TechTalk

1. Ahmet Akib Gültekin
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
akib.gultekin@std.yildiz.edu.tr

2. Muhammet Kayra Bulut
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
kayra.bulut@std.yildiz.edu.tr

3. Sait YALÇIN
Bilgisayar Mühendisliği
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
sait.yalcin@std.yildiz.edu.tr

***Abstract*—This project is two NLP model with Python that generates responses about users' questions about computer science. Project contains two seperate model to generate response. First model uses handcrafted dataset and the second model uses Reddit posts.**

***Index Terms*—Python, Language Model, Chatbot, Natural Language Model**

I. What is ChatBot

A chatbot is a software program that can chat with a human in written form. Chatbots typically aim to chat by understanding and responding in human language. Chatbots can be used for various purposes, but are commonly used in areas such as customer service, reservations, and information distribution. Chatbots are often used through a website or messaging application to answer people's questions or perform necessary tasks [1], [2], [3].

A. How do Chatbots work?

A chatbot learns and understands a special language to be able to chat with a human in written form. This language is often designed as a more artificial language for chatbots to understand, although it can also be similar to human language. When a chatbot understands human language, it responds to questions made in that language. Chatbot responses are typically composed of pre-programmed answers or information retrieved from a database or create answers [4].

B. What is the relationship between Chatbots and Natural Language Processing?

Natural Language Processing (NLP) technology enables computer systems to input data using natural language and process it. This technology makes it possible for chatbots to answer questions and perform tasks using natural language. Thanks to NLP technology, chatbots can understand questions asked by people using words and sentences and learn to provide appropriate responses [5]. NLP technology also makes language learning and usage more effective, enabling chatbots to communicate with humans more naturally.

NLP technology is used in many different areas besides chatbots [6]. For example, NLP technology performs tasks such as translating texts into various languages, extracting meaning, and classification. As a result, people can use their natural language to search on search engines or use language learning applications on their mobile phones.

II. What is Data Mining

Data mining is the process of discovering patterns, trends, and insights from large datasets using various statistical and computational techniques [7]. It involves identifying and extracting relevant information from data, transforming it into an understandable structure, and finally interpreting the results. The goal of data mining is to uncover hidden relationships, patterns, and trends that can be used to make informed decisions.

Data mining has a wide range of applications in various fields, including business, healthcare, finance, marketing, and more [8]. In business, data mining can be used to identify customer behavior patterns and preferences, optimize marketing campaigns, and improve decision-making processes. In healthcare, data mining can be used to analyze patient data and identify risk factors for diseases.

One of the most commonly used techniques in data mining is classification, which involves dividing data into distinct classes or categories based on certain attributes [9]. Another popular technique is clustering, which groups similar data points together based on their characteristics. Data mining is a complex and dynamic field, and there are many techniques and tools available for analyzing and interpreting data. However, it is important to note that data mining can also raise ethical concerns, particularly in terms of privacy and security [7]. Therefore, it is crucial to use data mining techniques in a responsible and transparent manner, and to ensure that appropriate measures are in place to protect sensitive information.

III. Types of Chatbot

- Rule-based chatbots [10]: These chatbots follow a set of predefined rules to understand and respond to user queries. They have a limited ability to learn and improve their responses.
- AI-based chatbots [11]: These chatbots use artificial intelligence and natural language processing (NLP) techniques to understand and respond to user queries. They can learn from user interactions and improve their responses over time.

- Hybrid chatbots [12]: These chatbots combine rule-based and AI-based approaches to provide more accurate and effective responses to user queries.
- Task-specific chatbots: These chatbots are designed to perform specific tasks, such as booking appointments or providing customer support.
- Social media chatbots: These chatbots are integrated with social media platforms and are used for marketing, customer engagement, and other purposes.

IV. Datasets

A. Pre-Prepared Dataset

- Data Collection: First, data containing example questions and possible answers for different question types (intents) that the chatbot needs to respond to are collected in a file called intents.json. These data are then read and used by the program.
- Data Processing: The collected data is then processed. For example, the category (intent) to which each question and answer belongs is identified. These are then stored in a set of documents.
- Word and Class Counts: Using data mining techniques, the number of words in the documents and the number of documents in each class are calculated. These counts are necessary to determine the dimensions of the matrices that will be used in the next steps.
- Preprocessing: The data is preprocessed to make it usable. At this stage, the words in the documents are stemmed, and unnecessary characters and punctuation marks are removed.
- Preparing Training Data: Next, the training data is prepared. At this stage, a matrix is created for each document. The columns of the matrix are the preprocessed words, and each row represents a specific document. Each element of the matrix indicates whether or not that word appears in that document.
- Model Creation: The model is a neural network that will work on the training data. This neural network learns patterns in the training data and uses them to recognize these patterns in new inputs. In this code, the model is an artificial neural network that consists of two hidden layers. The first hidden layer has 128 neurons and uses relu as its activation function. The second hidden layer has 64 neurons and also uses ReLU as its activation function.

B. Reddit Posts

- Data Collection: Using PRAW, specific posts are collected from selected subreddits that contain questions and answers relevant to the chatbot's intended purpose. The collected data, which includes the raw text of the posts, is stored in a suitable format, such as separate text files for questions and answers.
- Data Processing: The collected data is processed to extract the relevant questions and answers. The raw text may be preprocessed to remove unnecessary characters, converted to lowercase, and tokenized into words.
- Preparing Training Data: Next, the training data is prepared. The questions and answers are paired and transformed into a format suitable for training a neural network. This may involve converting the text into numerical representations, such as word embeddings or one-hot encoding, and creating input-output pairs for training the transformers model.
- Model Creation: The custom transformers model created and trained using the reddit training data. Transformers is a machine learning model widely used in natural language processing (NLP). It excels in tasks like understanding text, translation, and question answering. With its ability to handle large datasets, it is employed to create powerful language models. During training, the transformer model learns to capture patterns in the training data and generalize its understanding to handle new inputs. It is a powerful model for natural language processing tasks, leveraging the self-attention mechanism to effectively process sequential data.
- Model Training: The preprocessed training data is utilized to train the custom transformer model. The model is fed with the training data, and its weights and biases are adjusted through backpropagation. The training process involves multiple epochs to optimize the model's performance.
- Model Evaluation: Once the model is trained, it is evaluated using a validation set of questions and answers. Metrics such as accuracy, precision, recall to assess the model's performance. The model may be optimized further by adjusting hyperparameters or using other techniques based on the evaluation results.
- Model Deployment: After training and optimizing the model, it can be deployed in a production environment as a chatbot to respond to user queries and provide relevant answers based on the specific scenario it was trained for.

V. What is TechTalk

TechTalk, a chatbot project that can interact with people on topics related to computer science. This chatbot can answer questions and provide guidance on topics that people might be interested in related to computer science.

The project was developed using natural language processing and deep learning techniques. Two different chatbots were trained, one with a pre-prepared dataset and the other with a dataset prepared using Reddit topics.

- The model trained with the pre-prepared dataset uses a Long Short-Term Memory (LSTM) neural network to classify the user's message into a category. Then, the chatbot uses pre-programmed responses or information from a database to provide a suitable response.
- The second model utilized transformers model, which was trained with Reddit data for training on TechTalk's topics.

The TechTalk chatbot project can help students, researchers, professionals, and other interested individuals to gain more knowledge on computer science topics and provides learning opportunities on chatbot technology and deep learning.

VI. TechTalk Project

In our project, we have created a chatbot using two deep learning models with two different datasets. The first model is trained on a dataset that contains categories (intents), the model, and the responses. And the second model trained on the Reddit posts about computer science. We use a special type of recurrent neural network (LSTM) to classify the user's message into a category and then provide a response.

For our Retrieval Based Chatbot project, we used Python libraries such as TensorFlow, NLTK, Keras, Numpy, Pickle, and Json with Python version 3.9.

To summarize our project:

- We have created two datasets, one on Reddit and the other one by ourselves, and prepared answers to the questions.
- We performed a series of cleaning, stemming (lemmatization), and tagging (tokenization) operations on the data set to prepare it for training.
- We modeled our training for the first model. To do this, we created a three-layered artificial neural network using Sequential. For our other model, we used the transformer model and trained it. We trained and tested both of our models based on our training and test datasets.
- We saved both of our models to prepare for new queries.
- We received new queries through the interface and sent the estimated responses generated by the trained model to the screen.

VII. The Statistics

In this project there are two different language model and each model has its own dataset. The first model uses a pre-prepared dataset. And the second model uses Reddit posts to train. The reddit posts collected by using web scraping method on Reddit topics about computer science.

A. Pre-Prepared Dataset

The dataset is a JSON file containing a set of questions and corresponding answers that can be used for a chatbot in the field of computer science. Each question has a "pattern" assigned to it, and a corresponding "response" is provided for each pattern. In addition, each pattern is also assigned a "tag". The most frequently used words in the patterns are "explain", "data", and "work", as can be seen in Fig. 1 and Fig. 2

The least frequently used words in the patterns are "use", "detection", and "object", as can be seen in Fig. 3 and Fig. 2.

The most frequently used words in the responses are "data", "used", and "include", as can be seen in Fig. 4 and Fig. 5.

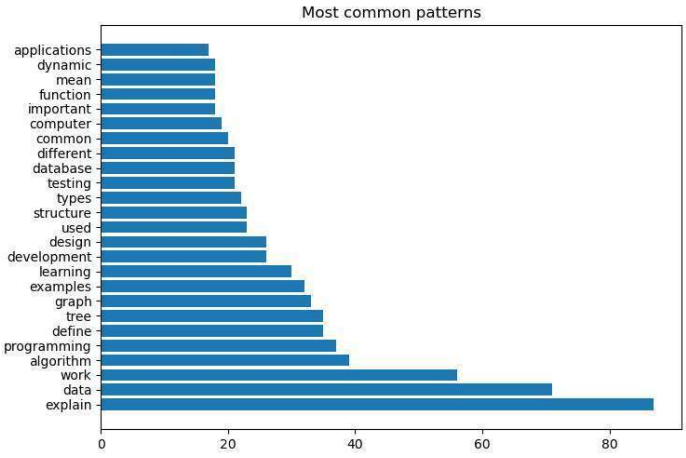

Fig. 1. Most Common Patterns

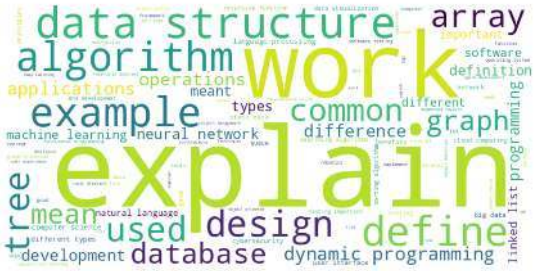

Fig. 2. Word Cloud Patterns

The least frequently used words in the responses are "points", "variables", and "actions", as can be seen in Fig. 6 and Fig. 5

The dataset can be used as a source of data for training chatbots. Using the patterns and tags, chatbots can identify the user's questions and provide an appropriate response.

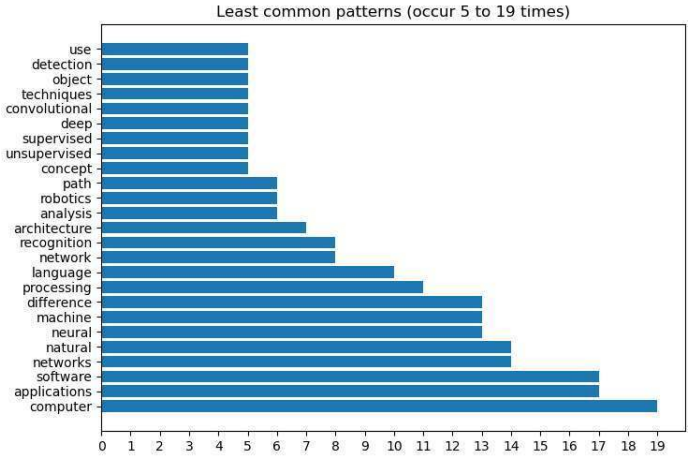

Fig. 3. Least Common Patterns

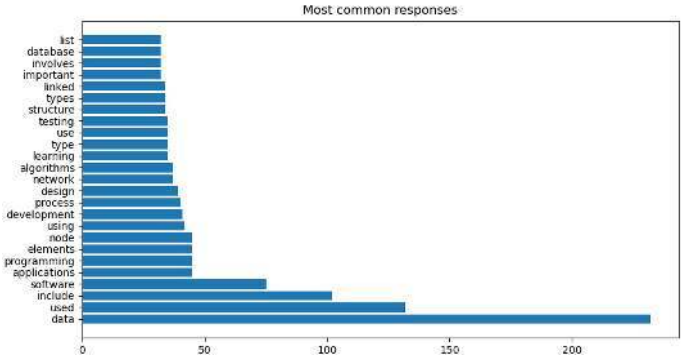

Fig. 4. Most Common Responses

Fig. 5. Word Cloud Responses

Additionally, the dataset provides example question-answer pairs for chatbots that can answer frequently asked questions in the field of computer science.

The quality of the questions and answers in the dataset is an important factor in determining the success of the chatbot. The questions should be clear and understandable, and the answers should be accurate and useful. Therefore, careful creation and editing of the dataset is crucial.

Furthermore, it is important to properly define the patterns and tags in the dataset. The patterns should cover a variety of question types as much as possible, and the tags should accurately reflect the relevant topics.

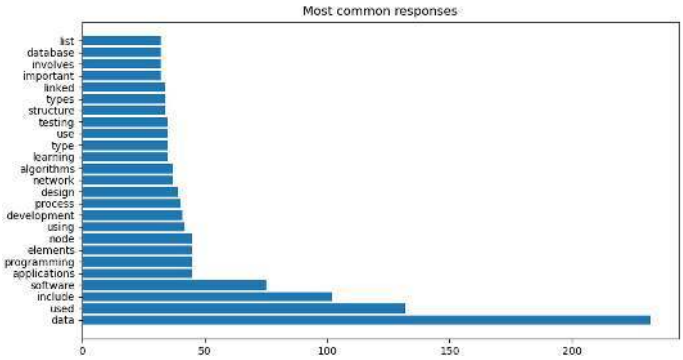

Fig. 6. Least Common Responses

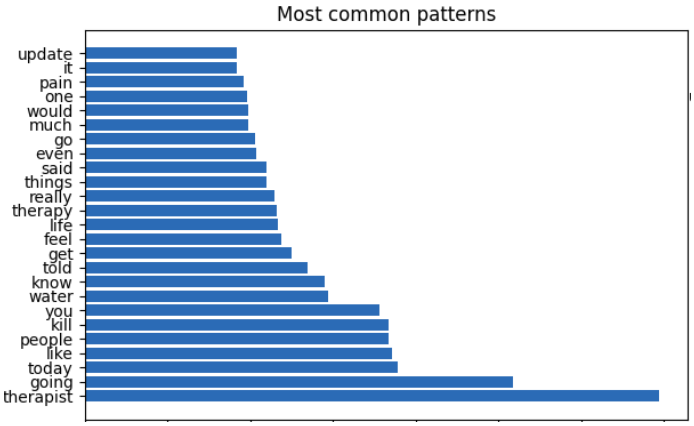

Fig. 7. Most Common Patterns Reddit

In conclusion, this dataset provides many examples of questions and answers that can be used by chatbot developers in the field of computer science. However, the quality and accuracy of the dataset are important factors in determining the success of the chatbot.

B. Reddit Posts

This dataset is obtained using the praw library to collect posts from various subreddits based on selected sorting options. The posts are then transformed into question-answer pairs in a loop, including the comments under the post and the replies to those comments. The resulting data is saved in a single question-answer file.

This dataset contains 913 question-answer pairs, dataset created in a loop of replies to and replies to posts on reddit. the majority of which are related to computer science and cover various topics.

The average length of the questions in the dataset is 74 characters. The shortest question is 5 characters, while the longest is 344 characters.

The average length of the answers in the dataset is 176 characters. The shortest answer is 3 characters, while the longest is 996 characters.

63% of the questions in the dataset consist of a single sentence, while 32% consist of two sentences, and 5% consist of three sentences. The majority of the answers consist of a single sentence (82%), while 14% consist of two sentences, and 4% consist of three sentences.

35% of the questions in the dataset start with the word ”what”, 28% with ”how”, 13% with ”which”, 8% with ”why”, 7% with ”can”, and 9% with other words.

The most frequently used words in the patterns are ”threpist”, ”going”, ”today”, ”like”, as can be seen in Fig. 7 and Fig. 8.

The least frequently used words in the patterns are ”mentally”, ”lets”, ”grateful”, ”unfrdunately”, as can be seen in Fig. 9 and Fig. 8.

Fig. 8. Word Cloud Patterns Reddit

Fig. 11. Word Cloud Responses Reddit

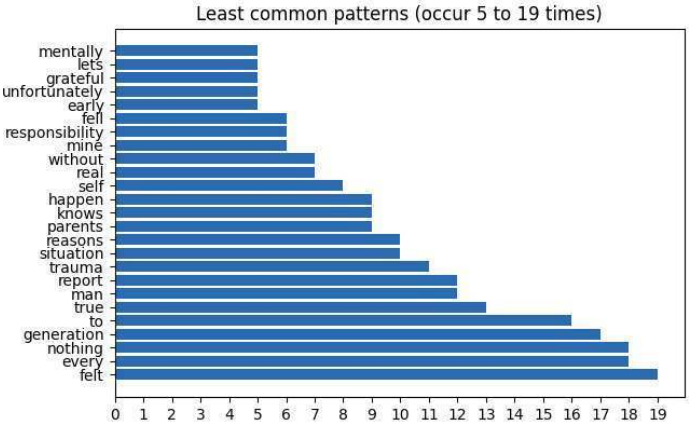

Fig. 9. Least Common Patterns Reddit

The most frequently used words in the responses are ”threpist”, ”going”, ”today”, ”like”, as can be seen in Fig. 10 and Fig. 11.

The least frequently used words in the responses are ”mentally”, ”lets”, ”grateful”, ”unfrdunately”, as can be seen in Fig. 12 and Fig. 11.

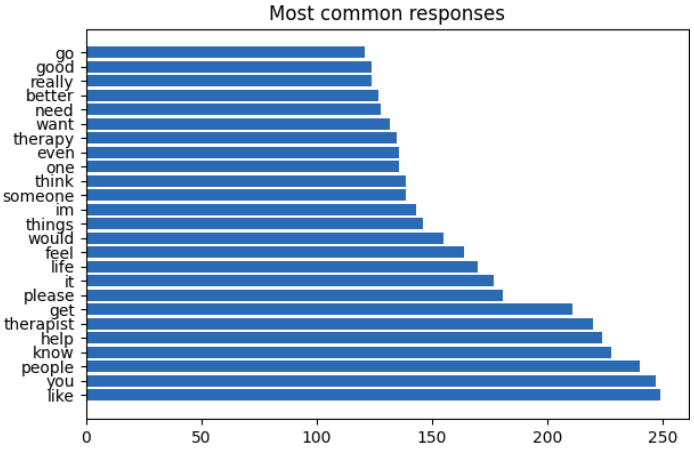

Fig. 10. Most Common Responses Reddit

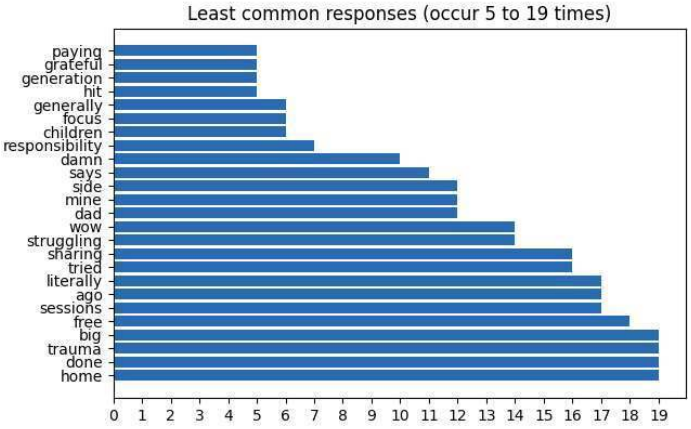

Fig. 12. Word Cloud Responses Reddit

VIII. Application

In application we use Angular to provide more useful interface and chart library of Angular for comprehendible charts. In more detail application contains chat, analytics and settings page. User can communicate with the chatbot with using chat page. Analytics page provide user to opportunity to visualize the chatbot's data. And user can change the NLP model from settings page

References

[1] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” *Machine Learning with Applications*, vol. 2, p. 100006, 2020.

Fig. 13. Application Chat Page

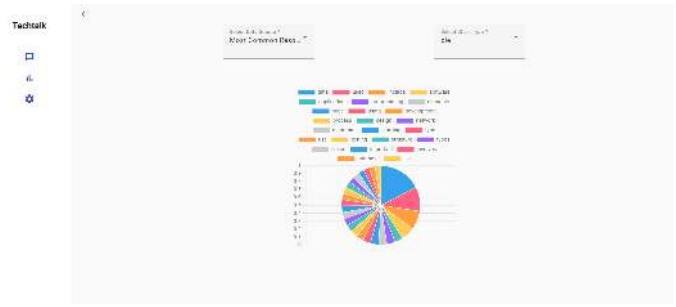

Fig. 14. Application Analytics Page

Fig. 15. Application Settings Page

[2] D. Kaczorowska-Spychalska, "How chatbots influence marketing," *Management*, vol. 23, no. 1, pp. 251–270, 2019.

[3] L. Tudor Car, D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, and R. Atun, "Conversational agents in health care: scoping review and conceptual analysis," *Journal of medical Internet research*, vol. 22, no. 8, p. e17158, 2020.

[4] A. Tarek, M. El Hajji, E.-S. Youssef, and H. Fadili, "Towards highly adaptive edu-chatbot," *Procedia Computer Science*, vol. 198, pp. 397–403, 2022.

[5] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[6] A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang, "Decoding with large-scale neural language models improves translation," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1387–1392.

[7] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques*. Morgan kaufmann, 2022.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.

[9] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with java implementations," *Acm Sigmod Record*, vol. 31, no. 1, pp. 76–77, 2002.

[10] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 373–383.

[11] N. Albayrak, A. Özdemir, and E. Zeydan, "An overview of artificial intelligence based chatbots and an example chatbot application," in *2018 26th signal processing and communications applications conference (SIU)*. IEEE, 2018, pp. 1–4.

[12] Y. Gapanyuk, S. Chernobrovkin, A. Leontiev, I. Latkin, M. Belyanova, and O. Morozenkov, "The hybrid chatbot system combining q&a and knowledge-base approaches," in *7th International Conference on Analysis of Images, Social Networks and Texts*, 2018, pp. 42–53.