JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.XXXX

HYBRID CONTENT-BASED FILTERING AND CLASSIFICATION RNN WITH PARTICLE SWARM OPTIMIZATION FOR TOURISM RECOMMENDATION SYSTEM

Syahdan Naufal Nur Ihsan[1*]; Erwin Budi Setiawan[2]

Informatics[1,2]
Telkom University[1,2]
http://telkomuniversity.ac.id/ [1,2]
syahdannaufal@student.telkomuniversity.ac.id[1*],erwinbudisetiawan@telkomuniversity.ac.id[2]

(*) Corresponding Author

Abstract -- *Economic recovery in the tourism sector after the COVID-19 pandemic is one of the main focuses of the Indonesian government at the moment, especially in Bandung City. This research aims to develop a personalized tourist spot recommendation system, by addressing the gaps in the existing literature through the integration of Content-Based Filtering (CBF) and Simple Recurrent Neural Network (RNN) methods that aim to improve recommendation accuracy. This study uses a hybrid approach that combines Term Frequency - Inverse Document Frequency (TF-IDF) and word embedding with the Robustly Optimized BERT (RoBERTa) model to identify similarities between tourist destinations based on their content characteristics. Simple RNN is used to analyze user preference patterns over time, which is then further optimized using Particle Swarm Optimization (PSO). As a result, the Simple RNN model that has been optimized with PSO shows an increased accuracy of up to 94.37%, outperforming other optimizations such as Adam and SGD. This research makes a novel contribution by applying advanced machine learning techniques to improve personalization in travel recommendation systems.*

Keywords: *Content-Based Filtering, PSO, RNN, Roberta, TF-IDF.*

Intisari -- Pemulihan ekonomi di sektor pariwisata pasca-pandemi COVID-19 menjadi salah satu fokus utama pemerintah Indonesia saat ini, terutama di Kota Bandung. Penelitian ini bertujuan untuk mengembangkan sistem rekomendasi tempat wisata yang sudah dipersonalisasi, dengan mengatasi kesenjangan dalam literatur yang ada melalui integrasi metode Content-Based Filtering (CBF) dan Recurrent Neural Network (RNN) sederhana (Simple RNN) yang bertujuan untuk meningkatkan akurasi rekomendasi. Studi ini menggunakan pendekatan hibrida yang menggabungkan Term Frequency - Inverse Document Frequency (TF-IDF) dan word embedding dengan model Robustly Optimized BERT (RoBERTa) untuk mengidentifikasi kesamaan antar destinasi wisata berdasarkan karakteristik kontennya. Simple RNN digunakan untuk menganalisis pola preferensi pengguna dari waktu ke waktu, yang kemudian dioptimalkan lebih lanjut menggunakan Particle Swarm Optimization (PSO). Hasilnya, model Simple RNN yang telah dioptimalkan dengan PSO menunjukkan peningkatan akurasi hingga 94,37%, mengungguli optimasi lain seperti Adam dan SGD. Penelitian ini memberikan kontribusi baru dengan mengaplikasikan teknik machine learning yang canggih untuk meningkatkan personalisasi dalam sistem rekomendasi wisata.

Kata Kunci: *Content-Based Filtering, PSO, RNN, Roberta, TF-IDF.*

INTRODUCTION

After two years of the COVID-19 pandemic, governments worldwide are intensively making post-pandemic recovery programs, including the government of the Republic of Indonesia. The primary recovery program was carried out in the tourism sector, considering that this sector contributed 5.5% to GDP and contributed 16,910 million USD to the country's foreign exchange in 2019. Unfortunately, during the COVID-19 pandemic, the tourism sector's contribution decreased significantly to 4% of GDP in 2020 and only contributed 540 million USD to the country's foreign exchange [1]. The policy has been taken by the Indonesian government to increase the tourism sector due to economic recovery was considered appropriate. The tourism sector has unlimited resources, such as nature, culture, and society [2]. This sector is also the most flexible sector with various international events such as G20, World Water Forum, and the Annual Commemoration of the Asia-Africa Conference in Bandung. This was confirmed by the workforce's contribution, which continues to increase, from 21.26 million people in 2021 to 22.89 million in 2022 [3].

The personalization of the user experience in most areas have often been touched by artificial intelligence (AI) tools, with tourism and entertainment sectors making noteworthy use AI-enabled interventions. A related trend that Ricci thoroughly describes [4] involves the use of AI-driven recommendation systems in the tourism sector to personalize destination, accommodation, and activity recommendations based on user preferences/performance. The main problem was AI in content based recommendation system and its extension to different domains. Abhinav showed how to recommanding movies by using cosine similarity and this approach can be reused for tourism to make the recommendation more accurate [5]. Likewise, in entertainment related field [5], Salsabil et al. showed the power of contentBasedFiltering (CBF) to generate relevant movie recommendations. They used RoBERTa-SGD (95.59%) which implements advanced NLP method (as TF-IDF, BERT…) and a RNN-based model for the task of paraphrase classification [6]. If AI-algorithms have been so successful in the field of entertainment engines it should be possible to take a similar approach on tourism recommender systems as well and provide an experience where people can receive offers that match their specific interests.

Another study was conducted by Jawarneh et al. using a collaborative filtering (CF), CF approach refers to one of the components in developing a recommendation system that considers contextual information and user preferences [7]. This research uses Large Language Models (LLMs) with a transfer learning approach to analyze customer reviews. The model has been used as a GPT (Generative Pre-trained Transformer). This research obtained an MAE value of 0.27, MSE of 0.26, R-squared of 0.69, and MAPE of 0.1 [8].

Another research carried out using the CF-CBF hybrid method approach by several researchers. CF-CBF Hybrid Method is a combination of the two methods to obtain personalized preferences collectively [9]. These studies also use hybrid deep learning methods to improve performance. However, this technique is not popular because, in future developments, there will be other approaches that have better performance. One of them was conducted by Xiaoyan et al., who used GloVe as a feature extraction algorithm and CNN-LSTM as a classification algorithm [10]. Another research was conducted by Amalia & Winarko who used Word Embedding as a feature extraction algorithm and CNN as a classification algorithm [11]. Another research was carried out by Hossain by combining CNN and LSTM [12].

In subsequent developments. apply the TF-IDF method to assign weights to words. Then, cosine similarity to calculate similarity based on TF-IDF weighted word vectors. One of the studies that used this method was conducted by Yunanda et al. This research uses the TF-IDF method and Cosine Similarity is an effective approach to overcome the problem of "information overload" in order to provide better recommendations for users[13]. Other studty was cunducted using Hybrid TF-IDF and Cosine Similarity Algorithms approach to create a movie recommendation system on Netflix [14]. Furthermore, the most recent study was conducted by Tan et al. using Hybrid RoBERTa-GRU [15]. The Hybrid method can be a solution to using Large Language Models (such as Roberta) on data with little context.

This research integrates the content-based filtering (CBF) approach with hybrid methods involving Large Language Models (LLMs) and weighting algorithms to develop an improved recommendation system for tourist destinations in Bandung. Specifically, the research employs a Hybrid Method combining Term Frequency-Inverse Document Frequency (TF-IDF) and word embedding using the Robustly Optimized BERT approach (Roberta). The classification process is managed through a Recurrent Neural Network (RNN) optimized by Particle Swarm Optimization (PSO).

The primary objective of this method is to enhance both the performance and quality of the recommendations generated, marking a novel contribution to this field. The PSO algorithm iteratively updates particle positions and velocities based on their individual fitness values and the globally best solution identified by the swarm. The bi-level structure of the algorithm enhances performance by promoting diversity, improving search efficiency, and preventing premature convergence [16].

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.XXXX

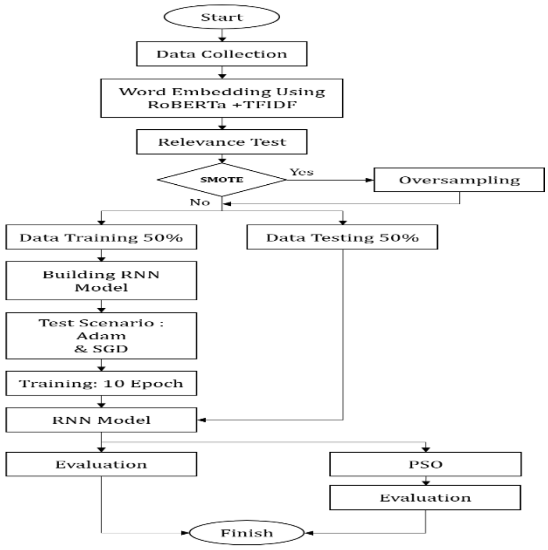
This optimization approach has demonstrated significant benefits in terms of cost efficiency, performance, and energy management [17]. In the tourism sector, similar hybrid models have successfully delivered personalized travel recommendations, analyzed tourist reviews, and predicted visitor satisfaction, thereby enhancing the overall tourist experience [18]. These methodologies are particularly effective in addressing the dynamic and heterogeneous nature of tourism data, making them highly relevant for destination recommendation systems [19]. The evaluation metrics employed in this study include MAE, RMSE, and accuracy.

This research proposes the integration of content-based approach (CBF) with a hybrid method involving Large Language Models (LLMs) and weighting algorithms to develop a better recommendation system for tourist destinations in Bandung. Specifically, this research uses a hybrid method that combines Term Frequency-Inverse Document Frequency (TF-IDF) and word embedding with the Robustly Optimized BERT (RoBERTa) approach, where the classification process is performed through a Recurrent Neural Network (RNN) optimized with Particle Swarm Optimization (PSO). This system is expected to provide significant benefits for tourists and the tourism industry, especially in increasing the visibility of lesser-known destinations and increasing the attractiveness of Bandung City as a competitive tourist destination.

MATERIALS & METHOD

A. Research Design

The research design of this study can be seen in Figure 1.



Source : (Research Results,2024)
Figure 1. Research Design

This research started by collecting data from verified sources such as reviews, tweets, and previous research. Then, the preprocessing stage is carried out by grouping the data into review values (points) from these sources. Then, the description content of tourist destinations from these sources enters the embedding process using Hybrid RoBERTa and TF-IDF. The relevance of the content was assessed based on the suitability of the embedding results and tested with Loss parameters using MAE & RMSE. In the testing phase using SMOTE (balancing data), the data will enter the oversampling process first. Then, the data is divided into training data and testing data in a 50:50 ratio.

The next stage is the modeling stage, which has been begun by creating an RNN model using the SimpleRNN layer. Then, the training process has been conducted on 10 itteration by implementing test scenarios using the Adam and SGD optimizers. The trained models then has been tested using testing data and then evaluated. The final stage was optimize the model using PSO. Then, compare each performance value. The research was considered successful if one of the test scenarios performed well.

B. Datasets

This research uses two main datasets, user dataset and destination dataset. The user dataset was obtained through a crawling process on Twitter, where we collected data from usage history based on user activities related to tourism, and we conducted web scraping to collect ratings of tourist destinations in the city of Bandung. In addition, we collected tourist attractions by the Bandung City Tourism Office, which includes place reviews related to tourist attractions in the city.

In total, 95 user data were collected. This data did not come from one single source, but from various sources, namely Twitter, Website and Tourism Office.

As shown in Table 1, here are the results of crawling user data on twitter that has passed the process of cleaning text, translate text.

Table 1 . Data Twitter

Users	Place_Name	Cleaned_text
Mapay bandung com	Braga	The origin of the name ... this drink
...	...	...
Jacktour bandung	Farm House	New tourist attraction ... sell milk

source : (Research Results,2024)

As shown in Table 2, he following is the process of analyzing the sentiment of cleaned_text and then converting the sentiment into a 1-5 rating scale. Sentiment analysis is performed using the VADER Sentiment Intensity Analyzer from the NLTK library. The result is a rating for each review that can be used for further analysis, such as to be able to generate recommendations.

Table2. Data Rating Twitter

Users	Place_Name	Cleaned_text	Rating
Mapay bandung com	Braga	The origin of the name ... this drink	4
...	...	...	
Jacktour bandung	Farm House	New tourist attraction ... sell milk	4

source : (Research Results,2024)

Table 3 . Users Datasets

Users	Braga St	...	Farm House
TripAdvisor	4.0	...	0.0
klook	0.0	...	4.6
solidify	4.5	...	0.0
...	...	...	...
Agoda	0.0	...	0.0

Source : (Research Results,2024)

As shown in Table 2 and Table 3, the dataset format between the crawling results from twitter and web scrapping already has the same format, namely each user already has a rating value for each tourist spot they have visited, which data is ready to be processed.

As shown in Table 4,The following is a dataset of tourism places in the city of Bandung that are ready to be processed to produce a recommendation system for users.

Table 4. Destination Datasets

Place_Name	Description
Gunung Tangkuban Perahu	Gunung Tangkuban Parahu adalah salah satu gunung ...
Braga St	Jalan Braga adalah nama sebuah jalan utama di ...
...	...
Kota Mini	Destinasi yang sangat menarik bernuansa ...
Chingu Cafe Little Seoul	Selain populer karena memiliki pemandangan yang ...

source : (Research Results,2024)

In total, there are 95 users and 99 tourist destinations in the city of Bandung.

C. Word Embedding

The word embedding method used in this research is an enhanced model, which is a hybrid model between Term Frequency - Inverse Document Frequency (TF-IDF) and Robustly Optimized BERT (RoBERTa) approach. TF-IDF is an algorithm that measures the importance or relevance of a word representation in a document among a set of documents (corpus) [14]. Roberta is a transformer-based Large Language Model (LLM) that uses a self-attention layer to process input sequences and generate contextual representations of words in a sentence [15]. The embedding process has been done twice. First, Roberta is used to get the embedded

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.XXXX

tokens, and then TF-IDF is used to get the weight of the embedded tokens.

The implementation of artificial intelligence in this research on the combination process of TF-IDF and RoBERTa, will result in a deeper ability to understand the context and relevance of the content. One of the implementations of this application in the tourism sector is the development of a tourist destination recommendation system that can automatically adjust recommendations based on individual preferences and current travel trends[20]. By analyzing the preferences of the user, the system can generate more relevant recommendations, guide tourists to destinations they might be interested in visiting, and ultimately increase the number of tourist visits [21]. Therefore, to achieve the goal of improving the tourism sector, it is very useful in terms of optimizing marketing strategies and dynamic pricing based on travel demand and trends, thus not only increasing tourist satisfaction but also maximizing revenue for tourism industry [22].

This system combines the use of TF-IDF and RoBERTa methods to identify and generate weights on relevant content features, which will then generate more personalized and contextual recommendations for tourists. The main benefit for travelers is that it allows the system to provide destination recommendations that are more relevant to individual preferences, allowing them to discover places that they may not have visited before but are highly compatible with their interests based on previous user profiles [15]. For the tourism industry in Bandung, the system helps to increase the visibility of lesser-known destinations and distribute tourist visits more evenly, which ultimately contributes to improving the local economy and strengthening the competitiveness of the tourism sector [11].

D. Relevance Test

Relevance tests were used to measure the performance of content-based filtering [6]. The method used was Linear Regression. The performance metrics used were Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE is the average absolute error value that measures the context difference between one content and another [8]. The formula for MAE is as follows:

$$MAE = \frac{1}{N} \sum | - \hat{y}| \qquad (1)$$

Meanwhile, RMSE is the root value of the average square error. This value also represents the context difference between content and another [6]. The formula for RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum (y - \hat{y})^2} \qquad (2)$$

The smaller the MAE and RMSE values, the more relevant the content will be to each other.

Confusion Matrix in this study is to analyze the results of RNN classification. Confusion Matrix is useful for calculating true positive, false positive, true negative, and false negative, which helps in understanding the effectiveness of content-based filtering models in correctly predicting relevant content [23]. The confusion matrix used in this research can be seen in the table. 5

Table.5. Confusion Matrix

	Predicted Relevant	Predicted Not Relevant
Actual Relevant	True Positive (TP)	False Positive (FP)
Actual not Relevant	False Negative (FN)	True Negative (TN)

From this Confusion Matrix, several metrics can be calculated to evaluate the model's performance, namely **accuracy**, **precision**, **recall**, and **F1-score**. These metrics are calculated using the following formulas:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

$$F1\ Score = 2\,x \frac{Precision\ x\ Recall}{Precision + Recall} \qquad (6)$$

Confusion Matrix is particularly useful in evaluating Recurrent Neural Networks (RNNs), which are often used for sequence prediction tasks, such as language modeling, time series forecasting, and sentiment analysis. For RNNs, Confusion Matrix helps in understanding how

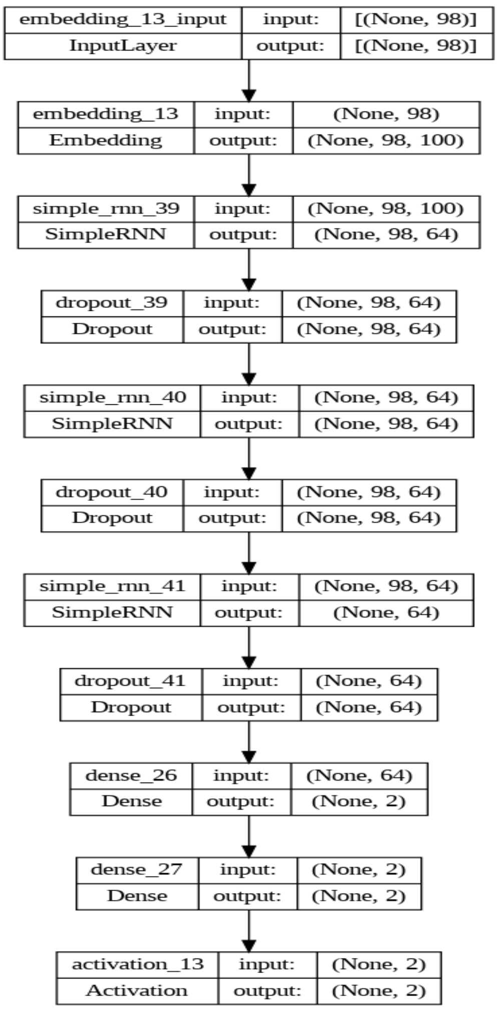
VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i1.XXXX

well the model distinguishes between different classes over time, especially in the context of predicting sequential data. It provides insight into how effectively the RNN captures dependencies in the sequence and accurately predicts the next element in the series [24]

By combining MAE, RMSE, and Confusion Matrix analysis, a robust and detailed evaluation of the performance of content-based filtering systems can be achieved.

E. Model Architecture

This research used several RNN layers and Dropout layers. The activation function used was Sigmoid. The complete used architecture can be figure out on Figure 2.



Source : (Research Results,2024)

Figure 2. Model Architecture

As shown in Figure 2. the model applied in this study using multiple layers of Recurrent Neural Networks (RNN) and Dropout, which has been designed to capture and enhance the temporal dependency of the input data which serves to reduce overfitting. There is an embedding layer that functions as an initial stage, i.e., converting the 98-long input sequence into a dense vector with a dimension of 100, which can effectively capture the semantic nuances of the data [25]. This transformation is important to generate a meaningful representation to the next RNN layer.

This architecture also includes three SimpleRNN layers, each with an output dimension of 64. SimpleRNN is fundamental in processing sequential data that serves to preserve and propagate temporal context across the sequence. The reason for using multiple RNN layers is to allow the model to learn a hierarchical temporal representation, where each layer captures a different level of temporal abstraction [26]. After each RNN layer, a Dropout layer is used, with the dropout rate set to reduce the risk of overfitting by randomly setting a portion of the input units to zero during training [27]. This regularization technique ensures that the model can be generalized well.

This architecture ends with a Dense layer, followed by a final activation layer that uses a Sigmoid function. The Dense layer maps the learned features from the RNN layer to the output space, which in this implementation consists of two units corresponding to binary classifications. Sigmoid activation is used here to generate a probability distribution between the two classes, ensuring that the output can be interpreted as a probability [28].

Overall, the combination of RNN and Dropout layers is essential in this architecture, as it allows the model to effectively learn and generalize from sequential data. The use of these layers, along with the embedding and dense layers, forms a powerful framework capable of handling the complexity of temporal sequence modeling.

F. Test Design

The test scenario was conducted several times by comparing the performance of data enriched with SMOTE Oversampling and those not enriched with SMOTE Oversampling. Then, test scenarios were also conducted by comparing Adam and SGD optimizers. Finally, the test scenario by applying the PSO algorithm. The selection of SMOTE is due to its ability to overcome class imbalance problems by generating new samples from minority classes,

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.XXXX

which has been shown to improve model performance on imbalanced data [29].

Furthermore, the test scenario was also conducted by comparing two different optimizers, namely Adam and SGD. The Adam optimizer was chosen due to the combined advantages of AdaGrad and RMSProp, which adaptively adjusts the learning rate for each parameter. In this experiment, the default hyperparameter for Adam is used, which is a learning rate of 0.001. On the other hand, the SGD optimizer was chosen because of its reliability in convergence on test scenarios with simpler data [30]

Finally, tests were conducted by applying the PSO (Particle Swarm Optimization) algorithm to optimize the model hyperparameters. The PSO algorithm was chosen due to its efficiency in global search space exploration, which is very useful in finding the optimal hyperparameter set [31]. In this study, the PSO parameters used include the number of particles of 200, the number of iterations of 500, and the inertia factor of 0.9.

By including the specific hyperparameters used in each test scenario as well as the reasons for their selection, this test design is expected to provide more comprehensive and reliable insights in evaluating model performance.

RESULT & DISCUSSION

A. Datasets

Table. 6. Hybrid Result

Username	Destination		
	Jalan Braga	...	Farm House
TripAdvisor	4.382321	...	2.068098
klook	3.390721	...	1.928820
holidify	3.878021	...	1.928820
...	...	...	...
Agoda	2.772581	...	2.989828

Source : (Research Results,2024)

As shown in Table 6, the following table is the result of combining the RoBERTa approach to tokenization with the TF-IDF technique to produce a vector representation of the text, which then generates a rating prediction from the user based on the user profile, the model has been trained, and produces normalized predictions that can produce recommendations that are tailored to the profile with the user.

MAE and RMSE values obtained based on the results of the rating prediction value obtained from hybrid TF-IDF and RoBERTa are MAE of 0.4208 and RMSE of 0.5733, the results obtained show that the error rate is very small.

Table 7. Classification Result

Username	Destination		
	Braga St	...	Farm House
aboutbdgcom	1	...	0
Suka_Bandung	0	...	0
...	...	...	...
bandung911	0	...	1
GEMAWAHYUH	0	...	1

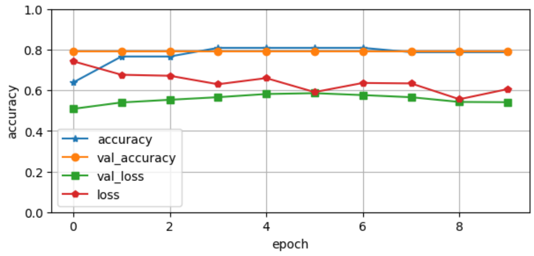
Source : (Research Results,2024)

Based on the results As shown in Table 6, the following e that will be processed to classification, numbers 1 and 0 which are binary values obtained from hybrid results that get a rating of 1-2 will be 0 and a rating of 3-4 will be 1.

A shown in Table 6, the value '1' is given to destinations that are recommended to users, and the value '0' is given to destinations that are not recommended to users.

B. Model With Adam Optimizer

The training process using Adam Optimizer was conducted using a learning rate of 0.001 and 10 iterations. The results of the training process in this test can be seen in Figure 3.



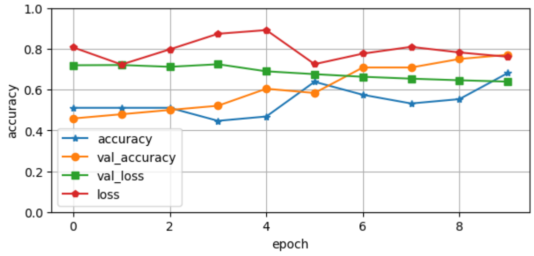
Source : (Research Results,2024)
Figure 3. Training Result Using Adam Optimizer

From Figure 3, it can be seen that the validation accuracy value has a stagnant trend at 74% and ends with a validation accuracy of 78% at the end of the iteration. The validation loss value tends to decrease from the middle of the training process to the end of the training process and produces a point of 0.5454.

After the testing process on the testing data, accuracy was obtained to 79.16%. The precision, recall and f1-score values for the '0' label are 79%, 100%, and 88%. Meanwhile, on the label '1', precision, recall, and f1-score values were found to be 0%.

C. Model With SGD Optimizer

SGD Optimizer training processes were conducted using a learning rate of 0.001 and 10 iterations. The results of the training process in this test can be seen in Figure 3.



Source : (Research Results,2024)
Figure 4. Training Result Using SGD Optimizer

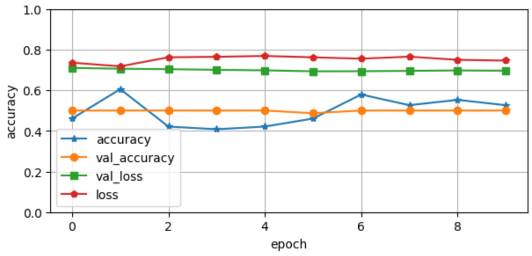
From Figure 4, it can be seen that the validation accuracy values tend to fluctuate, but the trend was increasing. The accuracy validation value in the last iteration was 70.83%. The validation loss tends to be stagnant. The final value obtained was 0.6707.

After the testing process on the testing data, an accuracy of 56.25% was obtained. The precision, recall and f1-score values for the '0' label are 79%, 89%, and 84%. In the label '1', precision, recall, and f1-score values were found to be 21%, 90%, and 35%.

D. Model With Adam Optimizer & SMOTE

The training process using SMOTE & Adam Optimizer was conducted using a learning rate of 0.001 and 10 iterations. The reason for implementing SMOTE as part of the research is to address the issue of data imbalance in the datasets used in this research and for comparison.

Because unbalanced datasets have a negative effect on the performance of machine learning models, especially in classification tasks, where the model may become biased towards the majority class, leading to poor performance in predicting the minority class. Using SMOTE, synthetic samples of the minority class are generated to balance the dataset, thus improving the model's ability to generalize and predict accurately across all classes. The results of the training process in this test as shown in Figure 5.



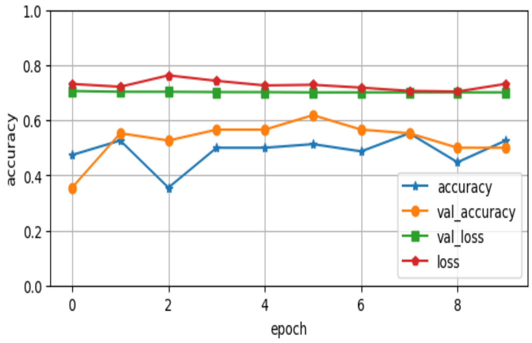
Source : (Research Results,2024)
Figure 5. Training Result Using SMOTE + Adam Optimizer

From Figure 5, it can be seen that the validation accuracy value was tends to stagnate. In the last iteration, validation accuracy was obtained at 57.89%. The validation loss also stagnated. The last loss value recorded was 0.7457.

After the testing process on the testing data, an accuracy of 27.63% was obtained. The precision, recall and f1-score values for the '0' label are 51%, 58%, and 54%. On the label '1', precision, recall, and f1-score values were found to be 51%, 92%, and 65%.

E. Model With SGD Optimizer & SMOTE

Training process using SMOTE & SGD Optimizer was carried out using a learning rate of 0.001 using 10 iterations. The results of the training process in this test can be seen in Figure 6.



Source : (Research Results,2024)
Figure 6. Training Result Using SMOTE + SGD Optimizer

From Figure 6, it can be seen that the validation accuracy value was tends to increase. The final accuracy validation value obtained was 64.47%. Meanwhile, the validation loss value tends to stagnate at 70%.

After the testing process on the testing data, an accuracy of 59.21% was obtained. The precision, recall and f1-score values for the '0' label are 73%, 42%, and 53%. Meanwhile, on

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.XXXX

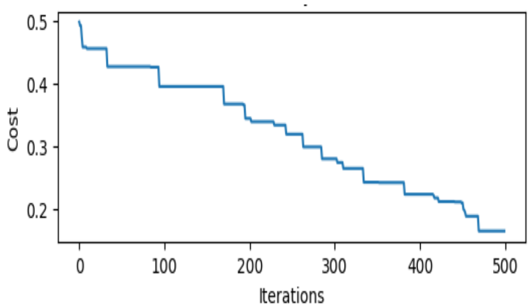
the label '1', precision, recall, and f1-score values were found to be 58%, 92%, and 71%.

Although our model shows strong performance, there are limitations that need to be noted. One of them is the model's bias towards the '0' label, which persists even after SMOTE has been applied. This is because class imbalance and overfitting can affect the generalization ability of the model. Other than that, the limitation of the dataset focusing on tourism in Bandung City may limit the applicability of the model to other contexts or domains. Future research needs to consider a wider variety of datasets to test the generalizability of the model.

F. PSO Optimization

Optimization with Particle Swarm Optimization (PSO) was conducted using 200 particles that move between the range 0 to 1. This range was adjusted to the output of the RNN model from the sigmoid activation function. The value of the load carried by the particles was set at 0.9. The number of samples is adjusted to the amount of data 'X' (independent variable). And optimization iterations were carried out 500 times.

This optimization succeeded in increasing the accuracy value from 79.16% in the model with the Adam optimizer to 94.37% using PSO. This result also impacts the cost of obtaining a recommendation decision, which becomes smaller with each iteration. The better the model, the smaller the cost required to draw conclusions, which impacts increasing the model's accuracy performance. The costs required for each iteration process show indications of decreasing, as shown in Figure 7.



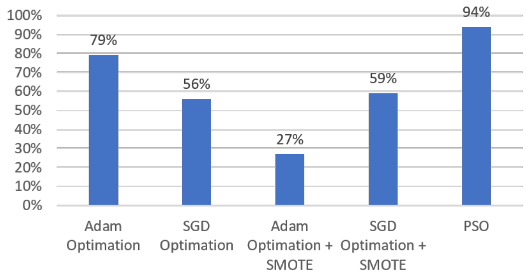
Source : (Research Results,2024)
Figure 7. Cost History of PSO

G. Relevance Test

In the relevance test, Roberta and TF-IDF obtained an MAE value of 0.4208 and an RMSE value of 0.5733. This value states the rating error level between recommended and non-recommended data. Because the error value was very small, it could be concluded that RoBERTa and TF-IDF can be used for research.

H. Analysis

A comparison table of performance results in each test scenario and PSO optimization can be seen in Figure 8.



Source : (Research Results,2024)
Figure 8. Comparison of Test Results

The data used in this research mostly has the label '0' (not recommended). So that the model has a tendency towards data that is not recommended. This results in the precision, recall and f1-score values tending to be high on the '0' label and very low on the '1' label. It can be seen in testing without using SMOTE. It can be actually overcome by using SMOTE. However, it actually resulted in a very significant decrease in accuracy in the Adam test and only resulted in an increase in accuracy of only 3% in the SGD test.

In Adam's test without SMOTE, the precision, recall, and f1-score results on the '0' label were all 100%. This means the model was only able to provide decisions on these labels. In another sense, not all tourist destinations are recommended. Even a reasonably high accuracy value can be considered invalid, considering that the percentage of data with the label '0' is 79%. In another sense, 79% is the percentage distribution of data on the '0' label.

Significant improvements occurred after using SMOTE, where the model could provide decisions both on label '0' and on label '1'. However, the decisions taken tend to be biased. This can be seen from the precision values on the labels '0' and '1' for Adam's test with SMOTE, which both have a value of 51%. This means that half of the data on each label is incorrectly recommended. As a result, the

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

model accuracy decreased drastically to reach 27%.

By using PSO optimization, the model experienced increased performance to reach 94.37% accuracy. The validity of this accuracy is proven by the cost function value, which continues to decrease with each accuracy.

CONCLUSION

This research has successfully produced a tourism recommendation system using a hybrid Content-Based Filtering (CBF) approach integrated with a Recurrent Neural Network (RNN) model that applies several optimizations such as Adam, Stochastic Gradient descent (SGD), and Particle Swarm Optimization (PSO). as well as implementing the RNN model for the classification process; this study aims to improve personalization in recommendations. The evaluation results show that RoBERTa's Hybrid TF-IDF model provides the best relevance with values of 0.4208 for MAE and 0.5733 for RMSE. In the application of classification using the RNN method and PSO optimization, the evaluation results using confusion matrix prove that the RNN-PSO model provides superior results compared to the use of ADAM and SGD , with accuracy reaching 94.37%. Therefore, the objectives of this study were successfully achieved by developing a hybrid CBF-based recommendation system for tourism in Bandung City, which showed significant performance improvement compared to traditional methods.

REFERENCE

[1] Badan Keahlian Sekretariat Jenderal DPR RI, "Analisis Ringkas Cepat: Urgensi Penguatan Daya Saing Pariwisata Untuk Meningkatkan Perekonomian Nasional," *Www.Bk.Dpr.Go.Id*, pp. 1–44, 2023, [Online]. Available: https://berkas.dpr.go.id/pa3kn/analisis-ringkas-cepat/public-file/analisis-ringkas-cepat-public-41.pdf.

[2] B. Utami, "SEKTOR PARIWISATA INDONESIA DI TENGAH PANDEMI COVID 19," *J. Din. Ekon. Pembang.*, vol. 4, 2021, doi: 10.33005/jdep.v4i1.198.

[3] I. M. Hasibuan, S. Mutthaqin, R. Erianto, and I. Harahap, "Kontribusi Sektor Pariwisata terhadap Perekonomian Nasional," *J. Masharif Al-Syariah J. Ekon. dan Perbank. Syariah*, vol. 8, no. 2, 2023, doi: 10.30651/jms.v8i2.19280.

[4] F. Ricci, "Recommender Systems in Tourism," in *Handbook of e-Tourism*, Z. Xiang, M. Fuchs, U. Gretzel, and W. Höpken, Eds. Cham: Springer International Publishing, 2022, pp. 457–474.

[5] N. Abhinav and K. Sujatha, "Content-based movie recommendation system using cosine similarity measure," *AIP Conf. Proc.*, vol. 2901, no. 1, p. 60035, 2023, doi: 10.1063/5.0178819.

[6] A. A. Salsabil, E. B. Setiawan, and I. Kurniawan, "Content-based Filtering Movie Recommender System Using Semantic Approach with Recurrent Neural Network Classification and SGD," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 9, no. 2, pp. 193–202, May 2024, doi: 10.22219/kinetik.v9i2.1940.

[7] I. M. Al Jawarneh *et al.*, "A Pre-Filtering Approach for Incorporating Contextual Information into Deep Learning Based Recommender Systems," *IEEE Access*, vol. 8, pp. 40485–40498, 2020, doi: 10.1109/ACCESS.2020.2975167.

[8] D. Magdaleno, M. Montes, B. Estrada, and A. Ochoa-Zezzatti, "A GPT-Based Approach for Sentiment Analysis and Bakery Rating Prediction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14502 LNAI, no. February, pp. 61–76, 2024, doi: 10.1007/978-3-031-51940-6_7.

[9] L. Li, Z. Zhang, and S. Zhang, "Hybrid Algorithm Based on Content and Collaborative Filtering in Recommendation System Optimization and Simulation," *Sci. Program.*, vol. 2021, pp. 1–11, 2021, doi: 10.1155/2021/7427409.

[10] L. Xiaoyan, R. C. Raga, and S. Xuemei, "GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews," *J. Sensors*, vol. 2022, 2022, doi: 10.1155/2022/7212366.

[11] P. R. Amalia and E. Winarko, "Aspect-Based Sentiment Analysis on Indonesian Restaurant Review Using a Combination of Convolutional Neural Network and Contextualized Word Embedding," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 3, p. 285, 2021, doi: 10.22146/ijccs.67306.

VOL. 10. NO. 1 AUGUST 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i1.XXXX

[12] N. Hossain, M. R. Bhuiyan, Z. N. Tumpa, and S. A. Hossain, "Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, no. July, 2020, doi: 10.1109/ICCCNT49239.2020.9225328.

[13] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation System from Microsoft News Data using TF-IDF and Cosine Similarity Methods," *Build. Informatics, Technol. Sci.*, vol. 4, no. 1, pp. 277–284, 2022, doi: 10.47065/bits.v4i1.1670.

[14] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms," *Proc. 2nd Int. Conf. Big Data, Model. Mach. Learn.*, 2021.

[15] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: 10.3390/app13063915.

[16] J. J. Jiang, W. X. Wei, W. L. Shao, Y. F. Liang, and Y. Y. Qu, "Research on Large-Scale Bi-Level Particle Swarm Optimization Algorithm," *IEEE Access*, vol. 9, pp. 56364–56375, 2021, doi: 10.1109/ACCESS.2021.3072199.

[17] Z. Cao, X. Han, W. Lyons, and F. O'Rourke, "Energy management optimisation using a combined Long Short-Term Memory recurrent neural network – Particle Swarm Optimisation model," *J. Clean. Prod.*, vol. 326, no. May, p. 129246, 2021, doi: 10.1016/j.jclepro.2021.129246.

[18] A. Javadian Sabet, M. Shekari, C. Guan, M. Rossi, F. Schreiber, and L. Tanca, "THOR: A Hybrid Recommender System for the Personalized Travel Experience," *Big Data Cogn. Comput.*, vol. 6, no. 4, 2022, doi: 10.3390/bdcc6040131.

[19] S. P. R. Asaithambi, R. Venkatraman, and S. Venkatraman, "A Thematic Travel Recommendation System Using an Augmented Big Data Analytical Model," *Technologies*, vol. 11, no. 1, 2023, doi: 10.3390/technologies11010028.

[20] J. H. Yoon and C. Choi, "Real-Time Context-Aware Recommendation System for Tourism," *Sensors*, vol. 23, no. 7, 2023, doi: 10.3390/s23073679.

[21] J. Lee, J. A. Shin, D. K. Chae, and S. C. Lee, "Personalized Tour Recommendation via Analyzing User Tastes for Travel Distance, Diversity and Popularity," *Electron.*, vol. 11, no. 7, pp. 1–14, 2022, doi: 10.3390/electronics11071120.

[22] M. Sigala, "Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ' s public news and information ," no. January, 2020.

[23] S. Babichev, "applied sciences Applying a Recurrent Neural Network-Based Deep Learning Model for Gene Expression Data Classification," 2023.

[24] H. Yang, C. Jiang, Y. Song, W. Fan, Z. Deng, and X. Bai, "TARGCN: temporal attention recurrent graph convolutional neural network for traffic prediction," *Complex Intell. Syst.*, 2024, doi: 10.1007/s40747-024-01601-1.

[25] M. Zitnik *et al.*, "Systems biology Current and future directions in network biology," 2024.

[26] S. Lun, Z. Sun, M. Li, and L. Wang, "Multiple-Reservoir Hierarchical Echo State Network," 2023.

[27] E. Chammas, "Fine-tuning Handwriting Recognition systems with Temporal Dropout."

[28] V. Shatravin, D. Shashev, and S. Shidlovskiy, "Sigmoid Activation Implementation for Neural Networks Hardware Accelerators Based on Reconfigurable Computing Environments for Low-Power Intelligent Systems," *Appl. Sci.*, vol. 12, no. 10, 2022, doi: 10.3390/app12105216.

[29] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: 10.3390/app13064006.

[30] E. Hassan, M. Y. Shams, N. A. Hikal, and S. Elmougy, *The effect of choosing optimizer algorithms to improve computer vision tasks: a comparative study*, vol. 82, no. 11. Multimedia Tools

and Applications, 2023.

[31] Y. Fan, Y. Zhang, B. Guo, X. Luo, and Q. Peng, "A Hybrid Sparrow Search Algorithm of the Hyperparameter Optimization in Deep Learning," 2022.