Question Answering Temanggung City Tourism using a Natural Language Processing Approach

Helga Raditia Ade Wijayanto
University of Technology Yogyakarta

Arief Hermawan
University of Technology Yogyakarta
Yogyakarta, Indonesia

ABSTRACT
Tourism is a valuable aspect for a region, and the more visitors it attracts, the more it can advance the walfare of the local community around tourist attractions. Temanggung regency has numerous tourist attractions, and the ease of searching and finding information is supported by the presence of search engines like Google, Yahoo, and others. However, not all information can be easly found, especially highly specific information, such as tourism related details. One way to address this issue is by leverging Natural Language Processing technology, particular a Question Answering System that enables computers to understand the intended meaning of user queries. This research develops a simple Question Answering System. The Preprocessing techniques employed include tokenization, lemmatization, and normalization. The outcome of the system is that it can answer questions inputted by the user if there are relevant keywords related to the desired tourism destinations.

Keywords
Tourism, Temanggung, Natural Language Processing, Question Answering System, Preprocessing

1. INTRODUCTION
The ease of searching and finding information is supported by the presence of search engines such as Google, Yahoo, and so on. Information on the internet can be accessed on various devices such as cellphones, laptops, computers and other devices that can be used.[2] However, not all information can be found easily and quickly, especially information about Temanggung city tourist attractions. Based on the many tourist attractions in Temanggung city, of course there are still many outsiders who don't know how many tourist attractions there are in Temanggung city. [5] Obstacles in obtaining incomplete information and providing information about tourism that is already well-known and not showing tourism that is still in the developing stage will certainly be an obstacle.[3] This means that tourism, which is still in its developing stage, is not well publicized.

Based on the problems above, the solution that can be offered is in the form of Question Answering (QA). So, with Temanggung city tourism question answering, people outside the city and local people can get information easily. [1]

The aim of this research is that, with question answering for Temanggung city tourism, people outside the city and local people can get information easily. And the benefits. So, with the existence of Temanggung city tourism question answering, people outside the city and local people can get information easily.

2. STUDY BACKGROUND
2.1 Question and Ansewering System
Question and Answering System (QAS) is a system that allows users to enter questions in natural language, namely the language used in everyday conversation and get answers quickly and concisely, or sometimes even accompanied by sentences that are sufficient to support the truth of the answer the.[11] The Question and Answer system is one solution to the many requests from users to obtain information quickly and accurately [9]

2.2 Natural Language Processing
Natural Language Processing (NLP) is a part of research and applications that examines how computers can be used to understand and manipulate natural language in the form of text or speech for useful purposes. [6] Text manipulation has been recognized as an important area of research in NLP. An NLP system that processes text starts with morphological analysis.[8] The text is converted, in a query or document, to obtain morphological variants of the words involved. Lexical and syntactic processing involves using a dictionary to determine the characteristics of words, recognize parts of speech, determine words and phrases, as well as for parsing sentences [4]

2.3 Preprocessing
Preprocessing is the process of changing the form of unstructured data into structured data according to needs. [10] There are several stages carried out in the preprocessing process, namely case folding, tokenizing, filtering, stemming, tagging and analyzing [7]. In this research, the preprocessing stages carried out were Case Folding, Tokenizing, Filtering and Stemming.

3. METHODOLOGY
3.1 Data
In this study, researchers used data taken from the mytrip123.com page, on this page there are 33 tourism data for the city of Temanggung which contain photos and explanations of each place. Examples of images from this page can be seen in the following image.

Fig 1: Posong Nature Tourism

3.2 Bussiness Rules

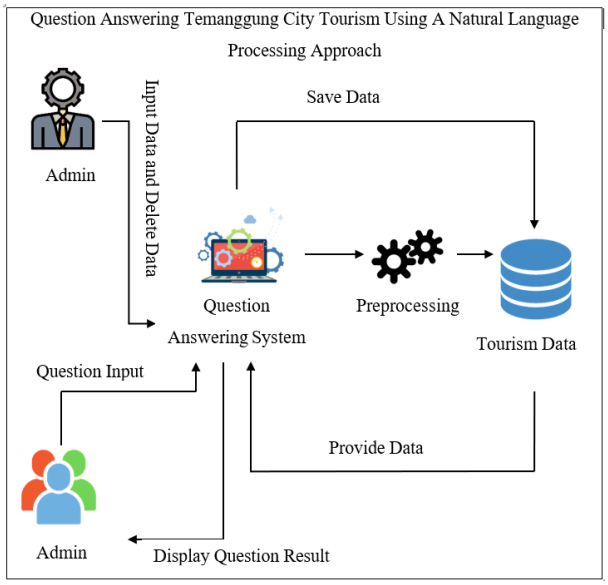

Fig 2: Bussiness Rules

As an illustration, flow data can be stored, processed, and displayed back to the user. This system can be used by local and out-of-town residents who want to know more about tourism in Temanggung city, so they can input questions which will then be processed by the system through a preprocessing process. From the preprocessing results, data will be taken from tourism data so that the system will display the results of questions input by the user. Admin can input and delete data on Temanggung tourism data.

3.3 Research Stages

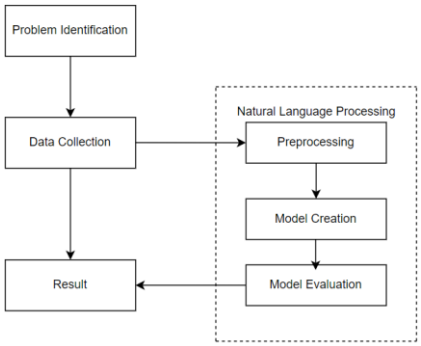

Fig 3: Research Stages

The image above is the research stages used in question answering for Temanggung city tourism using a natural language processing approach. The first stage is problem identification to find out what the problem is that is the basis for creating this system, the second is data collection to find information related to tourism in Temanggung city. After the data is collected, a preprocessing process is carried out to make a sequence of numbers in one of the tables as input for training the model, next is creating a model to be used as a parameter in the model training process, and model evaluation is used to determine the resulting accuracy. The final stage is the results to display what is produced by this system after carrying out the processes as described above.

4. SYSTEM ANALYSIS AND DESIGN

4.1 Requirement System

Analysis of functional requirements for the Temanggung City Tourism Question Answering system using the Natural Language Processing approach includes the features and capabilities of the system. Some examples of functional advantages that this system can have are question processing, information search, natural language processing and providing answers.

Meanwhile, non-functional analysis includes the equipment (hardware and software) used to run this system. The following are examples of equipment that can be used, such as computers/laptops, the Python programming language, Scikit-learn, and datasets.

4.2 Usecase Diagram

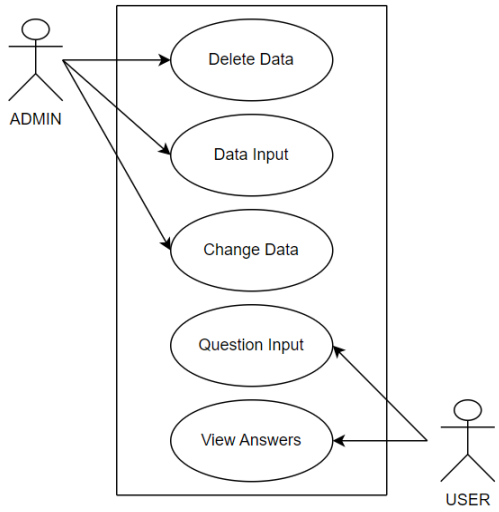

Fig 4: Usecase Diagram

Use Case Diagram describes the expected functionality of a system.

Table 1. Usecase Diagram description

No	Actor	Descripton
1	Admin	Admin is an actor who has full access rights to data. Admin can add, change and delete data.
2	User	Users are actors who only have access rights to ask questions

4.3 Activity Diagram

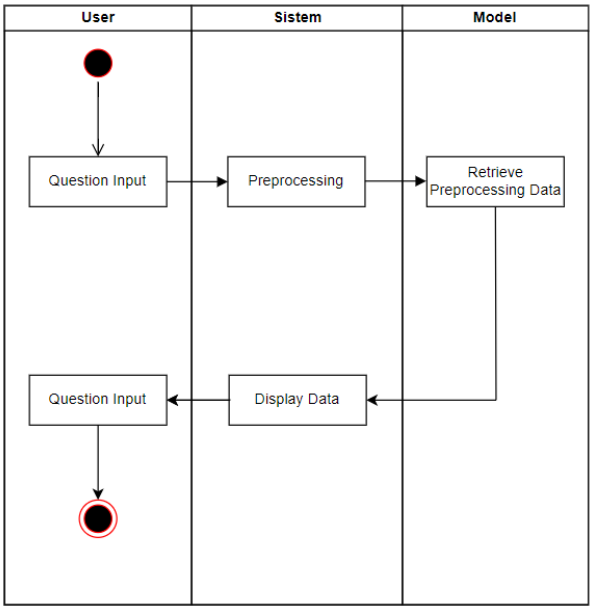

Fig 5: Activity DIagram

The image shows the activities carried out by the user on the system. The first step is that the user sends a question to the system, then the system will carry out preprocessing, the preprocessing results will continue to display the results of the question searched by the user.

4.4 Entitiy Relationship Diagram

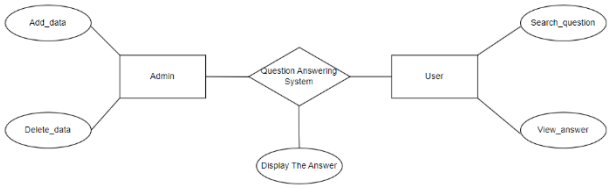

Fig 6: Entity Relationship DIagram

The entity relationship diagram used in creating this system contains 2 entities, namely admin and user. Admin is the object that manages this system while user is the person who uses this system. The attributes for the admin are add_data and delete_data, while the attributes for the user are search_questions and see_answers.

5. IMPLEMENTATION

5.1 Implementation

At the implementation stage, the web page interface was designed using the Streamlit framework. This interface design includes elements such as the page title, question input, "submit" button and display of the resulting answers. This interface design is designed to be easy for users to understand and use. The following is some of the source code used to build this Question Answering system.

To read the dataset used in this system, use the source which can be seen in the following figure. [15]

```
# Baca dataset
@st.cache_data()
def read_dataset():
    dataset=
pd.read_csv('C:\\Users\\Lenovo\\Downloads\\Streamlit2\\DATA_TMG.
csv')
    return dataset
```

Fig 7: Read dataset

To read the dataset used in this system, use the source which can be seen in Figure.

The preprocessing used to create the Question Answering system website display can be seen in the following figure.

```
# Preprocessing
@st.cache_data()
def preprocess_data(dataset):
    questions = dataset['pertanyaan'].values
    answers = dataset['jawaban'].values

    tokenizer = Tokenizer()
    tokenizer.fit_on_texts(questions)
    total_words = len(tokenizer.word_index) + 1

    question_sequences = tokenizer.texts_to_sequences(questions)
    padded_questions = pad_sequences(question_sequences)
```

Fig 8: Preprocessing model

The prerocessing used to create the Question Answering system website display can be seen in the figure.[12]

Modeling for this Question Answering system uses a simple model as can be seen in the following figure.

```
# Buat model sederhana
Def              create_model(total_words,num_classes,input_length,
activation):
    model = Sequential()
    model.add(Embedding(total_words,100,
input_length=input_length))
    model.add(LSTM(50))
    model.add(Dense(num_classes, activation=activation))
    return model
```

Fig 9: Model creation

Model creation is done by adding an embedding layer to the model. The embedding layer aims to convert a sequence of numbers into a vector with lower dimensions. 'total_words' is the number of unique words in the dataset.

Then, the model training process can be seen in the following figure.

```
# Training model
def train_model(model, X_train, y_train_encoded, epoch, batch_size,
verbose, optimizer):
    model.compile(loss='categorical_crossentropy',
optimizer=optimizer, metrics=['accuracy'])
    model.fit(X_train,        y_train_encoded,        epochs=epoch,
batch_size=batch_size, verbose=verbose)
    return model
```

Fig 10: The model training process

The train model function is used to compile the model with the specified loss and optimizer functions, train the model with training data, and return the trained model.[13]

Model evaluation is used to evaluate models using test data. The source code can be seen in the following image.

```
# Evaluasi model
def evaluate_model(model, X_test, y_test_encoded):
    loss, accuracy = model.evaluate(X_test, y_test_encoded)
    return loss, accuracy
```

Fig 11: Model evaluation

The train model above produces loss values and model accuracy which is used using the return loss, accuracy function.

Generating Answer or returning an answer is the final stage of this system after the previous process has been carried out. [16]

The source code to produce this answer can be seen in the following image.

```
# Fungsi untuk menghasilkan jawaban berdasarkan pertanyaan
def generate_answer(question, model, tokenizer, label_encoder,
X_train):
    question_sequence = tokenizer.texts_to_sequences([question])
    padded_question = pad_sequences(question_sequence,
maxlen=X_train.shape[1])
    predicted_index = np.argmax(model.predict(padded_question),
axis=-1)[0]
    predicted_label =
label_encoder.inverse_transform([predicted_index])[0]
    return predicted_label
```

Fig 12: Generating Answer

The generate answer function is to receive the results of all processes that have been carried out previously, including the model, tokenizer, label encoder, and training data. The tokenizer is used to convert questions into a sequence of tokens and the token sequence is used as padding using training data. The final result is to predict the corresponding answer index using the trained model and return the predicted answer label.

5.2 Result

This research produces a Question Answering System for Temanggung City Tourism Using a Natural Language Processing Approach. With this system, it is hoped that we will be able to find information about tourism in the city of Temanggung, which will really help potential visitors to find information about the tourism they want to visit. This system was created using the Python language and uses libraries such as numpy, pandas, nltk, and sklearn. To create the appearance of the website using the Streamlit framework. The final result of the system created can be seen in following figure.

Fig 13: System Result

The image above is the result of the Temanggung Tourism Question Answering system. Users must input questions in the column provided and the questions entered must contain words from the tourist spot they want to search for, otherwise the system will not recognize the question and will not answer the question the user is looking for.

5.3 Discussion And Testing

Testing this application uses several questions and checks the accuracy of the answers. This test was carried out using 4 sample questions and using different numbers of epochs, optimizers, batch sizes and types of activation.

Answers from the test results will be divided into 2 types, namely correct answers and incorrect answers. The correct answer is the output issued by the system in accordance with the question input by the user, while the incorrect answer is the answer issued by the system that does not match the question input by the user.

Test 1 was carried out using epoch 50, optimizer adam, batch size 8, verbose 1, and softmax activation type. The test results can be seen in the following table,

Table 2. Testing 1

Question	Correct answer	Wrong answer
What tourist attractions are there in Temanggung?	✓	
Where is Bejen Fruit Garden located?	✓	
What's interesting about Jumprit tourism?	✓	
How much is the ticket price to enter the Pikatan Water Park tourist attraction?	✓	

Test 2 uses a number of epochs of 75, optimizer SGD, batch size 16, verbose 1, and the type of activation used is relu. The table below shows the results of the tests that have been carried out.

Table 3. Testing 2

Question	Correct answer	Wrong answer
What tourist attractions are there in Temanggung?		✓
Where is Bejen Fruit Garden located?	✓	
What's interesting about Jumprit tourism?		✓
How much is the ticket price to enter the Pikatan Water Park tourist attraction?	✓	

Test 2 uses a number of epochs of 100, optimizer rmsprop, batch size 32, verbose 1, and sigmoid activation type. The test results can be seen in the following table.

Table 4. Testing 3

Question	Correct answer	Wrong answer
What tourist attractions are there in Temanggung?		✓
Where is Bejen Fruit Garden located?	✓	
What's interesting about Jumprit tourism?	✓	
How much is the ticket price to enter the Pikatan Water Park tourist attraction?	✓	

The results of the 3 tests above are not all the answers displayed by this system can be answered correctly because the system can only answer very specific questions about what the user is asking. Other things that influence the accuracy of the answers given are the number of epochs, optimizer, batch size, and verbose.

6. CONCLUSION

Based on the results of implementation and testing of this question answering system, it can be concluded that the question answering system can provide information to users about tourism in the city of Temanggung. To simplify the search, users must write the keywords of the tourism they are looking for. In this question answering system, there is still an error in the output displayed if the user does not enter the keywords of tourism sought. So that further development needs to be done in the future so that this system is able to answer questions appropriately and can add new features to the display the searched tourism images.

Meanwhile, the suggestion from this research is to add a dataset so that the system provides more information to the user. And Implementing more advanced NLP techniques, such as the use of transformer-based language models (e.g. BERT, GPT) or multilevel approaches, to improve understanding of context and nuances in user questions

7. ACKNOWLEDGMENTS

We would like to thank our universities for their contribution and support of our research.

8. REFERENCES

[1] Adams, B. T. (2019). Information Retrieval in Tourism Using NLP. Journal of Tourism and Hospitality Research, 32(3), 256-269..

[2] Anderson, M. J. (2019). Challenges and Opportunities in Question Answering for Tourism. Journal of Computational Linguistics, 36(3), 213-227.

[3] Brown, C. D. (2020). Applications of NLP in Tourism Information Retrieval. International Journal of Computational Linguistics, 15(2), 89-104.

[4] Carter, J. H. (2020). Evaluation Metrics for NLP-based Tourism Question Answering. Journal of Natural Language Understanding, 7(1), 45-58.

[5] Clark, F. G. (2020). NLP Techniques for Extracting Tourism Information. International Journal of Tourism Studies, 28(4), 311-324.

[6] Davis, H. S. (2019). Natural Language Processing for City Tourism. Journal of Tourism and Hospitality, 21(2), 156-170.

[7] Harris, R. M. (2020). NLP Approaches for Tourism Query Answering. Journal of Language Processing and Linguistics, 14(4), 321-335.

[8] Johnson, A. B. (2020). A Survey of Question Answering Systems. Journal of Information Science, 52(4), 567-582

[9] Martin, P. S. (2020). A Survey of NLP Techniques for Question Answering. International Journal of Computational Tourism, 12(3), 187-200.

[10] Smith, A. R. (2019). Language Models for Tourism Information Retrieval. International Journal of Artificial Intelligence, 42(2), 89-103.

[11] Smith, J. (2019). Advances in Natural Language Processing for Tourism. Journal of Tourism Research, 45(3), 123-136.

[12] Taylor, K. L. (2019). Text Mining for Tourism Information. Journal of Information Retrieval, 27(1), 56-70.

[13] Turner, D. C. (2019). Question Answering Systems in the Tourism Domain. Journal of Computational Tourism, 16(4), 278-292.

[14] Walker, L. A. (2020). Natural Language Processing for Tourism Knowledge Graphs. International Journal of Knowledge Engineering, 8(2), 120-134.

[15] White, L. P. (2020). Semantic Question Answering in Tourism. Journal of Language Technology, 10(1), 34-48.

[16] Wilson, E. R. (2019). An Overview of Question Answering in the Tourism Domain. Journal of Language Processing, 34(1), 45-58