Introduction

In the project, we implemented a novel-based chatbot, collecting 15 novels from the Gutenberg Website. The first step included extracting the text data from the novels and placing them into a format on which we trained our system on the novel's data. There were many different methods to implement and phrase the text data such that it would be easy for the chatbot as well as the user to read.

We have designed the UI through which users will interact with the chatbot. The chatbot understands based on the query provided by the user, whether it is novel-related or just a chit-chat model. On that basis, the chatbot interacts with the user. There are some features added to the chatbot project that provide a better understanding of how well the chatbot is performing while responding to queries. We have added visualization graphs, where we included response time as one of the main factors to evaluate the chatbot also to support how well the chatbot is we have documents retrieved per query graphs which describe in detail what all documents were considered in the backend to respond. Another visualization graph related to the distribution of queries has been added in that part, which specifically focuses on chitchat vs novel distribution, where when the user provides a query, the chatbot identifies whether the user wants to chitchat or requires information related to the novels.

In the project, we have also introduced hugging face APIs and OpenAI APIs which help to refine the responses of the chatbot. The chatbot uses a total of 15 novels, namely “The Adventures of Sherlock Holmes”,” The Art of War”, “The Time Machine”,” Common Sense”, “The Further Adventures of Zorro”,” The Wolf-Men: A Tale of Amazing Adventure in the Under_World, “From Zone to Zone”, “All the world Over Interesting stories”, “ The Lonely house”, “The Erotic Movie in Literature”, “Personal hygiene and physical training for women”, “Tokology”,” Home Education”, “Outside Saturn 4”, and “The Red Planet”.

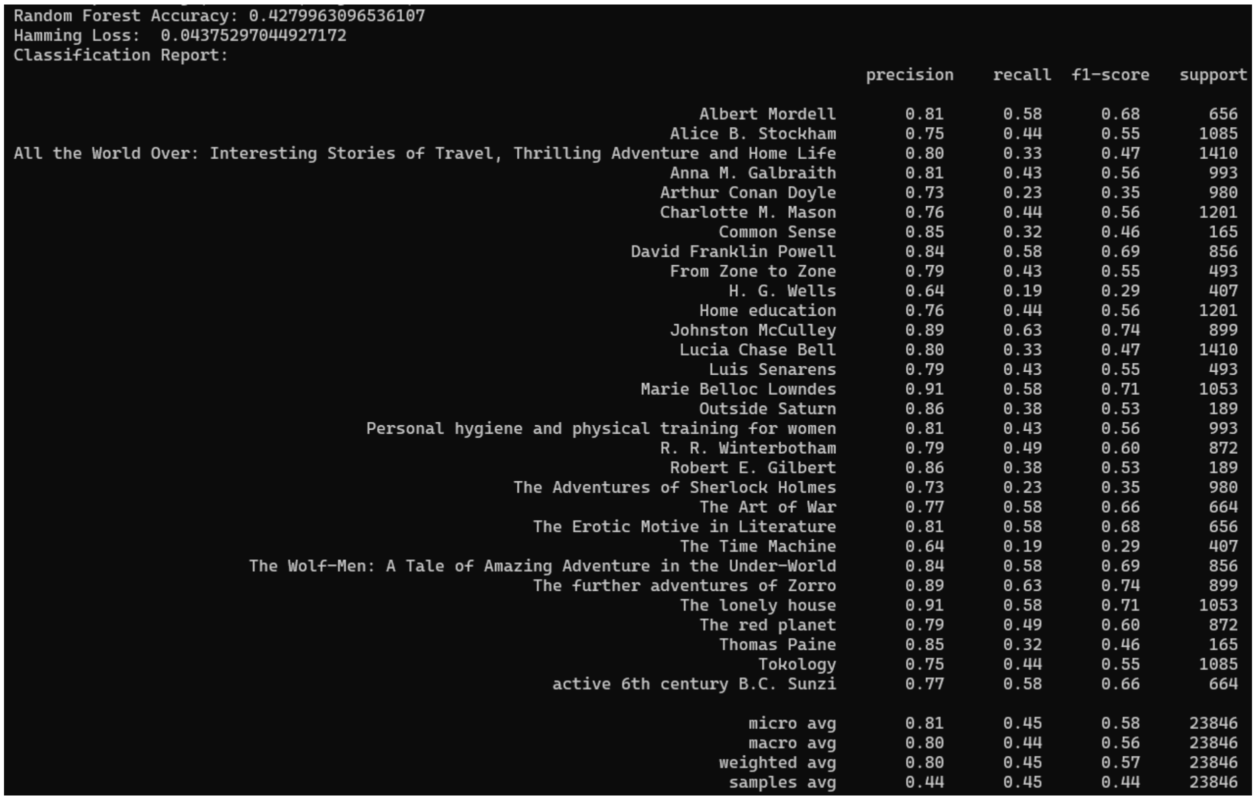
Methodology

As we have introduced the parts of the project, now let's dive deep into how everything is implemented. The whole process is divided into backend and frontend parts. At first, the system extracts the text data from novels and stores the data in a data frame. The system also extracts metadata from the novels such as title, author, and language. The data frame consisted of 5 columns, title, author, language, original content, and preprocessed content. The system is not using that preprocessed content further as we wanted the same preprocessing method for the query and the content, hence it is dropping this column later. We saved that file into csv file and accessed it for later use.
The next step included training a classifier, which will predict the title and author from the query provided by the user. The classifier we have used to train the data is the Random Forest classifier. We used term frequency and inverse document frequency vectorizer is used for converting text data into a matrix of TF-IDF features. The target variables (title and author) are combined into a multi-label format. MultiLabelBinarizer is used to encode these multi-labels into a binary format suitable for the classification model. The system preprocessed the content to train the classifier and stored it in different columns. For preprocessing, it uses lemmatization, removes stop words, and special characters, makes everything lowercase, and then splits the sentences into tokens of words. The reason for using lematization as it reduces words to their base or root form. A OneVsRestClassifier with RandomForestClassifier as the base classifier is initialized. The model's performance is evaluated using accuracy, hamming loss, and classification reports. Overall, the precision accuracy is good for the model. The whole purpose of the classifier is to predict the title and author from the query provided by the user. So further models can retrieve the data from novels by looking into the content. The system uses zero-shot classification API to determine if the query is "Chit-chat" or "Novel-related". For the chit-chat purpose, it is using Blenderbot API. If the query is "Novel-related", the system retrieves documents based on the user's query. It also checks whether the user has selected any specific topic if the user has selected then it checks into the selected topics.
For the UI part, we have developed a web application. In that, there is an option for users to select topics from which they want their response to be. The process of retrieving documents includes filtering the data according to topics provided or if there are no topics provided by the user then it predicts using the classifier. Creates an inverted index from the TF-IDF matrix. An inverted index maps each word to the documents that contain it. This method also uses document IDs to reference documents. It sorts the documents in descending order of the criteria. Calculates a score for a document based on the frequency of query terms in that document. Retrieves documents relevant to a novel's predicted topic or author. It filters documents based on an inverted index and further refines the search based on the topic or author. It then calculates cosine similarities between the query and the documents' TF-IDF vectors, combines these with

content scores, and selects the top k documents based on a combined score. Then we pass the top k documents to refine the response using openAI APIs.
The Retrieval Augmentation Generation(RAG) process includes Generating an answer to a given query based on the provided context. which uses OpenAI's GPT-3.5 API to generate answers to user queries based on a given context.
Then in the system, there are conditions such as the maximum response length of 150, and stopping conditions. The request is made to the OpenAI API endpoint, and the method checks for a successful response and returns a coherent response. This is the whole backend process we have used to respond to a query provided by the user. We integrated the backend with a UI where users can interact with the chatbot. The frontend web application consists of HTML, CSS, and JavaScript for our chatbot UI. The design of UI is user-friendly, it has a query container where the user will provide a query to the chatbot, a topic container that consists of a list of all the topics where the user can select if they want some specific information from the novel. The web application consists of a visualization button that will redirect to the visualization page where graphs are plotted, it depicts how well the chatbot has performed. We also provided a terminate chat option to the user, where when the user wishes to end the session, the chatbot terminates it. The user will be asked for confirmation before actually ending the session, after that user may no longer be able to chat with the chatbot. If the user wants to use it again, then need to refresh the page.
The system is also handling the error efficiently, A custom exception class, NoLabelsException, is defined to handle scenarios where the machine learning classifier does not return any labels. In addition, Blenderbot and Zero-shot classifier, these functions include error handling for HTTP response status codes. If the response status code is not 200 (indicating a successful request), an exception is raised, and an error message is generated detailing the status code and the response text. For any other exceptions, a generic error message is returned, and the exception details are logged.

Sample screenshots

Classifier:

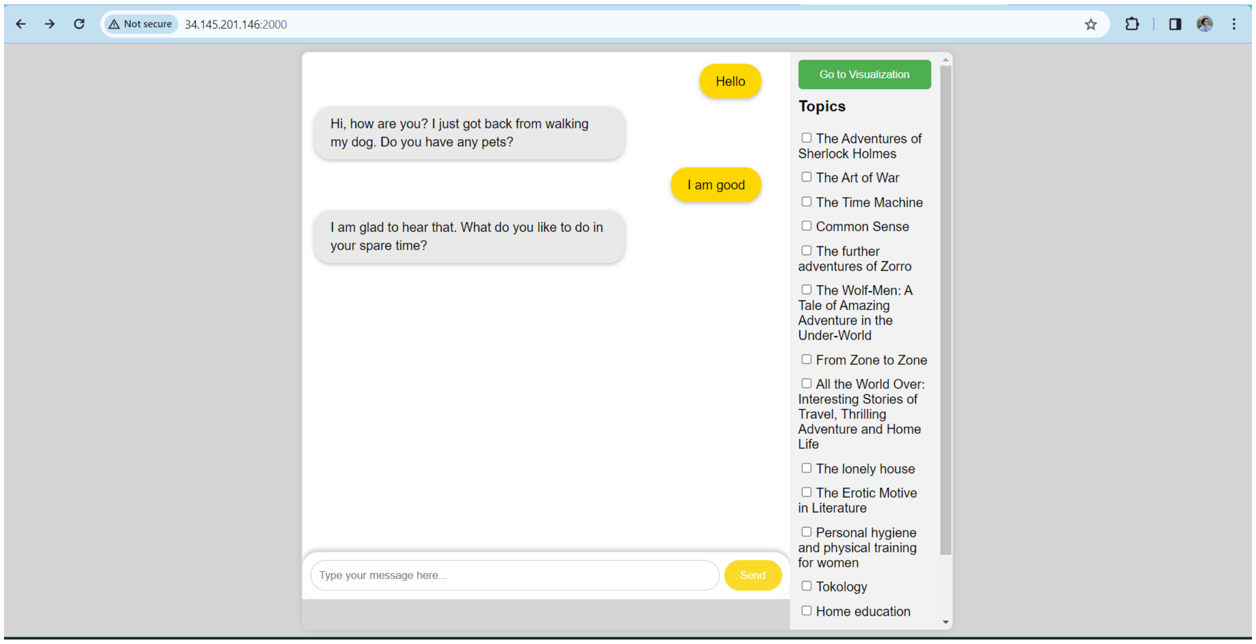


```
Random Forest Accuracy: 0.4279963096536107
Hamming Loss:  0.04375297044927172
Classification Report:
                                                                                    precision    recall  f1-score   support

                                                                   Albert Mordell       0.81      0.58      0.68       656
                                                                Alice B. Stockham       0.75      0.44      0.55      1085
All the World Over: Interesting Stories of Travel, Thrilling Adventure and Home Life       0.80      0.33      0.47      1410
                                                               Anna M. Galbraith       0.81      0.43      0.56       993
                                                              Arthur Conan Doyle       0.73      0.23      0.35       980
                                                              Charlotte M. Mason       0.76      0.44      0.56      1201
                                                                    Common Sense       0.85      0.32      0.46       165
                                                           David Franklin Powell       0.84      0.58      0.69       856
                                                               From Zone to Zone       0.79      0.43      0.55       493
                                                                    H. G. Wells       0.64      0.19      0.29       407
                                                                  Home education       0.76      0.44      0.56      1201
                                                               Johnston McCulley       0.89      0.63      0.74       899
                                                                Lucia Chase Bell       0.80      0.33      0.47      1410
                                                                  Luis Senarens       0.79      0.43      0.55       493
                                                            Marie Belloc Lowndes       0.91      0.58      0.71      1053
                                                                  Outside Saturn       0.86      0.38      0.53       189
                                       Personal hygiene and physical training for women       0.81      0.43      0.56       993
                                                             R. R. Winterbotham       0.79      0.49      0.60       872
                                                               Robert E. Gilbert       0.86      0.38      0.53       189
                                                 The Adventures of Sherlock Holmes       0.73      0.23      0.35       980
                                                                  The Art of War       0.77      0.58      0.66       664
                                                    The Erotic Motive in Literature       0.81      0.58      0.68       656
                                                                The Time Machine       0.64      0.19      0.29       407
                         The Wolf-Men: A Tale of Amazing Adventure in the Under-World       0.84      0.58      0.69       856
                                                   The further adventures of Zorro       0.89      0.63      0.74       899
                                                                The lonely house       0.91      0.58      0.71      1053
                                                                  The red planet       0.79      0.49      0.60       872
                                                                    Thomas Paine       0.85      0.32      0.46       165
                                                                        Tokology       0.75      0.44      0.55      1085
                                                    active 6th century B.C. Sunzi       0.77      0.58      0.66       664

                                                                       micro avg       0.81      0.45      0.58     23846
                                                                       macro avg       0.80      0.44      0.56     23846
                                                                    weighted avg       0.80      0.45      0.57     23846
                                                                     samples avg       0.44      0.45      0.44     23846
```

Chitchat:



Hello

Hi, how are you? I just got back from walking my dog. Do you have any pets?

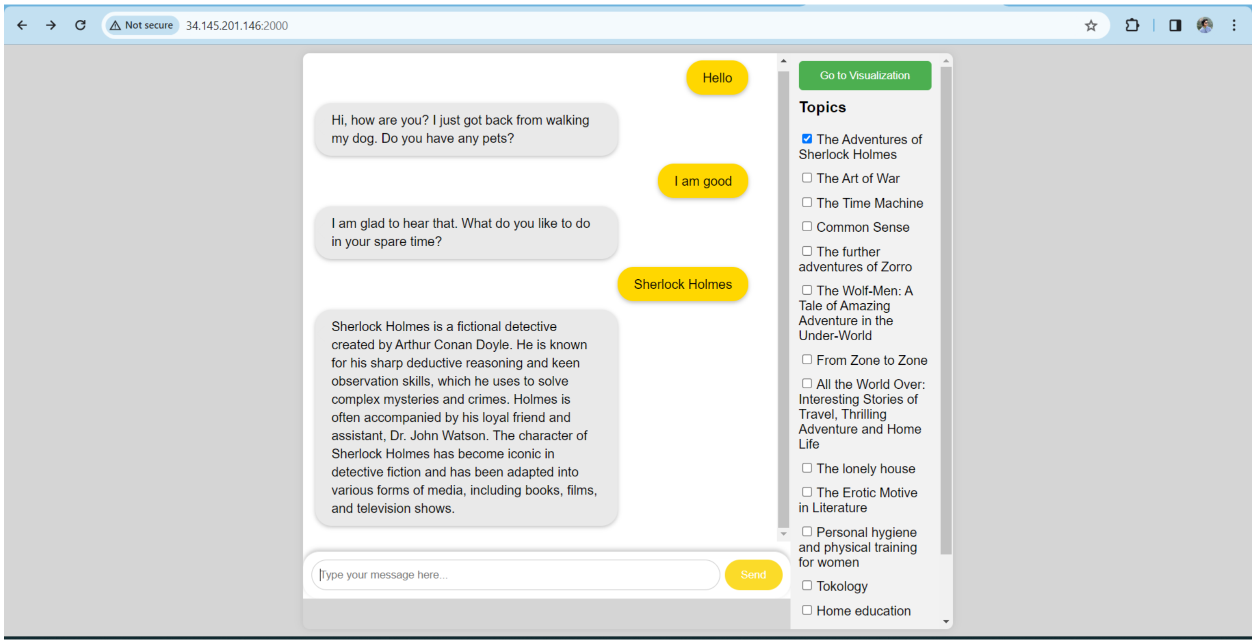
I am good

I am glad to hear that. What do you like to do in your spare time?

Topics

- ☐ The Adventures of Sherlock Holmes
- ☐ The Art of War
- ☐ The Time Machine
- ☐ Common Sense
- ☐ The further adventures of Zorro
- ☐ The Wolf-Men: A Tale of Amazing Adventure in the Under-World
- ☐ From Zone to Zone
- ☐ All the World Over: Interesting Stories of Travel, Thrilling Adventure and Home Life
- ☐ The lonely house
- ☐ The Erotic Motive in Literature
- ☐ Personal hygiene and physical training for women
- ☐ Tokology
- ☐ Home education

Novel Query:



Not secure 34.145.201.146:2000

Hello

Hi, how are you? I just got back from walking my dog. Do you have any pets?

I am good

I am glad to hear that. What do you like to do in your spare time?

Sherlock Holmes

Sherlock Holmes is a fictional detective created by Arthur Conan Doyle. He is known for his sharp deductive reasoning and keen observation skills, which he uses to solve complex mysteries and crimes. Holmes is often accompanied by his loyal friend and assistant, Dr. John Watson. The character of Sherlock Holmes has become iconic in detective fiction and has been adapted into various forms of media, including books, films, and television shows.

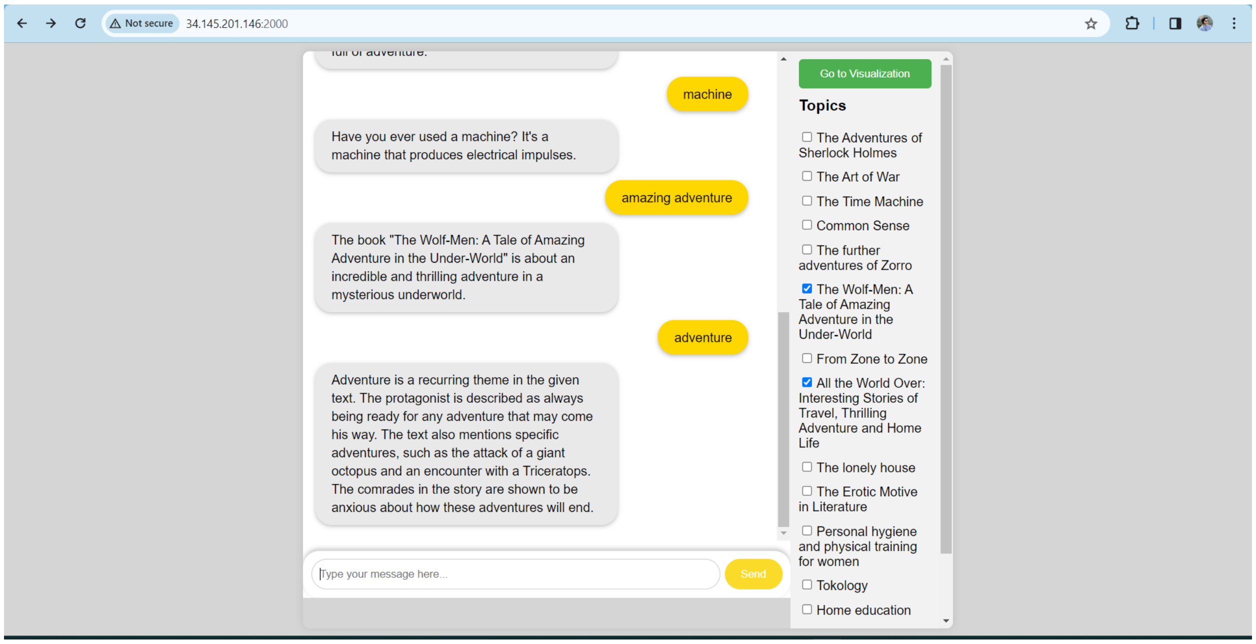
Type your message here... Send

Go to Visualization

Topics

- [✓] The Adventures of Sherlock Holmes
- [ ] The Art of War
- [ ] The Time Machine
- [ ] Common Sense
- [ ] The further adventures of Zorro
- [ ] The Wolf-Men: A Tale of Amazing Adventure in the Under-World
- [ ] From Zone to Zone
- [ ] All the World Over: Interesting Stories of Travel, Thrilling Adventure and Home Life
- [ ] The lonely house
- [ ] The Erotic Motive in Literature
- [ ] Personal hygiene and physical training for women
- [ ] Tokology
- [ ] Home education

Multi Novel:



Not secure 34.145.201.146:2000

full of adventure.

machine

Have you ever used a machine? It's a machine that produces electrical impulses.

amazing adventure

The book "The Wolf-Men: A Tale of Amazing Adventure in the Under-World" is about an incredible and thrilling adventure in a mysterious underworld.

adventure

Adventure is a recurring theme in the given text. The protagonist is described as always being ready for any adventure that may come his way. The text also mentions specific adventures, such as the attack of a giant octopus and an encounter with a Triceratops. The comrades in the story are shown to be anxious about how these adventures will end.

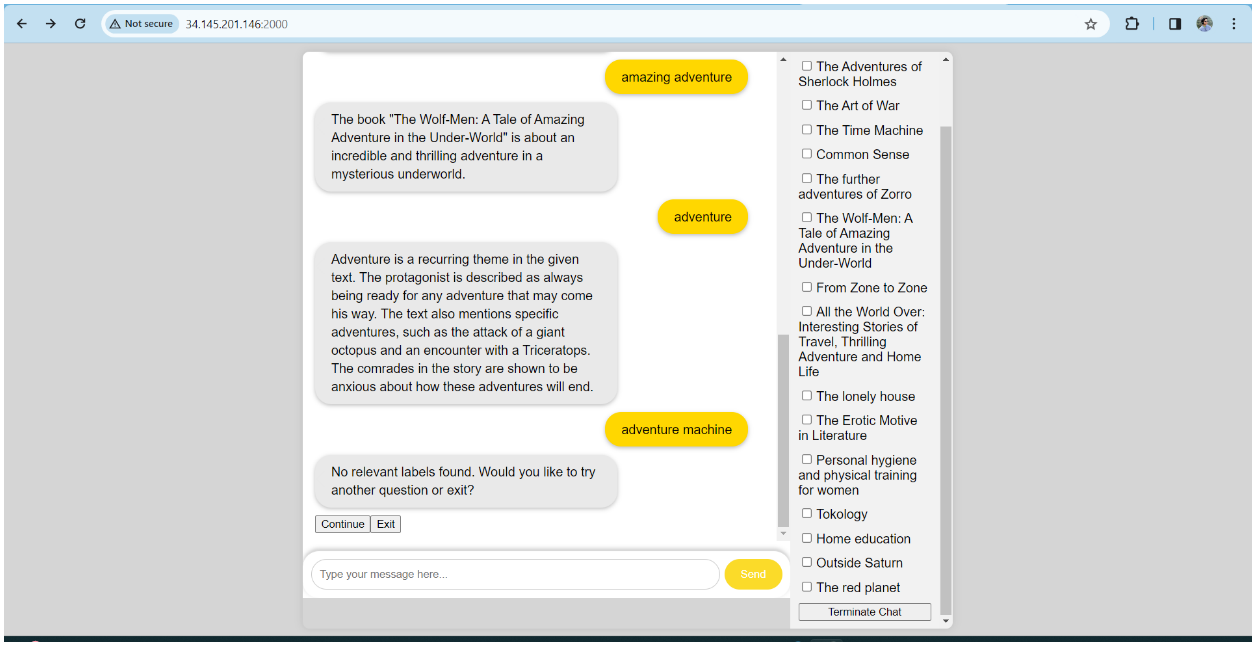
Type your message here... Send

Go to Visualization

Topics

- [ ] The Adventures of Sherlock Holmes
- [ ] The Art of War
- [ ] The Time Machine
- [ ] Common Sense
- [ ] The further adventures of Zorro
- [✓] The Wolf-Men: A Tale of Amazing Adventure in the Under-World
- [ ] From Zone to Zone
- [✓] All the World Over: Interesting Stories of Travel, Thrilling Adventure and Home Life
- [ ] The lonely house
- [ ] The Erotic Motive in Literature
- [ ] Personal hygiene and physical training for women
- [ ] Tokology
- [ ] Home education

Error:



Not secure 34.145.201.146:2000

amazing adventure

The book "The Wolf-Men: A Tale of Amazing Adventure in the Under-World" is about an incredible and thrilling adventure in a mysterious underworld.

adventure

Adventure is a recurring theme in the given text. The protagonist is described as always being ready for any adventure that may come his way. The text also mentions specific adventures, such as the attack of a giant octopus and an encounter with a Triceratops. The comrades in the story are shown to be anxious about how these adventures will end.

adventure machine

No relevant labels found. Would you like to try another question or exit?

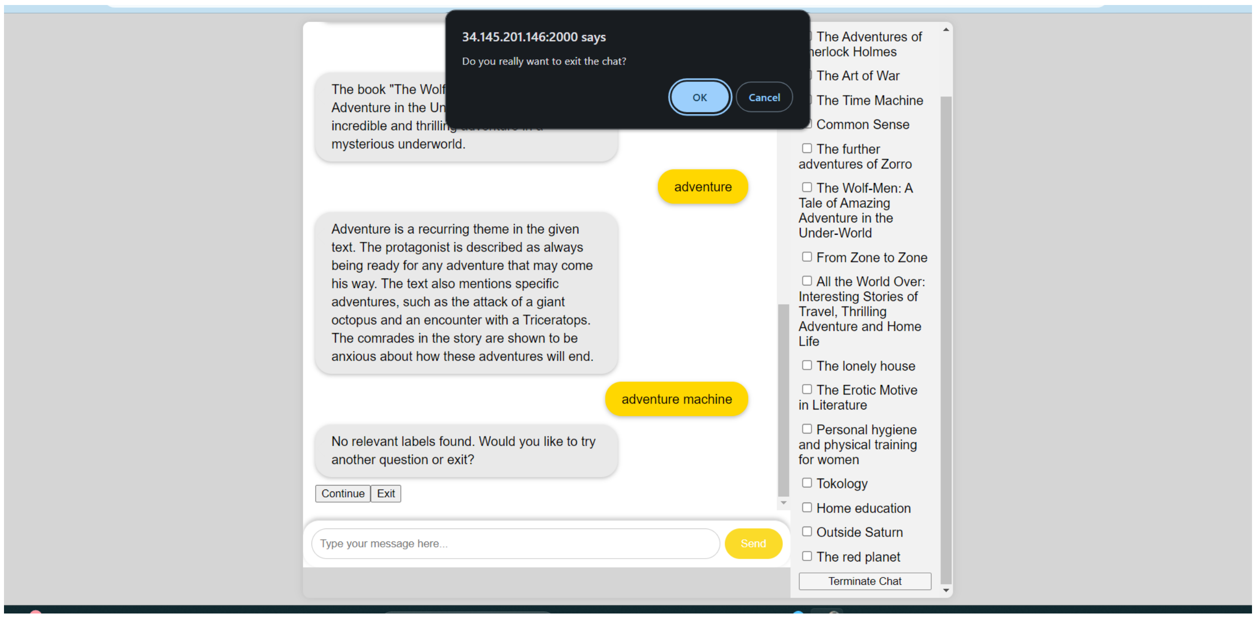
Continue | Exit

Type your message here... | Send

- ☐ The Adventures of Sherlock Holmes
- ☐ The Art of War
- ☐ The Time Machine
- ☐ Common Sense
- ☐ The further adventures of Zorro
- ☐ The Wolf-Men: A Tale of Amazing Adventure in the Under-World
- ☐ From Zone to Zone
- ☐ All the World Over: Interesting Stories of Travel, Thrilling Adventure and Home Life
- ☐ The lonely house
- ☐ The Erotic Motive in Literature
- ☐ Personal hygiene and physical training for women
- ☐ Tokology
- ☐ Home education
- ☐ Outside Saturn
- ☐ The red planet

Terminate Chat

After clicking on Terminate chat:



34.145.201.146:2000 says

Do you really want to exit the chat?

OK | Cancel

The book "The Wolf... Adventure in the Un... incredible and thrilling... mysterious underworld.

adventure

Adventure is a recurring theme in the given text. The protagonist is described as always being ready for any adventure that may come his way. The text also mentions specific adventures, such as the attack of a giant octopus and an encounter with a Triceratops. The comrades in the story are shown to be anxious about how these adventures will end.

adventure machine

No relevant labels found. Would you like to try another question or exit?

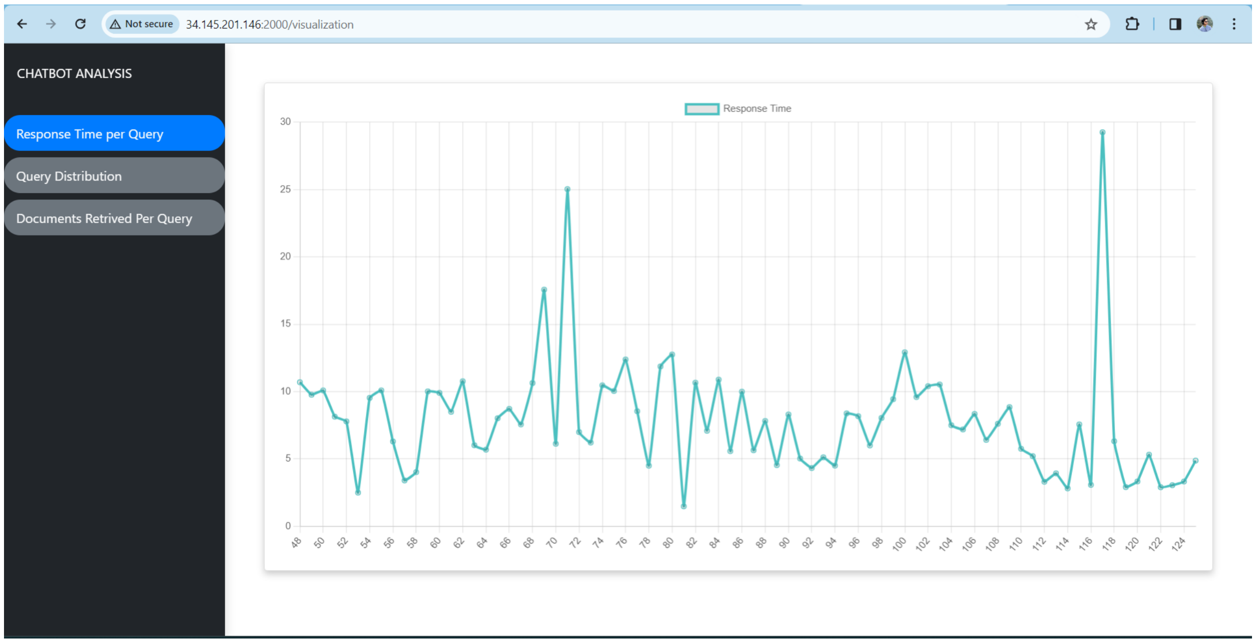
Continue | Exit

Type your message here... | Send

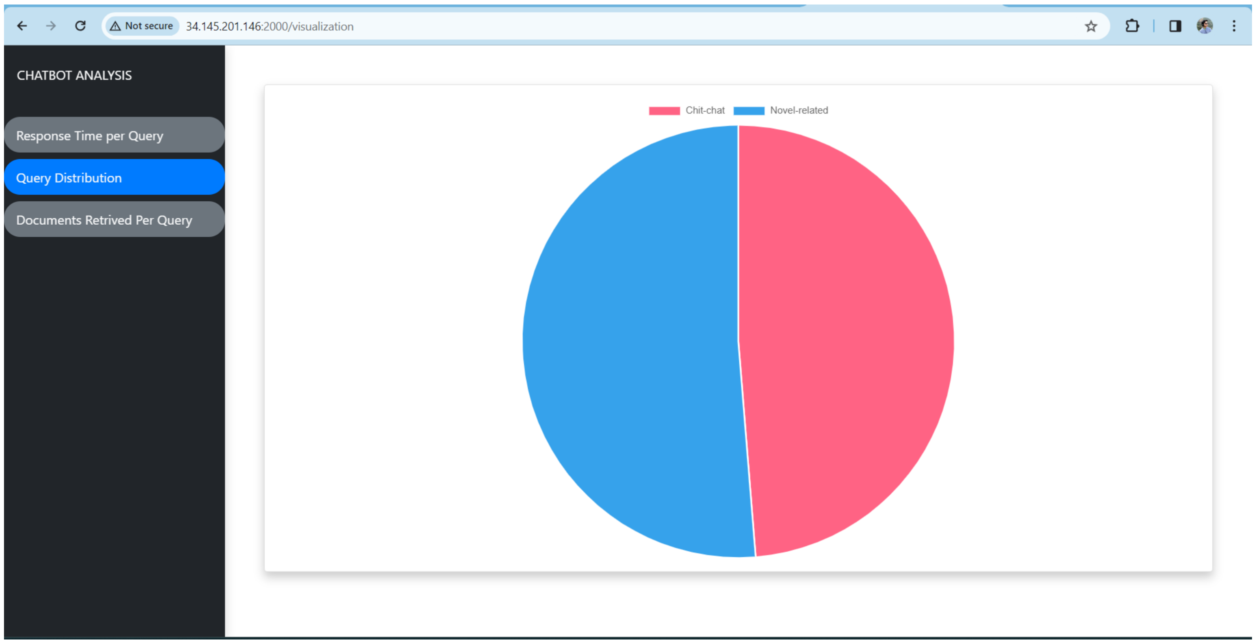
- The Adventures of Sherlock Holmes
- The Art of War
- The Time Machine
- Common Sense
- ☐ The further adventures of Zorro
- ☐ The Wolf-Men: A Tale of Amazing Adventure in the Under-World
- ☐ From Zone to Zone
- ☐ All the World Over: Interesting Stories of Travel, Thrilling Adventure and Home Life
- ☐ The lonely house
- ☐ The Erotic Motive in Literature
- ☐ Personal hygiene and physical training for women
- ☐ Tokology
- ☐ Home education
- ☐ Outside Saturn
- ☐ The red planet

Terminate Chat

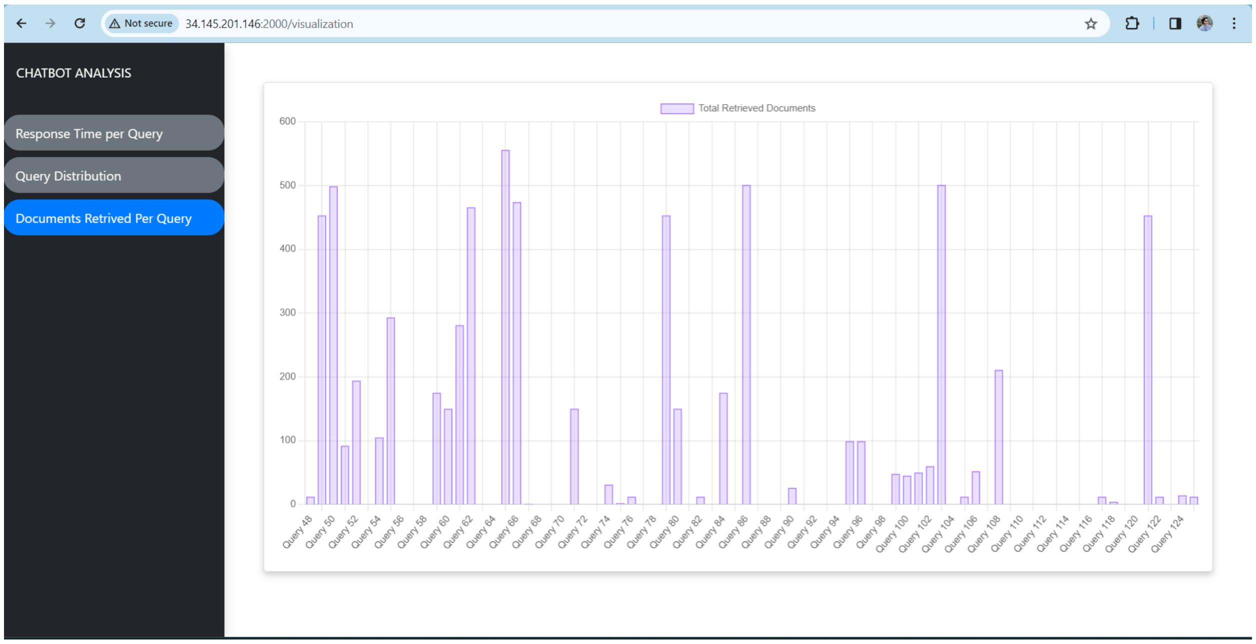
Visualization graph of Response time per query:



Visualization graph of Query Distribution:



Visualization graph of Documents Retrieved Per Query:


Not secure
34.145.201.146:2000/visualization
CHATBOT ANALYSIS
Response Time per Query
Query Distribution
Documents Retrived Per Query
Total Retrieved Documents
600
500
400
300
200
100
0
Query 48
Query 50
Query 52
Query 54
Query 56
Query 58
Query 60
Query 62
Query 64
Query 66
Query 68
Query 70
Query 72
Query 74
Query 76
Query 78
Query 80
Query 82
Query 84
Query 86
Query 88
Query 90
Query 92
Query 94
Query 96
Query 98
Query 100
Query 102
Query 104
Query 106
Query 108
Query 110
Query 112
Query 114
Query 116
Query 118
Query 120
Query 122
Query 124

Work breakdown by teammates

Work	Done By
Preprocessing	Bhushan
Novels Extraction	Bhushan
Classifier	Bhushan
Retrieving Documents	Bhushan
Chit Chat Model, Blender Bot	Deep
RAG(OpenAI)	Deep
Flask App and Hosting	Deep
UI	Deep
GCP	Deep/Bhushan

Conclusion

Overall, after using a chatbot for some queries, we found out that when the user selects multiple topics, the chatbot requires more time comparatively when there is no specific need. After using the API rather than importing libraries it was providing a much faster response, For the summary query we provided, the bot responded in 25 seconds which is the highest of all queries, and 1.48 seconds is the lowest response time, where the query was chitchat. The mean response time of the bot is 10 seconds. In the distribution of queries, the total number of chitchat queries and the total number of novels-related queries are increasing after each session. In the other graph where we interpreted the total number of documents retrieved, sometimes the number is high and when it goes into chitchat it does not retrieve any relevant documents. In the future scope, we can train models on different classifiers related to deep learning or transformers, reranking method can be based on tf-idf scores which might provide more satisfactory results.