MULTI-TOPIC INFORMATION RETRIEVAL CHATBOT

Aaqib Wadood Syed, Aditya Bothra, Atharva Patil, Gunjan Vivek Swamy
Department of Computer Science
State University of New York at Buffalo
Buffalo, NY, USA
{aaqibwad,abothra,atharvap,gunjanvi}@buffalo.edu

ABSTRACT

Searching and retrieving relevant information quickly is a very important feature for many applications. In this paper we have attempted to create a retrieval based chatbot on chitchat and reddit datasets using deep learning techniques, predefined inputs, response patterns and other approaches to fetch the appropriate response. The model and GUI is written in python and predefined patterns and responses are in JSON format.

1 INTRODUCTION

Chatbots are computer application that are intelligent enough to simulate conversations with humans. These are cost effective and hence getting more and more popular everyday. These agents are being used as experts providing responses and assisting users with routine tasks.

But implementing chatbot presents many challenges, biggest being variety of ways in which a query can be described. Chatbot also requires large collection of data to learn from.

In this report we are presenting our Retrieval based Chatbot Mr. Robot which is trained on two different datasets - Chitchat and Reddit to interact with user. It can converse about multiple topics - politics, environment, technology, healthcare and education. It uses deep learning techniques, predefined inputs, response patterns and other approaches to fetch the appropriate response. If there are several responses matching the user input then the best input is determined using BERT model. The chatbot is hosted on GCP server, the model and GUI is written in python and predefined patterns and responses are in JSON format.

2 METHODOLOGY

2.1 DATA PREPROCESSING

For chitchat dataset we have json file that has predefined patterns and responses.

Tokenization - We tokenized our text data to break it into small words. After this we append each word to the list of words by iterating through the patterns. Finally we created a list of classes. We used Python NLTK library for this task.

Lemmatization - We then lemmatized each word and removed duplicates, and created pickle files of both words and the classes.

2.2 TRAINING AND TEST DATA CREATION

We next created the training dataset by random shuffling of data. Our input is the pattern and our output is the class our input pattern belongs to.

2.3 Model

Now we used the training dataset to train our Machine Learning Neural Network Model on Chitchat dataset.

The model has 3 layers Input layer - 128 neurons, Hidden Layer - 64 neurons, and the Output Layer which has the same size as our number of intents. We have a frequency rate of 0.5 at each step during training time. Using Keras sequential API, we compiled our model with stochastic gradient descent, and with our training dataset in 200 epochs.

If the model classifies it as Chitchat dataset then we retrieve the appropriate response and show it to the user. If not then we assume that the query is for reddit and we hit Solr for Reddit Index. According to the given topic we invoke respective Solr API.

On Solr we have default BM25 model for retrieving the top ten results. We then send these results to BERT for re-ranking and getting the best possible result. BERT assigns the cosine similarity scores to these top ten documents, and we return the most relevant document.

3 Sample Screenshots

3.1 Chatbot Working

First greeting screen of the Chatbot.

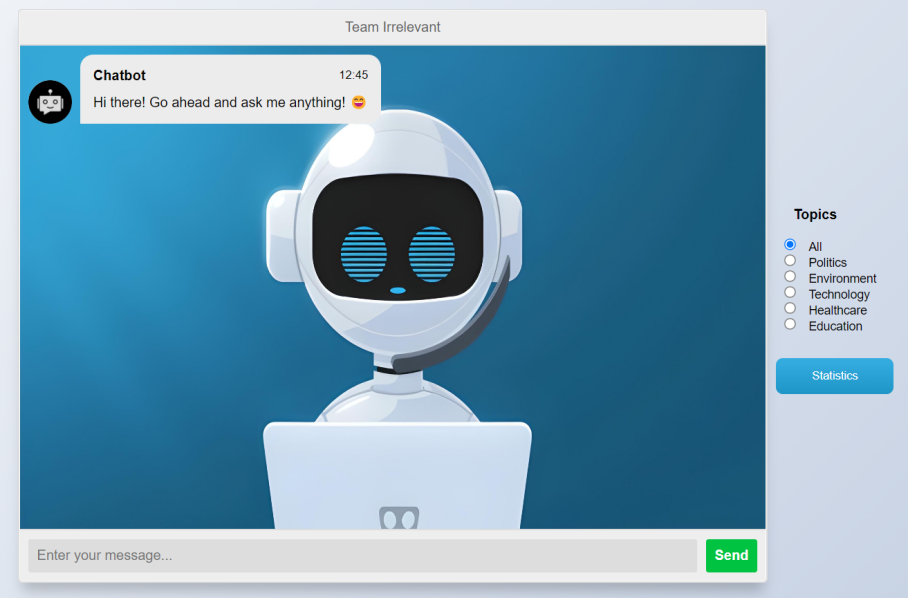

Figure 1: Chatbot Greeting

Chatbot retrieving responses from the Chitchat index.

Figure 2: Chatbot Chitchat

Chatbot retrieving responses from the Reddit Index for particular topic.

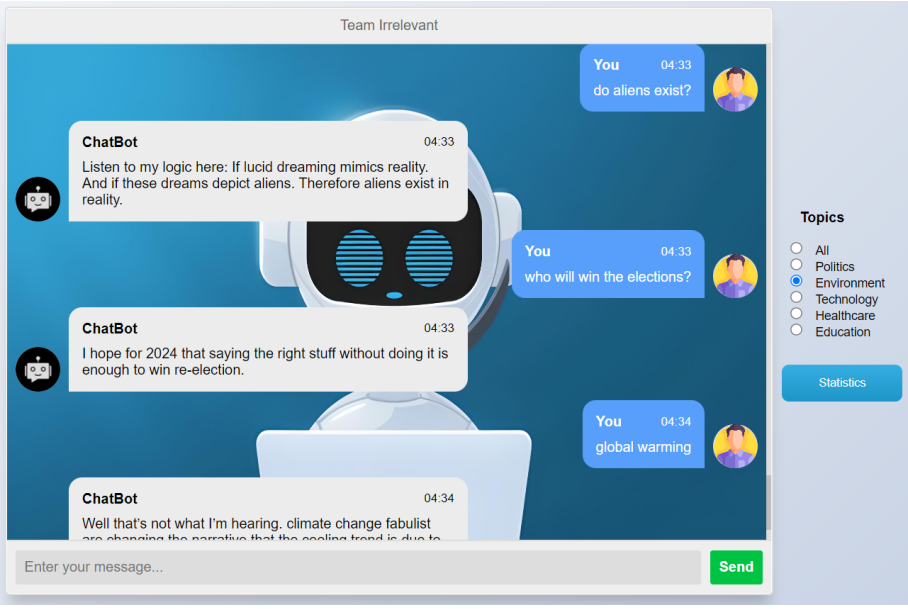

Figure 3: Chatbot Reddit

3.2 Statistics and Visualization

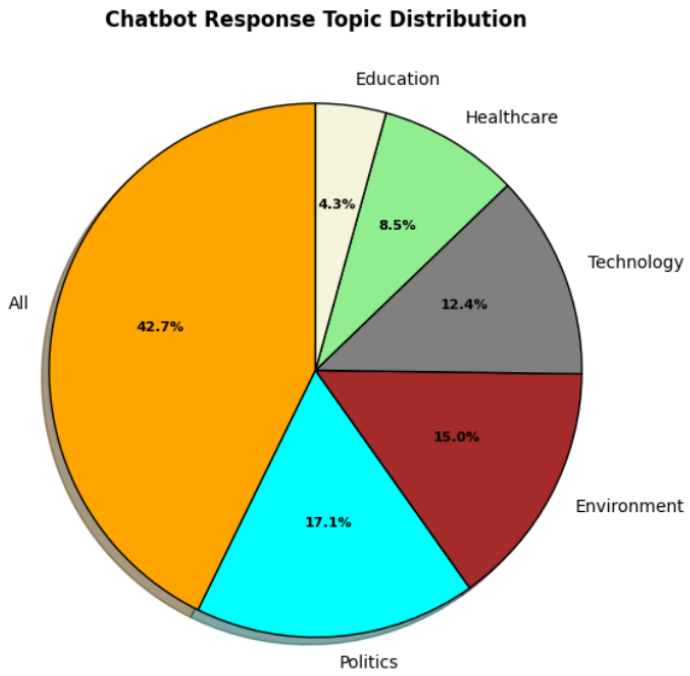

Figure 4: Chatbot Response Topic Distribution

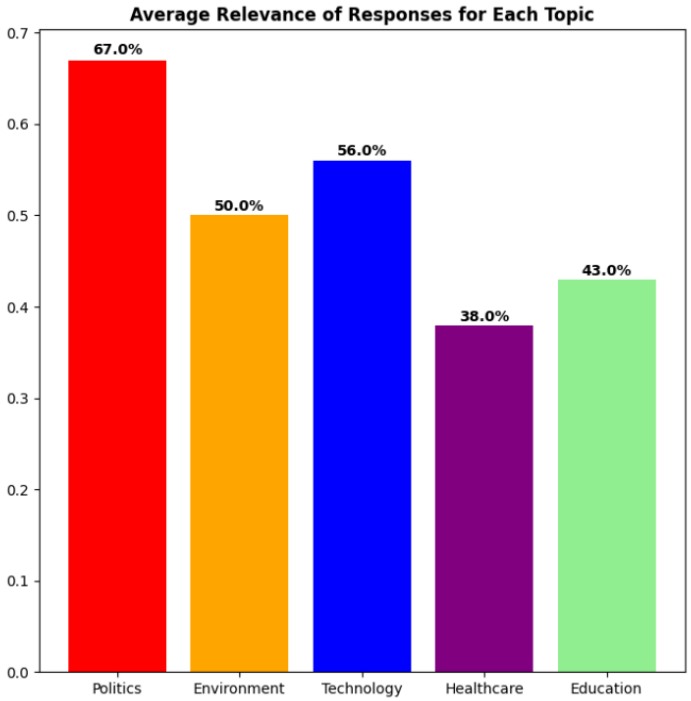

Figure 5: Relevance of Responses for each Topic

4 WORK BREAKDOWN

Aaqib Wadood Syed: Data Collection, Data Pre-processing, Chatbot GUI, Solr Indexing, BERT integration, ML model implementation, Visualization, Report

Aditya Bothra: Data Collection, Data Pre-processing, Chatbot GUI, BM25 model implementation, ML model implementation, BERT integration, Visualization, Report

Atharva Patil: Data Collection, Data Pre-processing, Chatbot GUI, Solr Indexing, Report, BERT integration, ML model implementation,Visualization

Gunjan Vivek Swamy: Data Collection, Data Pre-processing, Chatbot GUI, BM25 model implementation, Report, BERT integration, ML model implementation,Visualization

5 CONCLUSION

In this project we developed our chatbot framework for answering questions related to multiple topics - Politics, Environment, Technology, Healthcare, Education and day-to-day conversation. Chatbot is trained on large dataset of 200000 submissions from reddit and the Chitchat datasets. The retrieval-based approach works robustly with the ML and NLP frameworks.

REFERENCES

https://arxiv.org/abs/2108.01436

An Introduction to Information Retrieval - Christopher D. Manning Prabhakar Raghavan Hinrich Schütze

https://dmitry-kan.medium.com/neural-search-with-bert-and-solr-ea5ead060b28

https://www.nltk.org/