DOI: https://doi.org/10.31004/riggs.v1i1.5

2.3. TF IDF

TF-IDF (Term Frequency-Inverse Document) is a step in document weighting that will be used to extract information from a document. This algorithm is one that is frequently used to convert text into a meaningful value [8].. The TF-IDF equation is shown in the following equation (1).

$$\sum_{i \in_d} tf_{i,d} * \log(\frac{N}{df_i}) \quad (1)$$

$tf_{i,d}$ is a representation of the total number of occurrences of the i-th term in the document, df_i is a representation of the entirety of the document that contains the i-th term, and N is a representation of the total number of documents.

2.4. Dice Similarity

One of the ways that the degree of similarity between two things can be determined is through the use of the dice similarity method. The value of k-grams is calculated for documents that are compared with dice similarity in order to determine their level of similarity. The query and the document were measured against one another, and the returned document is the document that was obtained from that measurement [6]. The equation that describes the similarity between dice is shown down below in equation (2).

$$S = \frac{2|A \cap B|}{|A|+|B|} * 100 \quad (2)$$

Where s is the total similarity value, A and B are each document's inputs.

2.5. Evaluation

Evaluation is an essential part of this research project, as it allows the researchers to check whether or not the developed bot system is functioning appropriately and in line with their aims. In the course of this research, the test scenario was developed by posing 42 questions to the chatbot. Recall and precision methods are utilized in the process of self-measurement. The level of success that a system has in finding information is referred to as its recall, and the matching of a piece of data with the necessary information is referred to as its precision [1] [9] [10] [11]. In equation 3, you can see the equation for recall, and in equation 4, you can see the equation for precision.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Precission = \frac{TP}{TP+FP} \quad (4)$$

In this study, the term *True Positive* (TP) refers to the number of correct answers provided by the chatbot, *False Negative* (FN) is an answer that the chatbot is unable to answer, and *False Positive* (FP) is the answer provided by the chatbot, but the results do not match with what you want.

3. Results and Discussions

This research makes use of Python version 3.9 as its programming language. Additionally, the research makes use of the Telegram application and BotFather as a supplier of bots for use in Telegram. Python functions are used for preprocessing, TF-IDF weighting, dice similarity, and for bots to respond to messages. These functions also allow for dice similarity. Calculating the TF-IDF requires the following coding, which is as follows:

```
1.  def tfidf(query, data, stopWords, N):
2.      words_list = []
3.      for doc in data:
4.          doc = re.sub('[^A-Za-z0-9]+',' ', doc)
5.          words_list.append([word.lower() for word in nltk.word_tokenize(doc) if word not in stopWords])
6.      all_words = list(itertools.chain(*words_list))
7.      word_set = list(set(all_words))
8.
9.      tf_vecs = [0 for i in range(N)]
10.     for i in range(N):
11.         tf_vecs[i] = [words_list[i].count(w) for w in word_set]
12.
13.     idf_all_words = list(itertools.chain(*[set(doc_words) for doc_words in words_list]))
14.     idfs = [math.log(float(N) / idf_all_words.count(w), N) for w in word_set]
15.     tfidf = [0 for i in range(N)]
16.     nomDc=[];
17.     for i in range(N):
18.         tfidf[i] = [tf * idf for tf, idf in zip(tf_vecs[i], idfs)]
19.         nomD = math.sqrt(sum(x**2 for x in tfidf[i]))
20.         tfidf[i] = [x / nomD for x in tfidf[i]]
```

DOI: https://doi.org/10.31004/riggs.v1i1.5

```
    nomDc.append(tfidf[i])
  query = re.sub('[^A-Za-z0-9]+',' ', query)
  qwords = [word.lower() for word in query.split() if word not in stopWords]
        qvec = [qwords.count(w) for w in word_set]
        qvec = [tf * idf for tf, idf in zip(qvec, idfs)]
  nomQ = math.sqrt(sum(x**2 for x in qvec))

  if nomQ != 0.0:
    qvec = [x / nomQ for x in qvec]
  else:
    qvec = [0 for x in qvec]
        return qvec, nomDc
```

The coding to do the dice comparison comes next.

```
def dicesimiliarity(nDoc,nQue):
  result = []
  for x in range(len(nDoc)):
    dot = sum(a * b for a, b in zip(nDoc[x], nQue))
    norm_a = sum(a**2 for a in nDoc[x])
    norm_b = sum(b**2 for b in nQue)
    count = (2 * dot) / (norm_a + norm_b)
    result.append(count)
  return result
```

After that, write the code that will allow the bot to receive the message:

```
def answer(dice,data,filenames):
  docK = []
  stop = open('univpahlawan/lainnya.txt').read()
  sort = sorted(dice, reverse=True)
  print("-----------------------------------------------------------------")
  print("BOBOT tertinggi adalah :", sort[0])
  print("-----------------------------------------------------------------")
  if sort[0] <= 0.20:
    print('kelasnya lainnya')
    return stop
  for x in sort[:1]:
    index = dice.index(x)
    docK.append(data[index])
    print(data[index])
    print(x)
    splitname = filenames[index].split("\\")
    print('kelasnya', splitname[-1])
    filejawaban = 'jawaban/'+splitname[-1]
    jawab = open(filejawaban, "r")
  doc_ans=" ".join(jawab)
  return doc_ans
```

When everything is done, each of these functions will be saved in one file. Following the execution of the file, the bot that had been previously created and given the name @penerimaan_pahlawan will immediately reply to the message automatically

3.1. Weighting Results with TF-IDF and Dice Similarity

After the text provided by the prospective student has been preprocessed, the program will now process the sentence that is the most comparable to the one in the query. In this instance, the sentence will be the data that was previously entered. The actual test was administered 42 times, and in the table that was provided as an example, there were 15 different possible responses. Table 6 presents the findings of the examinations that were carried out, including the results of the measurement of the text's degree of similarity as well as the weight that was assigned to the calculation of dice similarity.

Table 6. Weighting of Dice Similarity

No	Question	Bot Answer	Weight
1	akreditasi program studi S1 Gizi yaitu B kak	akreditasi program studi S1 Gizi yaitu B kak	0.9318
2	akreditasi program studi S1 KESEHATAN MASYARAKAT yaitu B	akreditasi program studi S1 KESEHATAN MASYARAKAT yaitu B	0.9572
3	akreditasi program studi Diploma III Kebidanan yaitu B kak	akreditasi program studi Diploma III Kebidanan yaitu B kak	0.4641
4	kak apa akreditasi program studi DIV Kebidanan ?	akreditasi program studi DIV Kebidanan yaitu B kak	0.3334
5	kak apa akreditasi program studi S1	akreditasi program studi S1 KEPERAW	0.8992

DOI: https://doi.org/10.31004/riggs.v1i1.5

	KEPERAWATAN ?	ATAN yaitu B kak	
6	kak apa akreditasi program studi S1 Kewirausahaan ?	akreditasi program studi S1 Kewirausahaan yaitu Baik kak	0.8688
7	jawaban jika tidak di mengerti	Mohon Tanya Lebih Spesifik agar Hero Chatbot Mengerti. Terima Kasih	
8	kak apa akreditasi program studi S1 Pendidikan Matematika ?	akreditasi program studi S1 Pendidikan Matematika yaitu B kak	0.9229
9	kak apa akreditasi program studi S1 PENDIDIKAN JASMANI KESEHATAN DAN REKREASI (PENJASKESREK) ?	akreditasi program studi S1 PENDIDIKAN JASMANI KESEHATAN DAN REKREASI (PENJASKESREK) yaitu C kak	0.7331
10	kak apa akreditasi program studi S1 Peternakan ?	akreditasi program studi S1 Peternakan yaitu Baik kak	0.8688
11	kak apa akreditasi program studi S1 Pendidikan Guru Sekolah Dasar (PGSD) ?	akreditasi program studi S1 Pendidikan Guru Sekolah Dasar (PGSD) yaitu B kak	0.5814
12	kak apa itu JALUR TAHFIZ ?	Jalur Tahfiz merupakan jalur yang dikhususkan bagi siswa yang mempunyai hafalan Al-Quran. * Hafal 30 Juzz Potongan Uang Pembangunan 100%. * Hafal 25 Juzz Potongan Uang Pembangunan 75%. * Hafal 15 Juzz Potongan Uang Pembangunan 25%	0.3133
13	kak apa akreditasi program studi S1 Teknik Industri ?	akreditasi program studi S1 Teknik Industri yaitu C kak	0.9243
14	kak apa akreditasi program studi S1 Teknik Sipil ?	akreditasi program studi S1 Teknik Sipil yaitu C kak	0.9243
15	Kak kapan tes masuk gelombang 1 (Pertama / Satu) di UP ?	Tes Masuk UP gelombang 1 : 27 Juni 2022	0.6289

According to table 6, the cutoff point that is applied to each question is 0.10. If the threshold's weight is less than 0.10, then the response from the bot is 'Mohon Tanya Lebih Spesifik agar Hero Chatbot Mengerti. Terima Kasih'.

3.2. Evaluasi

The evaluation is based on the chatbot's responses. The analysis of chatbot responses is outlined in the *Confusion Matrix*, with examples of chatbot responses in table 7 of the confusion matrix.

Table 7. Confusion Matrix

Question	Bot Answer	TP	FP	FN	TN
akreditasi program studi S1 Gizi yaitu B kak	akreditasi program studi S1 Gizi yaitu B kak	1			
akreditasi program studi S1 KESEHATAN MASYARAKAT yaitu B	akreditasi program studi S1 KESEHATAN MASYARAKAT yaitu B	1			
akreditasi program studi Diploma III Kebidanan yaitu B kak	akreditasi program studi Diploma III Kebidanan yaitu B kak		1		
kak apa akreditasi program studi DIV Kebidanan ?	akreditasi program studi DIV Kebidanan yaitu B kak		1		
kak apa akreditasi program studi S1 KEPERAWATAN ?	akreditasi program studi S1 KEPERAWATAN yaitu B kak		1		
kak apa akreditasi program studi S1 Kewirausahaan ?	akreditasi program studi S1 Kewirausahaan yaitu Baik kak	1			
jawaban jika tidak di mengerti	Mohon Tanya Lebih Spesifik agar Hero Chatbot Mengerti. Terima Kasih	1			
kak apa akreditasi program studi S1 Pendidikan Matematika ?	akreditasi program studi S1 Pendidikan Matematika yaitu B kak	1			
kak apa akreditasi program studi S1 PENDIDIKAN	akreditasi program studi S1 PENDIDIKAN	1			

DOI: https://doi.org/10.31004/riggs.v1i1.5

JASMANI KESEHATAN DAN REKREASI (PENJASKESREK) ?	JASMANI KESEHATAN DAN REKREASI (PENJASKESREK) yaitu C kak		
kak apa akreditasi program studi S1 Peternakan ?	akreditasi program studi S1 Peternakan yaitu Baik kak	1	
kak apa akreditasi program studi S1 Pendidikan Guru Sekolah Dasar (PGSD) ?	akreditasi program studi S1 Pendidikan Guru Sekolah Dasar (PGSD) yaitu B kak	1	
	Jalur Tahfiz merupakan jalur yang dikhususkan bagi siswa yang mempunyai hafalan Al-Quran.		1
kak apa itu JALUR TAHFIZ ?	* Hafal 30 Juzz Potongan Uang Pembangunan 100%. * Hafal 25 Juzz Potongan Uang Pembangunan 75%. * Hafal 15 Juzz Potongan Uang Pembangunan 25%		
kak apa akreditasi program studi S1 Teknik Industri ?	akreditasi program studi S1 Teknik Industri yaitu C kak	1	
kak apa akreditasi program studi S1 Teknik Sipil ?	akreditasi program studi S1 Teknik Sipil yaitu C kak	1	
Kak kapan tes masuk gelombang 1 (Pertama / Satu) di UP ?	Tes Masuk UP gelombang 1 : 27 Juni 2022	1	

The Confusion Matrix method was utilized in order to carry out the evaluation, and the parameters that were looked at were precision and recall. Recall rates of one hundred percent and precision rates of 76.92 percent were found to be the most impressive aspects of the chatbot's performance in the evaluation based on the results of several questions asked of it.

4. Conclusion

For the purpose of this investigation, a chatbot application was developed with the assistance of Telegram for the procedure of Pahlawan Tuanku Tambusai University's admission of new students.

The conclusion that can be drawn from the outcomes of the tests that were conducted using the chatbot and the responses that were obtained from it is that the recall rate reaches 100% while the precision level is approximately 76.92%. In light of these findings, the development of a chatbot as part of the procedure for admitting new students to Pahlawan Tuanku Tambusai University can be utilized to provide responses to questions posed by prospective new students.

Reference

[1] S. Adam and E. Lulianthy, "Frequenly Ask Question (FAQ) Chatbot for New Student Admission System Using Natural Language Processing at Politeknik Aisyiyah Pontianak," *Jtksi*, vol. 04, no. 03, 2021.

[2] D. S. Hormansyah and Y. P. Utama, "Aplikasi Chatbot Berbasis Web Pada Sistem Informasi Layanan Publik Kesehatan Di Malang Dengan Menggunakan Metode Tf-Idf," *J. Inform. Polinema*, vol. 4, no. 3, p. 224, 2018, doi: 10.33795/jip.v4i3.211.

[3] H. Agus Santoso *et al.*, "Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services," *Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018*, pp. 417–423, 2018, doi: 10.1109/ISEMANTIC.2018.8549797.

[4] E. Elshan, N. Zierau, C. Engel, A. Janson, and J. M. Leimeister, *Understanding the Design Elements Affecting User Acceptance of Intelligent Agents: Past, Present and Future*, no. 0123456789. Springer US, 2022. doi: 10.1007/s10796-021-10230-9.

[5] T. wahyuningsih, "Text Mining an Automatic Short Answer Grading (ASAG), Comparison of Three Methods of Cosine Similarity, Jaccard Similarity and Dice's Coefficient," *J. Appl. Data Sci.*, vol. 2, no. 2, pp. 45–54, 2021, doi: 10.47738/jads.v2i2.31.

[6] S. Purwaningrum, A. Susanto, and ..., "Comparation of Dice Similarity and Jaccard Coefficience Against Winnowing Algorithm For Similarity Detection of Indonesian Text Documents," *J. Appl. …*, vol. 6, no. 1, pp. 10–22, 2021.

[7] A. F. Shobirin, D. Puspitasari, and A. Prasetyo, "Aplikasi Chatbot untuk Reservasi Pijat Bayi dengan Metode Cosine Similarity," in *Seminar Informatika Aplikatif Polinema*, 2020, pp. 150–156.

[8] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability," *Int. Conf. Artif. Intell. Transform. Bus. Soc. AITB 2019*, vol. 1, pp. 1–4, 2019, doi: 10.1109/AITB48515.2019.8947433.

[9] A. M. Priyatno, F. M. Putra, P. Cholidhazia, and L. Ningsih, "Combination of extraction features based on texture and colour feature for beef and pork classification," *J. Phys. Conf. Ser.*, vol. 1563, no. 1, p. 012007, Jun. 2020, doi: 10.1088/1742-6596/1563/1/012007.

[10] A. M. Priyatno, "Spammer Detection Based on Account, Tweet, and Community Activity on Twitter," *J. Ilmu Komput. dan Inf.*, vol. 13, no. 2, pp. 97–107, Jul. 2020, doi: 10.21609/jiki.v13i2.871.

[11] A. M. Priyatno, M. M. Muttaqi, F. Syuhada, and A. Z. Arifin, "Deteksi bot spammer twitter berbasis time interval entropy dan global vectors for word representations tweet's hashtag," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 5, no. 1, pp. 37–46, Jan. 2019, doi: 10.26594/register.v5i1.1382.

DOI: https://doi.org/10.31004/riggs.v1i1.5