# CSE 535 – Information Retrieval
# RAN BOT

**Anuja Wani (anujacha: 50479694)**
**Nikitha Sadananda (nsadanan: 50471079)**
**Rohan Eli (rohaneli: 50468238)**
State University of New York at Buffalo

## 1 Introduction

A chatbot system using cutting-edge AI technologies was developed in this project, where conversational AI and specialized information retrieval were combined for enhanced user interaction. A sophisticated chatbot served as the backbone of this endeavor, allowing users to ask a wide range of queries, from general inquiries to targeted questions. A robust tech stack was used to build the chatbot, including Python as the core programming language. Among the key components were the Hugging Face Blenderbot model for seamless conversation and a combination of Multinomial Naive Bayes and Random Forest classifiers for efficient query categorization. Additionally, the T5 large model ensures that users receive precise and succinct responses by summarizing extensive information. Using diverse AI technologies, this project strives to bring users relevant, personalized, and insightful interactions.

## 2 Methodology

Using the Solr API, 64,316 documents were meticulously sourced to build the chatbot's foundation. There are several subtopics within this vast repository of data, each representing a key area of knowledge. Among them are:

### 2.1 Data Acquisition

- Sports: A collection of 7,152 documents.
- Education: Collection of 6,408 documents.
- Environment: Comprising 5,020 documents.
- Health: Comprising 5,139 documents.
- Technology: With 5,029 documents.
- Economy: Encompassing 5,088 documents.
- Entertainment: Including 5,355 documents.
- Politics: A significant 7,902 documents.
- Travel: Collection of 7,398 documents.
- Food: Collection of 9,825 documents.

### 2.2 Classification

For processing user queries, the chatbot used two primary classifiers:

- Multinomial Naive Bayes Classifier:This classifier differentiates between chit-chat and topic-specific queries. In addition to recognizing and categorizing user input, the model was trained to identify and respond appropriately to each type of input
- Random Forest Classifier: Employed for topic classification, this model was pivotal in identifying the specific subject matter of user queries. Its strength in handling multi-class classification tasks allowed for precise topic categorization, directing the chatbot to relevant data resources.

### 2.3 Conversational AI Integration

Hugging Face's {facebook/blenderbot_small-90M} model enabled the chatbot to have conversations that were natural and human-like. To facilitate conversational exchanges and guarantee that responses were coherent and pertinent to the context, this model was implemented.

### 2.4 Information Retrieval and Summarization Process

We used a cosine similarity algorithm to build an enhanced retrieval system. In response to user queries, this system was built to extract the most pertinent documents from the collection. The content was summarised using Hugging Face's T5 big model after retrieval. In order to provide users with brief and intelligible summaries, the summary procedure was essential in extracting the main ideas from the retrieved documents.

### 2.5 System Architecture and Workflow

The main programming language used to create the chatbot was Python, which was selected for its adaptability and interoperability with a wide range of AI and NLP tools. The several components of the system architecture were intended to be smoothly integrated(Figure 1):

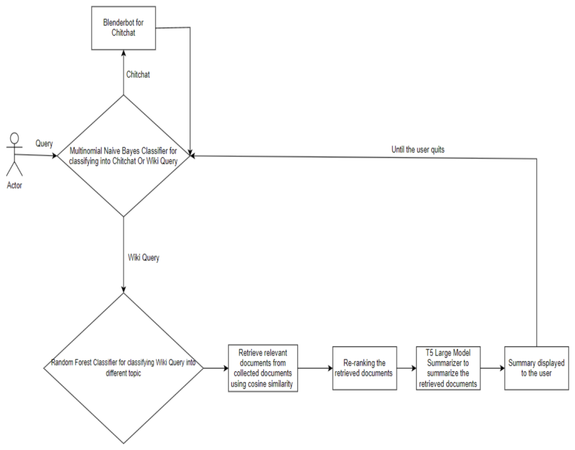

Blenderbot for Chitchat
Chitchat
Query
Multinomial Naive Bayes Classifier for classifying into Chitchat Or Wiki Query
Actor
Until the user quits
Wiki Query
Random Forest Classifier for classifying Wiki Query into different topic
Retrieve relevant documents from collected documents using cosine similarity
Re-ranking the retrieved documents
T5 Large Model Summarizer to summarize the retrieved documents
Summary displayed to the user

Figure 1: Architecture diagram

- User Input Processing:The chatbot used the Multinomial Naive Bayes Classifier to initially identify the type of query it had received. If the query was classified as idle chatter, the Blenderbot model assumed control and produced a dialogue-based answer, else the query was further processed for topic specific query handling.

- Topic-Specific Query Handling: We first preprocessed the user input to extract keywords, for getting the query context. The complete data that we retrieved from wikipedia was used to train the Random Forest Classifier to determine the pertinent subtopics for queries that were categorized as topic-specific. We used a threshold of 0.3 for a query to be classified into a particular topic. Using this we also could capture multi-topic queries. The machine then extracted documents from the dataset that were associated with these subtopics using cosine similarity.

- Response Generation: After retrieval for the relevant documents of each topic, the T5 large model summarized the information from these documents, providing the user with a succinct and informative response.

- Analytical Features: The system also had capabilities to analyze user engagement, like the probability of chit-chat against topic-specific queries and different themes, the query response time and the length of the query versus the answer time.

### 2.6 Development process

Numerous difficulties were faced during the development process, namely in merging several models and making sure that data flowed smoothly between components. Through testing and iterative system architecture improvement were used to address these. Optimizing the effectiveness of the information retrieval and summarization processes as well as the classifiers' performance received particular focus.

### 2.7 Visualization

The visualization techniques implemented in the Flask application significantly enhance user experience and data interpretation. The "ChitChat vs Query Probability Plot" visualizes the probability distribution between casual conversations and informational queries using scipy's ‘make_interp_spline‘ for smooth curve plotting. Similarly, the "Topic Classification Probabilities Plot" displays varying probabilities of topics based on user queries, aiding in understanding user focus areas by plotting probabilities for all topics. The "Query Response Times Plot," through a scatter and line plot, illustrates the time taken by the system to respond to queries, which is vital for assessing system performance and identifying efficiency issues.
Additionally, the "Length of Query vs Time Plot" correlates query length with response times, providing insights into how longer queries may affect processing time. These visualization techniques collectively improve user interaction by providing visual context to textual data, enable quick and effective analysis of system performance and user interaction patterns, and assist administrators or developers in making informed decisions about system enhancements and optimizations.

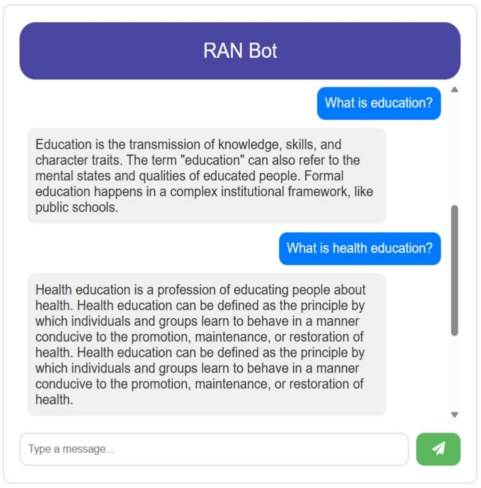
### 2.8 Error handling

In the design of our chatbot system, robust error handling plays a crucial role, particularly in dealing with unexpected or random user inputs. For instance, when a user inputs a nonsensical or irrelevant string like 'abcdegfjcdnjk', instead of causing a system crash or returning a generic error message, our chatbot gracefully handles such scenarios. It responds with a predefined fallback message such

as 'Have you heard of...,' ensuring the chat flow remains uninterrupted and user-friendly.
This approach is part of our commitment to ensuring a resilient and user-centric experience. The error handling mechanism is designed to recognize inputs that don't match any known patterns or topics and to respond in a way that invites the user to rephrase or ask something else, instead of ending the chat abruptly. This feature not only enhances the robustness of our chatbot but also aligns with best practices in user experience design, where maintaining a conversational flow is essential, even in the face of unexpected user behavior.

## 3 Screenshots

Please refer following screenshots - Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8.



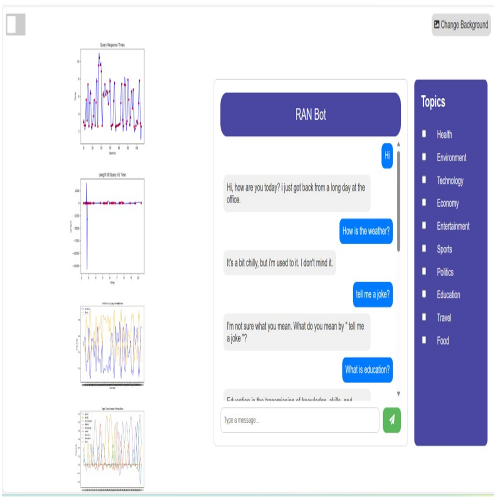
What is education?

Education is the transmission of knowledge, skills, and character traits. The term "education" can also refer to the mental states and qualities of educated people. Formal education happens in a complex institutional framework, like public schools.

What is health education?

Health education is a profession of educating people about health. Health education can be defined as the principle by which individuals and groups learn to behave in a manner conducive to the promotion, maintenance, or restoration of health. Health education can be defined as the principle by which individuals and groups learn to behave in a manner conducive to the promotion, maintenance, or restoration of health.

Figure 2: Topic specific query



Hi

Hi, how are you today? i just got back from a long day at the office.

How is the weather?

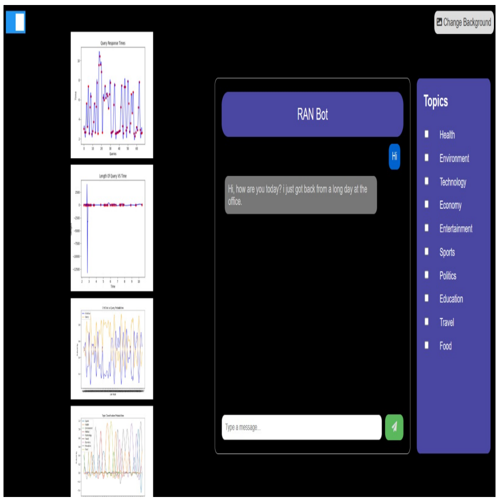
It's a bit chilly, but i'm used to it. I don't mind it.

tell me a joke?

I'm not sure what you mean. What do you mean by " tell me a joke "?

What is education?

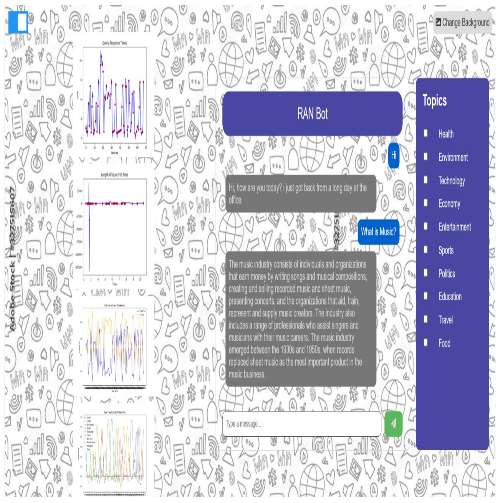
Figure 3: Chitchat query



Hi

Hi, how are you today? i just got back from a long day at the office.

Figure 4: UI darkmode



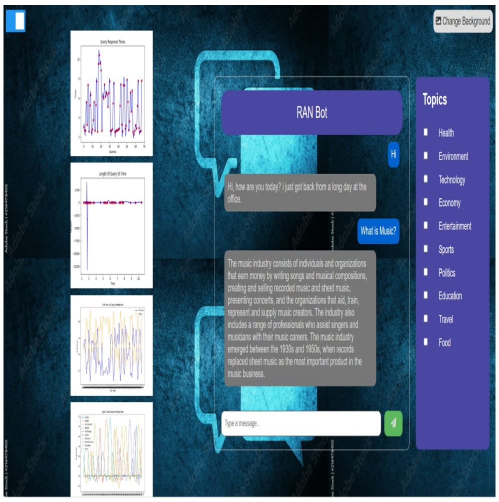
Hi

Hi, how are you today? i just got back from a long day at the office.

What is Music?

The music industry consists of individuals and organizations that earn money by writing songs and musical compositions, creating and selling recorded music and sheet music, presenting concerts, and the organizations that aid, train, represent and supply music creators. The industry also includes a range of professionals who assist singers and musicians with their music careers. The music industry emerged between the 1930s and 1950s, when records replaced sheet music as the most important product in the music business.

Figure 5: Background image for UI



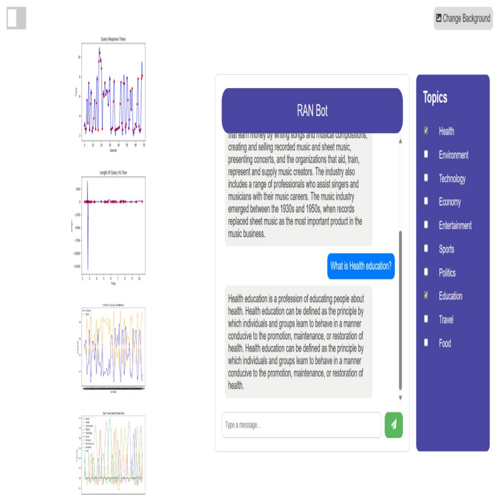
Hi

Hi, how are you today? i just got back from a long day at the office.

What is Music?

The music industry consists of individuals and organizations that earn money by writing songs and musical compositions, creating and selling recorded music and sheet music, presenting concerts, and the organizations that aid, train, represent and supply music creators. The industry also includes a range of professionals who assist singers and musicians with their music careers. The music industry emerged between the 1930s and 1950s, when records replaced sheet music as the most important product in the music business.

Figure 6: UI Image


Change Background

RAN Bot

that earn money by writing songs and musical compositions, creating and selling recorded music and sheet music, presenting concerts, and the organizations that aid, train, represent and supply music creators. The industry also includes a range of professionals who assist singers and musicians with their music careers. The music industry emerged between the 1930s and 1950s, when records replaced sheet music as the most important product in the music business.

What is Health education?

Health education is a profession of educating people about health. Health education can be defined as the principle by which individuals and groups learn to behave in a manner conducive to the promotion, maintenance, or restoration of health. Health education can be defined as the principle by which individuals and groups learn to behave in a manner conducive to the promotion, maintenance, or restoration of health.

Type a message...

Topics
- ✓ Health
- Environment
- Technology
- Economy
- Entertainment
- Sports
- Politics
- ✓ Education
- Travel
- Food

Figure 7: Checkbox selection for topics

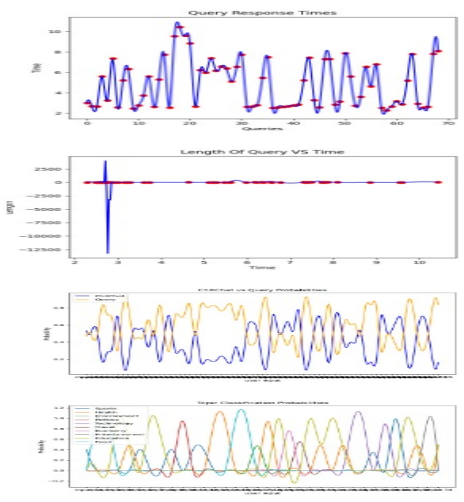


Figure 8: UI Visualization

## 4 Contribution

**Anuja Wani:** Document retrieval, Topic classification, T5 Summarizer, Multi-topic implementation, Report, Visualization, UI/UX

**Nikitha Sadananda:** Document retrieval, ChitChatvsQuery Classification, Re-ranking, Report, Visualization, UI/UX

**Rohan Eli:** Topic classification, Multi-topic implementation, T5 Summarizer, Re-ranking, Error Handling, UI/UX, Integration

## 5 Conclusion

An important advancement in AI-driven interactive technology has been made with the creation of this chatbot system. The system exhibits impressive versatility and accuracy in responding to user queries by combining complex classification algorithms with a large, organised dataset and cutting-edge conversational AI. The potential of AI in creating dynamic and responsive chat systems is demonstrated by the smooth integration of various AI models, such as the T5 model for summarization and the Blenderbot for conversation. This project not only establishes a standard for upcoming advancements in AI-chatbot technology, which aims to provide more intuitive, context-aware, and user-centric experiences, but it also serves as an example of how AI can be used practically to create intelligent systems.

## References

- Ramakrishna Kumar. 2008. A Review on Chatbot Design and Implementation Techniques
- Alfirna Rizqi Lahitani, Adhistya Erna Permanasari,Noor Akhmad Setiawan. April 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment
- Multinomial Naive Bayes classifier sklearn.naive_bayes.MultinomialNB
- Random Forest classifier sklearn.ensemble.RandomForestClassifier
- Facebook Blenderbot Blenderbot
- T5 Summarizer t5-large