

Chatbot-Based Movie Recommender System with Latent Semantic Analysis on Telegram Platform Using Dialog Flow

Antonius Randy Arjun, Z K Abdurahman Baizal*

Faculty of Informatics, Informatics, Telkom University, Bandung, Indonesia

Email: ¹randyarjun@student.telkomuniversity.ac.id, ^{2,*}baizal@telkomuniversity.ac.id

Submitted: 29/07/2022; Accepted: 20/08/2022; Published: 30/08/2022

Abstract—The growth in the number of movies continues to be experienced every year, making it difficult for users to choose the right movie. The recommender system is an alternative to being able to solve the problem. In many studies, recommender systems have been developed, but in their use, they do not apply intensive interaction between users and the system created. In this study, we developed a chatbot to help implement a movie recommender system that ensures users can interact intensively with the system with natural language. The chatbot was created using Dialog flow to enable the system to recognize the natural language. One way to understand a text concept is to find the relationship between the text concepts. Latent Semantic Analysis (LSA) can implement this, where LSA has the advantage of extracting a text and making a statistical representation in the form of a term-document matrix (TF-IDF) using a lower dimension (low-rank approximation). Singular Value Decomposition (SVD) can help decompose a large matrix into a matrix with small dimensions to determine a text's overall meaning. The relationship between text concepts with the highest or almost the same probability value can be used as an output to respond to the user. From the test results, the chatbot shows that the match rate between system and user responses is 86%. Thus, the developed chatbot can be used well in providing interactive movie recommendations.

Keywords: Chatbot; Dialog Flow; Natural Language Processing; Recommender System; Latent Semantic Analysis

1. INTRODUCTION

Rapid development continues to occur in the world of technology, which positively impacts the field of information and communication. The increased internet usage can prove this in people's lives. The survey results of the Indonesian Internet Network Operator Association (APJII) in 2019-2020 show that the penetration of internet usage is 196.71 out of 266.91 million people or around 73.7% of the total population of Indonesia [1], and it is estimated that this figure will continue to increase. Not only the internet but the world of cinema also continues to increase. This is because movies have become one of the most popular entertainment media in the community. However, the growth in the number of movies creates difficulties for users in choosing the right movie [1]. Therefore, a recommender system is needed to overcome this so that users can choose the right movie without having to experience difficulties. Also, with a recommender system, people do not need to spend money to find and choose the desired movie.

A recommendation system is a system that aims to estimate information that is of interest to users and also helps users in making their choices [2]. Several methods can be used to build recommender systems, such as collaborative filtering, content based filtering, and hybrid filtering [3]. The system needs information about user wants and needs to provide a recommendation. Various methods can be used to obtain this information by establishing a conversation with the user [4]. With a conversation, the system can directly interact with users and quickly find out information about the wants and needs of its users. To be able to implement it requires a system that can do this, such as a chatbot. The chatbot is usually used in a messaging application and can encourage human conversation using text chat, voice commands, or even both [5].

To date, many studies are related to recommender systems development. Rishabh Ahuja, et al., in 2019, develop a movie recommender system using KNN and k-means clustering [6]. The system was implemented using python language. This research proposes various recognition concepts related to machine learning and recommender systems. The results given in this study are that the system created using the KNN and k-means clustering can produce an RMSE value of 1.081648, which is better than existing techniques.

CheonSol Lee, et al., in 2022, developed a movie recommender system by utilizing the information contained in the user [7]. User information is then extracted using the BERT algorithm. The results given in this study are that the BERT algorithm with the IGMCM-based model is superior to the sophisticated graph-based conventional model. Geetha G, et al., in 2018, developed a system that can recommend a movie to both new and old users [8]. In this study, the system was developed by collecting all important information related to movie information. The developed recommender system uses collaborative filtering, content based filtering, and hybrid filtering. The results given in this study are that the developed system can provide precise movie recommendations to users. Syed M. Ali, et al., in 2018, developed a recommender system to help users choose movies [9]. This research developed the recommender system using a hybrid filtering model. This model utilizes tags from a movie combined using content-based filtering. The model was developed using Pearson correlation and principal component analysis. This research proves that movie tags can provide good results in finding similar movies, thus providing more accurate recommendations. Mayur Rahul, et al. in 2020, explained that in conducting research, problems are often found when processing large amounts of data [10]. In this study, a k-means method is proposed to perform classification and a singular value decomposition model that can reduce

the dimensions of large amounts of data. The developed method is then compared with existing methods based on the performance results generated from the MAE and RMSE values. The results of this study show that the developed method can provide better results than existing methods. From these studies, recommender systems have been developed using different methods and have proven their superiority. However, in the developed system, users are not allowed to interact intensively with the system, thus not allowing users to interact more with the system.

In this research, we built a movie recommender system as a chatbot on the telegram platform. First, we use the natural language processing approach as a process to make chatbots able to interact using natural language. Then the latent semantic analysis method is used to analyze the search for similarity values between concepts from a text. This test is carried out using the beta testing method, where beta testing is objective and can be done directly to users using a questionnaire related to user satisfaction with the chatbot that was built. With this chatbot-based movie recommender system, users can easily choose a movie according to their wishes without experiencing any difficulties.

2. RESEARCH METHODOLOGY

2.1 Research Stages

In this study, we developed a chatbot-based movie recommender system using latent semantic analysis. Overall, the flowchart of the chatbot is shown in Figure 1.

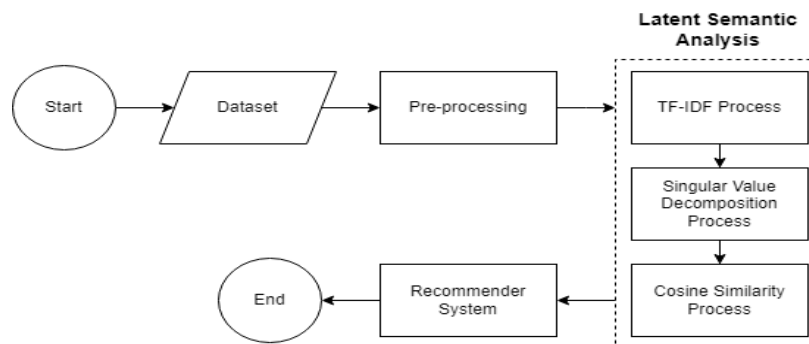


Figure 1. Flowchart System

2.2 Dataset

The first step in this research was to find a dataset. We get the movie dataset from the Kaggle platform. The dataset we use consists of 4803 rows and 20 columns. The next step is pre-processing, which aims to make the data we use easy to manage.

2.3 Pre-processing

Pre-processing is a process carried out to convert raw data into a form of data that is easy to understand. In this research, pre-processing consists of removing unused columns, converting data into strings, lowercasing to overcome sparsity problems [11], tokenizing as a process for text breakdown [12], stopwords removal, punctuation removal, and lemmatization.

2.4 Latent Semantic Analysis

Latent Semantic Analysis is a method that utilizes mathematical or statistical models to analyze and infer the semantic structure of a text. LSA can be used to assess essays by first converting essays into matrices which are then given a value on each provision to find similarities with the reference provided. In this research, the LSA method includes TF-IDF, SVD, and cosine similarity.

- Create a matrix A ($m \times n$) containing TF-IDF values where,
 - m : number of documents
 - n : unique word count
- Reducing dimension with SVD
Figure 2 is the LSA matrix where

$$\begin{array}{c} \overleftarrow{n} \\ \overrightarrow{m} \end{array} \begin{array}{|c|} \hline A \\ \hline \end{array} = \begin{array}{c} \overleftarrow{k} \\ \overrightarrow{k} \end{array} \begin{array}{|c|} \hline U_k \\ \hline \end{array} \begin{array}{c} \overleftarrow{k} \\ \overrightarrow{k} \end{array} \begin{array}{|c|} \hline S \\ \hline \end{array} \begin{array}{c} \overleftarrow{n} \\ \overrightarrow{k} \end{array} \begin{array}{|c|} \hline V_k^T \\ \hline \end{array}$$

Figure 2. LSA Matrix

U_k is the document term matrix, where k is the length of the vector.

V_k is a topic term matrix which is a vector representation of terms.

S is a diagonal matrix with non-negative numbers on the diagonal used for scaling

- c. Thus, SVD generates vectors for each term and document. These vectors will be used to find similar words and documents by calculating cosine similarity.

2.4.1 TF-IDF

TF-IDF is an algorithm used to calculate the weight of each word. TF-IDF will look for important terms to represent a document [13]. This research uses TF-IDF to generate a matrix to match the input and produce relevant recommendations [14]. Equation (1) is the formula for TF, and equation (2) is the formula for IDF.

$$TF = \frac{\text{Number of frequencies of the selected word}}{\text{word count}} \quad (1)$$

$$IDF = \log\left(\frac{\text{Number of documents}}{\text{Number of frequencies of the selected word}}\right) \quad (2)$$

2.4.2 Singular Value Decomposition

Singular Value Decomposition is used to factorize the TF-IDF matrix by decomposing a matrix into three parts. For example, suppose matrix A ($m \times n$) is decomposed using SVD; it will be decomposed into matrix U , matrix S , and VT as the equation (3).

$$A = U \times S \times V^T \quad (3)$$

2.4.3 Cosine Similarity

Cosine Similarity is used to estimate the level of similarity between documents and texts that have been made into a matrix. The cosine similarity process is done by calculating text similarity [15]. The calculation of cosine similarity can be done using the formula in equation (4).

$$\cos \theta = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| ||\vec{B}||} \quad (4)$$

2.5 Recommender System

A recommender system is an information filtering tool that estimates user preferences and makes recommendations based on those preferences [3]. With a recommender system, users do not need to spend much money to find and select the correct items [16]. The recommender system has three approaches: Collaborative Filtering, Content based Filtering, and Hybrid Filtering [3]. Collaborative Filtering provides item recommendations based on the similarity of preferences between users. Collaborative Filtering has two categories: model-based, used to estimate user ratings from unrated items, and memory-based, to provide ratings based on similarities between users or items [17]. Content based Filtering deals with profiles that have information about user preferences based on their assessment of items [18]. Hybrid Filtering combines content-based and Collaborative Filtering that can provide more effective and accurate recommendations compared to single recommendations[19].

2.6 Dialogflow

Dialogflow is a platform Google offers to develop computer-human interaction [20]. Dialogflow was initially known as Speaktioit and Api.ai as a platform to manage APIs for chatbots [21]. Google bought the company in September 2016 and then renamed it Dialogflow in 2017. Dialogflow supports several languages such as Indonesian, English, Japanese, Korean, and others [21]. In addition, Dialogflow integrates with almost any platform, including wearables, home apps, smart devices, and others [20]. This makes it one of the most popular platforms for developing chatbots.

3. RESULTS AND DISCUSSION

3.1 System Built

This research builds a chatbot-based movie recommender system using latent semantic analysis. First, the chatbot is implemented on the telegram platform and then developed using Dialog flow. The chatbot interaction mechanism has a flow, as shown in Figure 3.

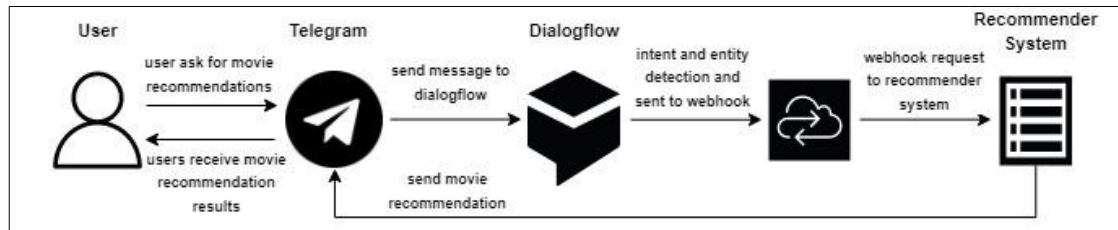


Figure 3. Chatbot Interaction Flow

3.2 Dataset

In this study, we used a movie dataset from a Kaggle platform. This dataset consists of 4803 rows and 20 columns. Table 1 is the dataset used in this study.

Table 1. Movie Dataset

Budget	Genres	Original_language	Title	Overview		Vote_count
237000000	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}]	en	Avatar	In the 22nd century, a paraplegic Marine is dispatched on a mission...		11800
300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}]	en	Pirates of the Caribbean: At World's End	Captain Jack Sparrow, long believed to be dead, has returned...		4500
0	[{"id": 1523, "name": "obsession"}, {"id": 224, "name": "Love"}]	en	My Date with Drew	Ever since the second grade when he first saw ...		16

3.3 Pre-processing

Pre-processing is a process carried out to convert raw data into a form of data that is easy to understand. In this research, pre-processing consists of removing unused columns, converting data into strings, lowercasing to overcome sparsity problems, tokenizing as a process for text breakdown, stop words removal, punctuation removal, and lemmatization. Table 6 is the result of pre-processing the dataset.

3.3.1 Lowercasing

After removing unused columns and converting data into strings, the next step is to perform lowercasing. Lowercasing is a stage used to convert a set of text into lowercase letters. Table 2 is the result of the lowercasing process.

Table 2. Lowercasing

Overview	Title
in the century, paraplegic marine is dispatch...	Avatar
captain barbossa, long believed to be dead, ha... cryptic message from bond's past sends him on..	Pirates of the Caribbean: At World's End Spectre
....	
ever since the second grade when he first saw ...	My Date with Drew

3.3.2 Tokenization

After lowercasing, the next step is tokenization. Tokenization is a stage for breaking text into several tokens or parts. At this stage, we perform tokenization to separate each word in the text, making it easier to find words. Table 3 is the result of the tokenization process.

Table 3. Tokenization

Tokens	Title
[in, the, century, paraplegic, marine, is, dis...	Avatar

Tokens	Title
[captain, barbossa, long, believed, to, be, de... [cryptic, message, from, bond, s, past, sends,... [ever, since, the, second, grade, when, he, fi...	Pirates of the Caribbean: At World's End Spectre My Date with Drew

3.3.3 Stopwords Removal

After performing Tokenization, the next step is to perform stopwords removal. Stopwords removal is a stage to remove words that have no meaning. At this stage, we perform stopwords removal to focus on essential words. Table 4 is the result of the stopwords removal process.

Table 4. Stopwords Removal

Tokens	Title
[century, paraplegic, marine, dispatched, moon... [captain, barbossa, long, believed, dead, come... [cryptic, message, bond, past, sends, trail, u... [ever, since, second, grade, first, saw, extra...	Avatar Pirates of the Caribbean: At World's End Spectre My Date with Drew

3.3.4 Lemmatization

After performing stopwords removal, the next step is to perform lemmatization. Lemmatization is a stage to find the root word in each word. At this stage, we perform lemmatization to make it easy to find similarities between words in a text. Table 5 is the result of the lemmatization process.

Table 5. Lemmatization

Tokens	Title
[century, paraplegic, marine, dispatch, moon, ... [captain, barbossa, long, believe, dead, come,... [cryptic, message, bond, past, send, trail, un... [ever, since, second, grade, first, saw, extra...	Avatar Pirates of the Caribbean: At World's End Spectre My Date with Drew

3.3.5 Data Pre-processing Results

In data pre-processing, we have done all the necessary steps such as deleting unused columns, converting data into strings, lowercasing, tokenizing, stopwords removal, punctuation removal, and lemmatization. This is so that the data we have can be easily understood. Table 6 is the result of the data pre-processing process.

Table 6. Data Pre-processing Results

Tokens	Title
century paraplegic marine dispatch moon pando... captain barbossa long believe dead come back ... cryptic message bond past send trail uncover ever since second grade first saw extraterres...	Avatar Pirates of the Caribbean: At World's End Spectre My Date with Drew

3.4 TF-IDF

TF-IDF is a statistical-based technique used to calculate the weight of each word. TF-IDF will look for important terms to represent a document. This research uses TF-IDF to generate a matrix to match the input and produce relevant recommendations. TF-IDF consists of the terms Term Frequency (TF) and Inverse Document Frequency (IDF). Table 7 is the result of the TF-IDF matrix.

Table 7. TF-IDF Matrix

Abandon	Ability	Aboard	Absence	Abuse	Abusive		Zoo
0,174251016	0	0	0,227694402	0	0		0
0	0	0,191541253	0	0	0		0
....							
0,106053864	0	0	0	0	0		0

3.5 Latent Semantic Analysis

Latent Semantic Analysis or commonly called LSA, is a method that utilizes mathematical or statistical models to analyze and infer the semantic structure of a text. For example, LSA can be used to assess essays by first

converting essays into matrices which are then given a value on each provision to find similarities with the reference provided. In this research, the LSA method includes TF-IDF, SVD, and cosine similarity. Table 8 is the result of the latent semantic analysis process.

Table 8. TF-IDF and SVD Results

Tokens	Abandon	Ability		Zoo
superman return discover absence allow lex lu...	0,174251016	0		0
captain jack sparrow cross path woman past su...	0	0		0
....				
peter parker outcast high schooler abandon pa...	0,106053864	0		0

3.6 Cosine Similarity

Cosine Similarity is a method used to find similarity by calculating the cosine angle between two vectors. The cosine similarity value has a range between 0 and 1. A value of 0 indicates a weak similarity between two vectors. Conversely, a value of 1 indicates a strong similarity between two vectors. In this research, cosine similarity is used to find the similarity of TF-IDF calculation in the token column. Table 9 is the result of cosine similarity in the tokens column that TF-IDF and SVD have processed.

Table 9. Cosine Similarity Results

0,715181638	0,004371183	0,000129671	0,064105143		0,033306369
0,004371183	0,595770531	0,020962331	0,004888159		0,002624688
0,000129671	0,020962331	0,609227379	0,000225337		-0,00484967
....					
0,011287336	0,033187058	0,001589898	0,008563063		0,003856583

3.7 Chatbot

A chatbot is an online conversation between human computers using natural language. The chatbot is an application that manages conversation services with instant messaging and the right way. Using a chatbot is faster and less confusing than other applications. To start, it is effortless, without the need to install a new application. In this research, the recommender system is implemented using a chatbot. Then the chatbot is developed using Dialogflow, which is integrated with the telegram platform. Figure 4 is the result of the chatbot system that was built.

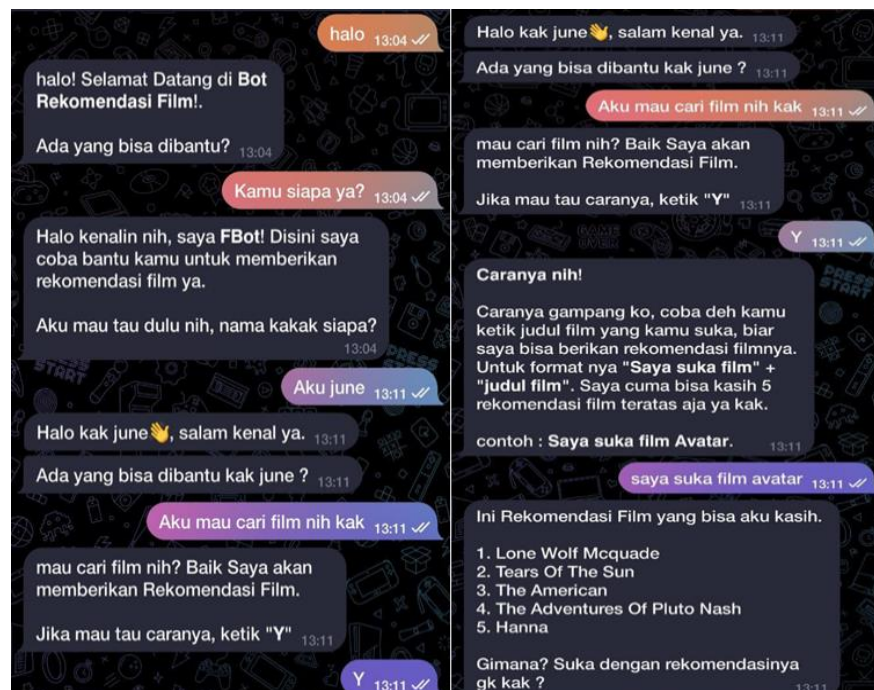


Figure 4. Chatbot

3.8 Recommender System

A recommender system is an information filtering tool that estimates user preferences and makes recommendations based on those preferences. In this research, the recommender system used is content-based filtering, where the assessment deals with profiles that have information about user preferences based on their

assessment of items. Figure 5 is the result of implementing the recommender system using content-based filtering. Users get movie recommendations based on the input given to the chatbot system.

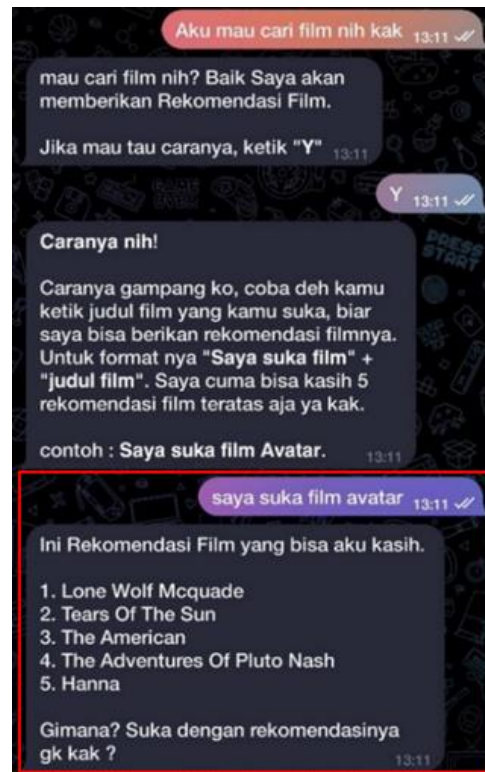


Figure 5. Movie Recommendation Results

3.9 Testing

The tests carried out in this study were carried out using the beta testing method. This beta testing is intended so the system can be tested directly to users. In this case, beta testing is carried out using a questionnaire to find user satisfaction and feedback on the system being built. Calculations to assess the results of questionnaires given to users can be calculated using the formula found in (5).

$$\text{Final Results} = \frac{\text{Total Score}}{\text{Maximum Score}} \times 100\% \quad (5)$$

Table 10 contains a list of questions asked to users consisting of several assessment weights. Finally, Table 11 explains the score assessment based on predetermined weights.

Table 10. Question

Points	Question	Score Assessment				
		TS	KS	CS	S	SS
1.	Do you know what to do when opening a chatbot for the first time?					
2.	Are you quickly getting used to chatbot?					
3.	Are you able to express your needs flexibly?					
4.	Do the movie recommendations match your preferences?					
5.	Do you find chatbot helpful?					

Table 11. Assessment Weight

Weight	Description	Score
TS	Disagree	1
KS	Less Agree	2
CS	Fairly Agreed	3
S	Agree	4
SS	Strongly Agree	5

3.10 Testing Results

In this study, testing was carried out using beta testing. This test is carried out by asking several questions to users related to the user experience when interacting directly with the chatbot. Chatbot testing is carried out

directly by the user and will produce objective feedback, and then the data results are used to assess the test results. Table 12 is the test result assessing each answer submitted to the user. The assessment results are calculated using formula (5).

Table 12. Testing Results

Question	Score Assessment					
	TS	KS	CS	S	SS	Percentage (%)
Do you know what to do when opening a chatbot for the first time?			3	24	15	(42/50)*100%=84%
Are you quickly getting used to chatbot?				8	40	(48/50)*100%=96%
Are you able to express your needs flexibly?			9	16	15	(40/50)*100%=80%
Do the movie recommendations match your preferences?			6	24	10	(40/50)*100%=80%
Do you find chatbot helpful?				20	25	(45/50)*100%=90%
Final Results						86%

3.11 Analysis of Test Results

Based on the results of the tests that have been carried out, it can be concluded that the chatbot recommender system can answer user questions and provide a good response, with a final percentage of 86%. The difference from the final result of 14% is due to the limited number of words that can detect the use of slang, so the chatbot system does not understand what the user is saying.

4. CONCLUSIONS

This research concludes that the recommender system implemented into the chatbot and developed using Dialogflow can provide good response for users. This can be seen in Table 12, the fourth point with a percentage of 80% chatbot can provide movie recommendations that match user preferences. Furthermore, the second and third points have a difference of 16%, whereas the second point has a percentage of 96%. In comparison, the third point has a percentage of 80%, proving that the chatbot provides convenience for its users because the use of natural language supports it.

REFERENCES

- [1] APJII, "Hasil Survei Internet APJII 2019-2020-Q2."
- [2] P.-C. Lin, "Movie Recommender Chatbot Based on Kansei Engineering," *Journal of Soft Computing and Data Mining*, vol. 1, no. 1, pp. 36–45, 2020.
- [3] P. G. Padti, K. Hegde, and P. Kumar, "Hybrid Movie Recommender System," vol. 4, no. 7, pp. 311–314, 2021.
- [4] D. Theosaksomo and D. H. Widyantoro, "Conversational Recommender System Chatbot Based on Functional Requirement," *TSSA 2019 - 13th International Conference on Telecommunication Systems, Services, and Applications, Proceedings*, pp. 154–159, 2019.
- [5] S. Shivanand, K. S. Pavan Kamini, M. M. Bai N, R. Ramesh, S. H. R, and U. Student, "Chatbot with Music and Movie Recommendation based on Mood," *International Journal of Engineering Research & Technology*, vol. 8, no. 15, 2020.
- [6] R. Ahuja, A. Solanki, and A. Nayyar, "Movie recommender system using k-means clustering and k-nearest neighbor," *Proceedings of the 9th International Conference On Cloud Computing, Data Science and Engineering, Confluence 2019*, pp. 263–268, 2019.
- [7] C. Lee, D. Han, K. Han, and M. Yi, "Improving Graph-Based Movie Recommender System Using Cinematic Experience," *Applied Sciences (Switzerland)*, vol. 12, no. 3, 2022.
- [8] G. Geetha, M. Safa, C. Fancy, and D. Saranya, "A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System," *Journal of Physics: Conference Series*, vol. 1000, no. 1, 2018.
- [9] S. M. Ali, G. K. Nayak, R. K. Lenka, and R. K. Barik, *Movie Recommendation System Using Genome Tags and Content-Based Filtering*, vol. 38. Springer Singapore, 2018.
- [10] M. Rahul, V. Kumar, V. Yadav, and Rishabh, "Movie recommender system using single value decomposition and K-means clustering," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, 2021.
- [11] A. Dahale, "A Natural Language Processing Approach for Musical Instruments Recommendation System Abhishek Dahale National College of Ireland Supervisor :," *Dublin, National College of Ireland*, 2019.
- [12] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017.
- [13] H. Zhou, "Research of Text Classification Based on TF-IDF and CNN-LSTM," *Journal of Physics: Conference Series*, vol. 2171, no. 1, pp. 218–222, 2022.
- [14] T. A. W. Tyas, Z. K. A. Baizal, and R. Dharayani, "Tourist Places Recommender System Using Cosine Similarity and Singular Value Decomposition Methods," *Jurnal Media Informatika Budidarma*, vol. 5, no. 4, p. 1201, 2021.
- [15] U. Open, S. Frameworks, A. Singh, K. Ramasubramanian, and S. Shivam, *Building an Enterprise Chatbot Building an Enterprise*.
- [16] A. T. Ciaputra and S. Hansun, "Rekomendasi Pemilihan Film Dengan Hybrid Filtering Dan Knearest Neighbor," *Jurnal Rekayasa Informasi*, vol. 9, no. 2, pp. 101–109, 2020.
- [17] M. Rizqi Az Zayyad and A. Kurniawardhani, "Penerapan Metode Deep Learning pada Sistem Rekomendasi Film," *Automata*, vol. 2(1), 2021.

- [18] G. Gadikar, “Towards a Hybrid Personalized Movie Recommender System,” vol. 9879, no. 978, pp. 71–74, 2018.
- [19] V. Vellaichamy and V. Kalimuthu, “Hybrid collaborative movie recommender system using clustering and bat optimization,” *International Journal of Intelligent Engineering and Systems*, vol. 10, no. 5, pp. 38–47, 2017.
- [20] A. Eikonsalo, “Utilizing Bots in Delivering Content from Kentico Cloud and Kentico Antti Eikonsalo,” no. October, 2017.
- [21] A. Y. Chandra, D. Kurniawan, and R. Musa, “Perancangan Chatbot Menggunakan Dialogflow Natural Language Processing (Studi Kasus: Sistem Pemesanan pada Coffee Shop),” *Jurnal Media Informatika Budidarma*, vol. 4, no. 1, p. 208, 2020.