

## IMPLEMENTATION OF VIRTUAL CONVERSATION WITH THE COSINE SIMILARITY METHOD IN TOURISM SERVICE APPLICATIONS IN EAST KALIMANTAN

Nikolaus Rio Saputra<sup>1</sup> and Viny C Mawardi<sup>2\*</sup>

<sup>1,2</sup> Faculty of Information Technology, Universitas Tarumanagara, Jakarta, Indonesia

Email : Nikolaus.535219101@stu.untar.ac.id,

Email : vinyam@fti.untar.ac.id

Submitted: 27-09-2023, Revised: 27-10-2023, Accepted: 11-12-2023

---

### ABSTRACT

*Chatbot is an application designed to communicate with machines. This communication helps users in searching for information. The information provided varies, such as information media regarding Tourism Public Services in the East Kalimantan Region. Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) that focuses on natural language processing. Natural language is language that is generally used by humans to communicate with each other. With current artificial intelligence technology, natural language can be processed into various forms, such as chatbots using methods, one of which is using Tf-Idf and Cosine Similarity. Cosine Similarity is a method for the Question Answering system by calculating the weight of each word in a question which will then be matched to the dataset. The use of this method will be applied to a chatbot system which is intended as an information medium regarding Tourism Public Services in the East Kalimantan Region, as a substitute for customer service, in addition to changing the delivery of information so that it is easy to understand. Based on tests results overall, provides a conclusion that the application built can be said to have met the requirements functionally, even though there are some cases that have not been accepted, this is caused by data that has not been used at this time, this limitation will fade as data is added by the application development team.*

### 1. Introduction

The rapid development of information systems in technology has had an impact on human life. Technological progress is something that cannot be avoided, because technology develops in accordance with scientific progress. With technology, it is hoped that news and information can be delivered and obtained much more quickly and easily. Public services are a medium provided by the government to provide the latest information to the public.

Information technology has now spread to various fields, one of which is tourist attraction information. In the field of tourism, every interesting thing to see can be said to be a tourist attraction. In Indonesia, especially in the province of East Kalimantan, it is one of the tourist areas that can be visited. This province has quite interesting potential compared to other places because it has 90% natural tourist attractions and the remaining 10% are artificial tourist attractions.



Figure 1 East Kalimantan Tourism

As the capital of East Kalimantan province, the city of East Kalimantan has quite a lot of different tourist and cultural potential. Various East Kalimantan tourist attraction information system websites are available and can be accessed to obtain information. Unfortunately, the existing information usually feels inadequate because of the large amount of information that is absorbed and can confuse users. To overcome this, a system is needed that can create clear and concise information for its users [1].

Based on this, a system was created to create clear information through a tourist information question and answer system. The implementation of this system is intended to help and make it easier for users to obtain information related to East Kalimantan tourism. The question and answer system created will be used to replace customer service in serving users who want to ask questions and answers about tourism. To create this system, a chatbot application system was created.

Chatbot itself is a program designed with the aim of simulating good communication with users. Chatbot is an artificial intelligence system or Artificial Intelligence (AI) which is produced from natural language processing or Natural Language Processing (NLP) which studies communication between computers and humans [2].

Communication that occurs in chatbots is a conversation using written media with responses from programs that have been declared in the program database on the computer. The ability of computers to store large amounts of data without forgetting the stored information combined with the practicality of asking direct information sources compared to searching for information themselves and their learning abilities makes chatbots reliable customer service.

Based on the description above, in the final assignment of this thesis a virtual conversation system in the form of a chatbot will be created for tourism services in the East Kalimantan area. The system will play a role in serving and responding to every question from the user through user input regarding East Kalimantan tourist attractions. The methods used to create this chatbot system are the Term Frequency-Inverse Document Frequency (TF-IDF) method and the Cosine Similarity method.

Term Frequency-Inverse Document Frequency (TF-IDF) is a way of giving weight to the relationship of a word (term) in a document. In the scheme, there is a process related to word weighting locally and globally. Local weighting is only guided by the frequency of words appearing in documents, and cannot look at the frequency of words appearing in other documents. Term frequency is the local approach that is most widely applied compared to other approaches [3].

Cosine Similarity is a Vector Space Model technique for measuring the similarity between a query and a document. Queries and documents are considered as vectors in  $n$ -dimensional space where  $n$  is the number of all words in the lexicon. The lexicon is a list of all the words in the index. One way to overcome this in the vector space model is to expand the vector with a process carried out on the query vector, document vector, or both vectors.

## 2. Material and Methods

### 2.1. Related Works

*"Implementation of the Levenshtein Distance Algorithm as a Line Application-Based Tourism Agent Chatbot"* by Fahri Firdausillah, Arieansyah [4]. The system designed is in the form of an online chatbot that helps automate the process for potential tourists to carry out activities easily at a place. The algorithm used is the Levenshtein distance algorithm which gets a score of more than 30%, namely 40% of the total investigators who tested the chatbot.

*"Application of the Scaffolding Rental Information System Chatbot Using the TF-IDF Method"* by Dimas Wahyu Wibowo, Habibie Ed Dien, Trianta Almira Ramadhani [5]. Chatbot provides fast information services for CVs. Scaffolding East Kalimantan by separating the constituent words of a document (tokenizing) and calculating the TF-IDF and Cosine Similarity weights to search for answers in the system.

*"WEB-Based Chatbot Application Using the DIALOGFLOW Method"* by Dicki Wahyudi Harahap, Liza Fitria [6]. The chatbot was designed in the form of a helpdesk at the Binjai Pratama Tax office to monitor the public regarding taxes according to the data entered. The mini system was designed with dialog flow consisting of agent intent and training phrases, and has worked well by providing responses that match keywords.

*"Implementation of the BOOYER-MOORE Algorithm on the Yogyakarta Tourism Chatbot"* by Bunga Permata Sari, Tuhfatussalisah, Yogi Yulianto, Anggit Dwi Hartanto [7]. The system designed is a chatbot with Telegram as a virtual assistant for tourists. This system uses the Booyer-Moore algorithm by comparing patterns from right to left so that this algorithm becomes an efficient solution for searching. Based on verification, validity and prototype testing, the chatbot system can run well according to plan.

*"Creating an Online Store Website Equipped with Chatbot"* by Fredickson Dinata, Viny Christanti Mawardi, Janson Hendryli [8]. The application of the Vector Space Model method has been successful and runs according to the manual calculations that have been carried out to obtain results from the available question bank.

### 2.2. Artificial Intelligence and Natural Language Processing

Artificial intelligence refers to the ability of machines or computers to imitate some aspects of human intelligence. This involves developing algorithms and computational models that enable systems to perform tasks that typically require human intelligence, such as natural language understanding, decision making, pattern recognition. In addition, artificial intelligence has great potential to change fields including industry, health, transportation and others.

Natural Language Processing is a branch of artificial intelligence that focuses on the ability of computers to understand, process, and interact with natural human language. NLP involves the use of computational techniques and algorithms to analyze, parse and produce text or human speech in a form that computers understand.

Text Preprocessing is the initial stage in the Natural Language Processing method for text documents. Text Preprocessing prepares unstructured text into data that is structured and ready to be processed. The text preprocessing stages are shown in Figure 1.

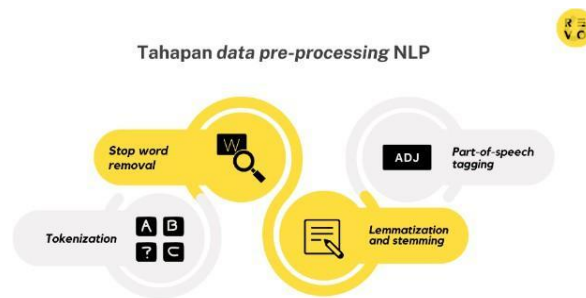


Figure 2 Stages of NLP Data Preprocessing [9]

### 2.3. Chatbot

Chatbot is short for “chat robot.” It refers to computer programs designed to interact with humans through conversation or chat, often in text form. Chatbots can use various technologies such as artificial intelligence (AI) and natural language processing (NLP) to understand and respond to human questions and statements. Chatbots can be found on various platforms such as websites, messaging apps, social media, and more. They can have varying levels of complexity, ranging from simple with fixed responses to very sophisticated with the ability to understand context, retrieve information from multiple sources, and provide more detailed answers. Examples of chatbot uses include automated customer service, virtual assistants, ordering services or products, initial medical consultations, and more. Chatbots have become popular due to their ability to provide fast and repeatable responses to users without the need for direct human intervention. There are two types of chatbots used, namely rule-based chatbots or retrieval chatbots and generative chatbots

### 2.4. Term Frequency – Inverse Document Frequency (TF-IDF)

The TF-IDF method is a way to weight the relationship of a word (term) to a document. This method combines two concepts for calculating weights, namely the frequency of occurrence of a word in a particular document and the inverse frequency of documents containing that word. The frequency with which a word appears in a given document indicates how important it is in that document. The frequency of documents containing the word indicates how common the word is. So the weight of the relationship between a word and a document will be high if the frequency of the word is high in the document and the overall frequency of documents containing that word is low in the document collection.

1. Calculation of Term Frequency (tf) using the equation

$$tf = tf_{ij} \quad (2.1)$$

$tf_{ij}$  is the number of occurrences of term  $t_i$  in document  $d_i$ . Term frequency (tf) is calculated by counting the number of occurrences of term  $t_i$  in document  $d_i$ .

2. Calculation of Inverse Document Frequency (idf), using the equation

$$idf = \log \frac{N}{df_i} \quad (2.2)$$

Where  $idf_i$  is the Inverse Document frequency, N is the journal document retrieved by the system and  $df_i$  is the number of documents in the collection where the term  $t_i$  appears in it.

3. Calculation of Term Frequency Inverse Document Frequency (Tf-Idf), using the equation

$$w_{ij} = tf_{ij} \times \log \frac{N}{df_i} \quad (2.3)$$

Where  $w_{ij}$  is the weight of the document,  $N$  is the number of documents retrieved by the system,  $tf_{ij}$  is the number of occurrences of term  $t_i$  in document  $d_j$  and  $df_i$  is the number of documents  $w_{ij}$  calculated to obtain the resulting weight multiplication or combination of term frequency  $tf_{ij}$  and inverse document frequency  $df_i$

The TF.IDF weight calculation is carried out by multiplying equations 1 with 2 to produce equation 3

$$W_{TD.IDF}(t_i, d_j) = (t_i \cdot d_j) \times (1 + \log(\frac{D}{df_i})) \quad (2.3)$$

Where  $W_{TD.IDF}(t_i, d_j)$  This is the section that describes how often the term  $t_i$  appears in document  $d_j$  (Term Frequency) and how rare it is in the document collection (Inverse Document Frequency).  $(1 + \log(\frac{D}{df_i}))$  This is the part that measures how often the term

$t_i$  appears in document  $d_j$  compared to the total number of words in document  $d_j$ , and also how rarely the term appears in the entire collection of documents ( $D$  is the total number document).

## 2.5. Cosine Similarity

Cosine Similarity is a type of Vector Space Model technique that is used to measure the similarity between a document and a query. In this model, queries and documents are considered as vectors in  $n$ -dimensional space, where  $n$  is the number of all terms in the lexicon. The lexicon is a list of all the terms in the index. One way to overcome this in the vector space model is to expand the vector. The expansion process can be carried out on the query vector, document vector, or both vectors [3]. Documents can be described as the following vector form:

$$Sim(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^n A_i \times B_i}{|\vec{q}| |\vec{d}|} = \frac{\sum_{i=1}^n w_{Ai} \cdot w_{Bi}}{\sqrt{\sum_{i=1}^n w_{Ai}^2} \sqrt{\sum_{i=1}^n w_{Bi}^2}} \quad (2.4)$$

Where  $Sim(\vec{q}, \vec{d})$  This is the cosine similarity score between vector  $q$  and vector  $d$ .  $\sum_{i=1}^n$

$$\frac{\sum_{i=1}^n A_i \times B_i}{|\vec{q}| |\vec{d}|}$$

This is the dot product between two vectors  $A$  and  $B$  having length  $n$ .  $\sqrt{\sum_{i=1}^n (w_{Ai})^2} \cdot \sqrt{\sum_{i=1}^n (w_{Bi})^2}$

This is the product of the lengths of vector  $A$  (represented by  $w_{Ai}$ ) and the length of the vector  $B$  (represented by  $w_{Bi}$ ), calculated using the square root of the sum of the squares of its components.  
t: Number of Vectors

## 3. Result and Discussion

### 3.1. System Architecture Design

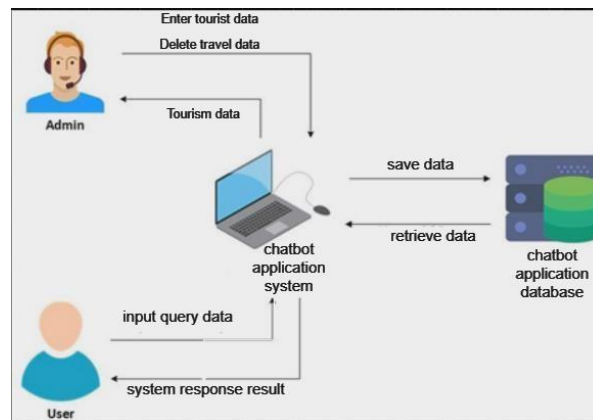


Figure 1. System Architecture Design

The image above is the architectural design for implementing Virtual Conversation using the Cosine Similarity Method in Tourism Service Applications East Kalimantan as an illustration of the flow of data that can be stored, processed and displayed back to the user.

### 3.2. System and Application Design

At this stage, system design is carried out by forming the overall system architecture. System design is described by creating a system flow design. The system flow concept can be seen in the process flow in Figure 3.2. The system flow concept consists of several parts, namely processing incoming data from the user using TF-IDF, then processing continues with weighting using TF-IDF and Cosine Similarity. The following is an explanation of the process flow in Figure 2.

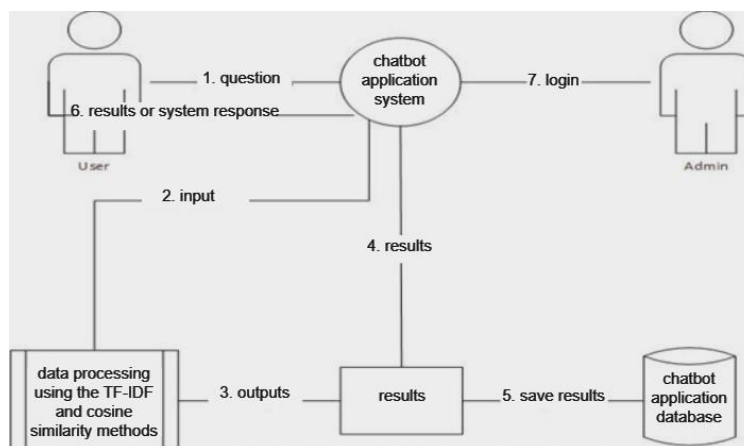


Figure 3. Chatbot System Process Flow

1. User enters a question into the chatbot system
2. Data processing using the Tf-Idf, Cosine Similarity method. Input data from the user where the system will read all input from the user is a question. Incoming questions will be processed where the same query is retrieved in the system database as the query question, then collected for the next process, namely Pre-Processing. Further data processing is in order, namely using Tf-Idf, Cosine Similarity as weighting.
3. The response output from the Chatbot is obtained from data processing using the Tf-Idf and Cosine Similarity methods.
4. From these results, the next process is entered into the chatbot system.
5. From the Chatbot system, the results of word processing will be stored in the Chatbot application

system database.

6. In this section the Chatbot system provides responses to the user in the form of answers to questions entered by the user.
7. In this section the admin can log in to the chatbot system

### 3.3. Data Flow Diagram (DFD)

Data Flow Diagram is a graphic representation of a system which is a system design tool oriented towards data flow. Data Flow Diagrams can be used as a modeling tool that allows systems professionals to describe the system as a network of functional processes that are connected to each other by data flow either manually or computerized. The following is an explanation of the DFD flow of the chatbot application system:

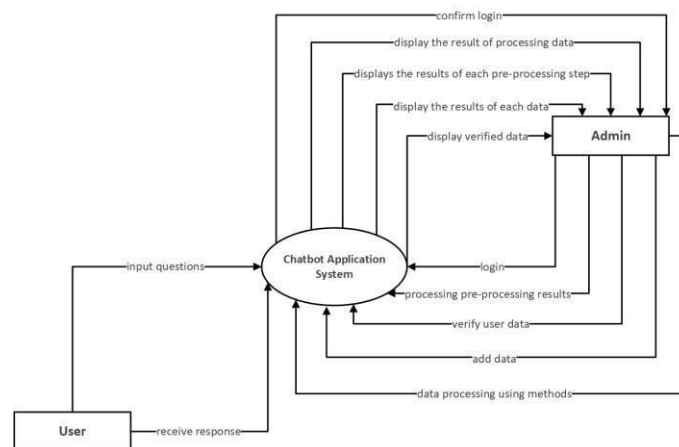


Figure 4. DFD Flow Of Chatbot System Process

The diagram above illustrates how the admin can add data (includes all data in the manual context required by the system), process data using methods (in this application system using supporting methods, namely Tf-Idf, Cosine Similarity), verify incoming data in the additional menu. on the Chatbot UI which is used as input information from the user), processing pre-processing results (sequential processing starting from the user query -> retrieving a query dataset that is similar to the user query -> tokenizing all queries -> stopwords), adding admin data, and login. Meanwhile, users can input questions and input the latest information.

### 3.4. System Implementation

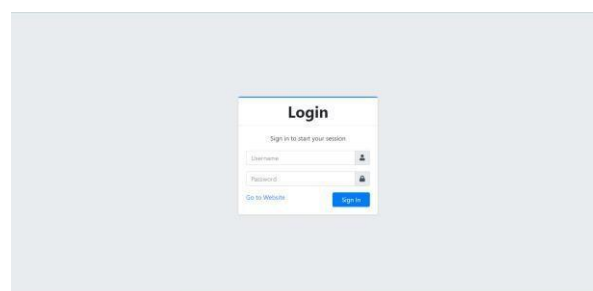
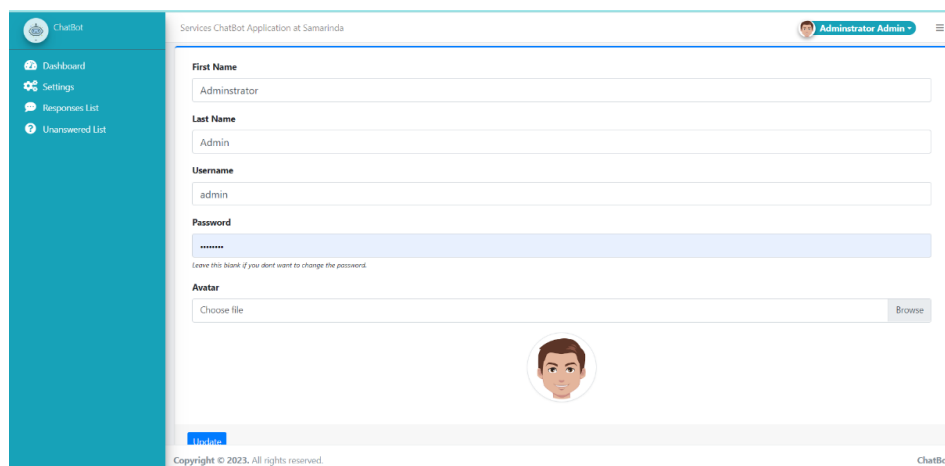


Figure 5. Login

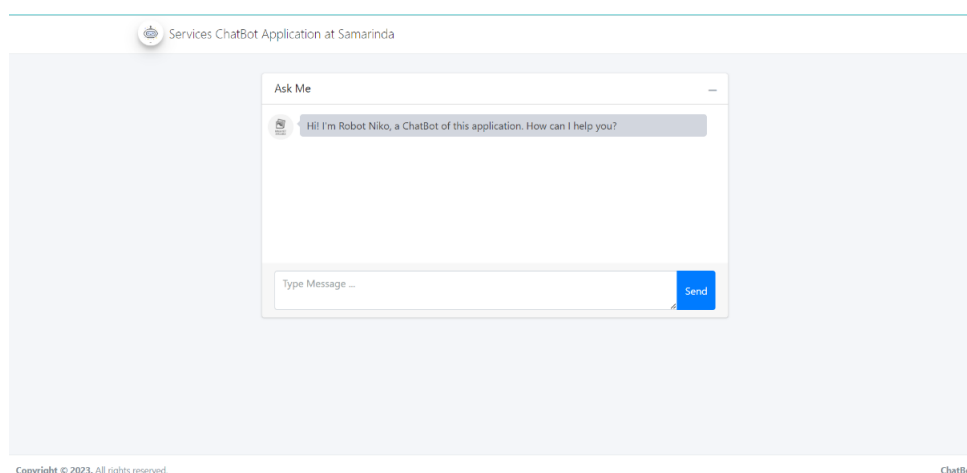
Login Module, is the initial application login display imagined by the user when the user runs the application. In this module there is a username, password and sign in button.



The screenshot displays the 'Admin chatbot' interface. On the left is a teal sidebar with navigation links: 'Dashboard', 'Settings', 'Responses List', and 'Unanswered List'. The main content area is titled 'Services ChatBot Application at Samarinda' and shows a user registration form. The form includes input fields for 'First Name' (containing 'Administrator'), 'Last Name' (containing 'Admin'), and 'Username' (containing 'admin'). The 'Password' field is masked with dots and has a small note below it: 'Leave this blank if you don't want to change the password.' Below the password field is an 'Avatar' section with a 'Choose file' button and a 'Browse' button. A circular avatar placeholder with a cartoon character is shown below the form. At the bottom left of the form is a blue 'Login' button. The footer contains 'Copyright © 2023. All rights reserved.' and 'ChatBot'.

Figure 6. Admin chatbot

This Admin module is a module that functions to provide an overview of the application. Making this module addresses the chatbot admin application as a chatbot setting.



The screenshot shows the 'Services Chatbot' interface. At the top, it says 'Services ChatBot Application at Samarinda'. The main area features a chat window titled 'Ask Me'. Inside the chat window, a message from the bot says: 'Hi! I'm Robot Niko, a ChatBot of this application. How can I help you?'. Below the chat window is a text input field labeled 'Type Message ...' and a blue 'Send' button. The footer contains 'Copyright © 2023. All rights reserved.' and 'ChatBot'.

Figure 7. Services Chatbot

This Services Chatbot module is a module where users interact in the form of questions about tourism in East Kalimantan.



### 3.5. Testing uses the Chatbot application

Cases and Test Result (True Case)			
Question	Expected Result	Observation	Conclusion
<ul style="list-style-type: none"> <li>- Tempat wisata di Kalimantan timur</li> <li>- Wisata apa yang ada pada Kalimantan timur?</li> <li>- Wisata Kalimantan timur?</li> <li>- Wisata apa saja di Kalimantan timur</li> <li>- Kalimantan timur</li> <li>- kalimantan</li> <li>- kaLiMantaN</li> <li>- Apa yang menarik di Kalimantan timur</li> <li>- Kalimantan Timur?</li> </ul>	<p>Kalimantan timur memiliki beberapa tempat wisata berupa Air terjun, bendungan, danau, goa, jembatan, kampung, Desa wisata, daerah wisata, masjid, wisata alam, museum, Pantai, pegunungan, pulau, Sungai, Taman, Telaga, air terjun, danau, gereja, lamin, taman, bukit dll</p>	<p>Bot answered correctly</p>	<p>[*] accepted [] rejected</p>
<ul style="list-style-type: none"> <li>- Apa saja wisata bendungan</li> <li>- Wisata yang bendungan?</li> <li>- Bendungan apa saja yang di Kalimantan timur?</li> <li>- Bendungan?</li> </ul>	<p>Wisata bendungan yang ada di Kalimantan timur berupa bendungan labanan dan merancang</p>	<p>Bot answered correctly</p>	<p>[*] accepted [] rejected</p>
Cases and Test Result (Wrong Case)			
Question	Expected Result	Observation	Conclusion
<p>Timur kalimantan</p>	<p>I am sorry. I can't understand your question. Please rephrase your question</p>	<p>Bots don't understand</p>	<p>[*] accepted [] rejected</p>

	and make sure it is related to this site. Thank you :)	and reply with a response default	
kalimantan selatan	I am sorry. I can't understand your question. Please rephrase your question and make sure it is related to this site. Thank you :)	Bots don't understand and reply with a response default	[*] accepted [] rejected

### 3.6. Result and anylist

Based on the test results overall, provides a conclusion that the application built can is said to have met the requirements functionally, even though there are some cases that have not been accepted, this is caused by data that has not been used at this time, this limitation will fade as data is added by the application development team

## 4. Conclusions

The conclusions from implementing Virtual Conversation using the Cosine Similarity method in tourism service applications in East Kalimantan are as follows:

1. Improved User Experience: The application of the Cosine Similarity method in Virtual Conversation can improve the user experience in searching for information and recommendations regarding tourism services in East Kalimantan. The ability of these algorithms to understand user preferences and provide appropriate recommendations can make the traveler experience more satisfying.
2. Service Efficiency: Travel service applications with Virtual Conversation can help optimize customer service. Travelers can quickly and easily get answers to their questions without having to wait for human customer service, thereby reducing waiting times and increasing service efficiency.
3. Potential Development: The Cosine Similarity method can be improved and further developed with improved data and AI models. This can help travel service applications become smarter in providing more accurate recommendations and information.
4. Language and Context Limitations: Although the Cosine Similarity method can provide good results, there are still limitations in understanding more complex languages and contexts. Therefore, please note that Virtual Conversation with this method may not always be able to address very complex or specific questions or requests.
5. User Satisfaction: The success of implementing Virtual Conversation using the Cosine Similarity method can be measured based on the level of user satisfaction. Regular evaluation of user satisfaction levels can help in improving these applications.

Thus, the application of Virtual Conversation using the Cosine Similarity method in tourism service applications in East Kalimantan has the potential to improve user experience, service efficiency and further development. However, it should be remembered that this implementation is not completely free from limitations and continuous evaluation needs to be carried out to ensure optimal service quality.

## References

- [1] Liza Chairunnisa, Wahyuni Eka Sari, Dawamul Arifin, "SISTEM INFORMASI GEOGRAFIS PEMETAAN TEMPAT WISATA DI KOTA SAMARIDA BERBASIS WEB," in *Teknologi Rekayasa Perangkat Lunak, Manajemen Pertanian East Kalimantan, Indonesia*, 2020.
- [2] Rishabh Shah, Siddhant Lahoti, dan Prof. Lavanya. K. , "'An Intelligent Chat-bot using Natural Language Processing'"," in *Department of Computer Engineering VIT University.*, 2017.
- [3] Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L., "'Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity.'," in *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2018.
- [4] Fahri Firdausillah, Arieansyah, "Implementasi Algoritma Levenshtein Distance Sebagai," in *Seminar Nasional APTIKOM (SEMNASITIK)*, 2019.
- [5] Dimas Wahyu Wibowo, Habibie Ed Dien, Trianta Almira Ramadhani, "Aplikasi Chatbot pada Sistem Informasi Penyewaan Scaffolding Menggunakan Metode TF-IDF," in *SEMINAR INFORMATIKA APLIKATIF POLINEMA (SIAP)*, 2020.
- [6] Dicki Wahyudi Harahap, Liza Fitria, "APLIKASI CHATBOT BERBASIS WEB MENGGUNAKAN," in *Jurnal Informatika dan Teknologi Komputer*, 2020.
- [7] Bunga Permata Sari, Tuhfatussalisah, Yogi Yulianto, Anggit Dwi Hartanto, "Implementasi Algoritma Booyer-Moore Pada Chatbot Wisata Yogyakarta," in *Technomedia Journal (TMJ)*, 2020.
- [8] Fredickson Dinata, Viny Christanti Mawardi, Janson Hendryli , "Pembuatan Website Online Store Dilengkapi dengan Chatbot," *Jurnal Ilmu Komputer dan Sistem Informasi*, 2021.
- [9] "<https://www.tempatwisata.pro/wisata/Kalimantan-Timur>".