CHATBOT FOR FAQ

Submitted by

AMAL LEON (00867541)

SELVA VIGNESHWAR (00866540)

SYSTEM ARCHITECTURE

In this chapter, the System Architecture for chatbot for FAQ's using deep learning isrepresented and the modules are explained.

4.1 ARCHITECTURE DESIGN

In system architecture the detailed description about the system modules and the working of each module is discussed as shown in figure.

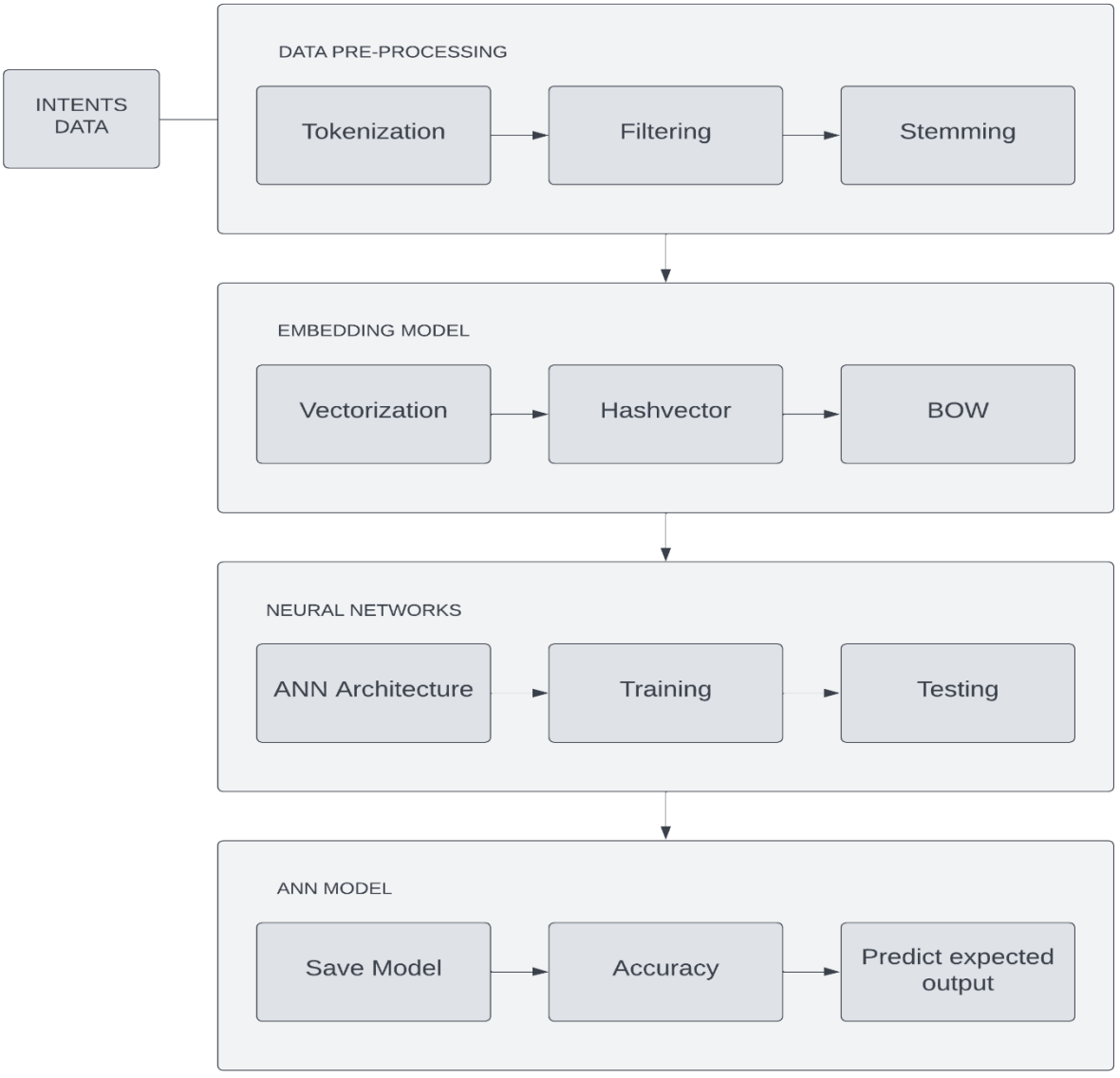


Figure 1 System Architecture of chatbot application for online FAQ using deep learning

ARCHITECTURE DESCRIPTION

1 INTENTS DATA

Intent refers to the goal the user has in mind when typing in a question or comment. While entity refers to the modifier the user uses to describe their issue, intent is what they really mean. Entities are predefined categories of names, organizations, time expressions, quantities, and other general groups of objects that make sense.

2 DATA PREPROCESSING

Tokenisation

Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.

Filtering

Filtering is the process of removing stop words or any unnecessary data from the sentence. Removing special characters and removing numbers and removing punctuations.

Stemming

Stemming is one of the most common data pre-processing operations we do in almost all Natural Language Processing (NLP). stemming is the process of removing a part of a word, or reducing a word to its stem or root. This might not necessarily mean we’re reducing a word to its dictionary root.

3 EMBEDDING MODEL

Vectorization

Word Embeddings or Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics. The process of converting words into numbers are called Vectorization.

Hashvector

Convert words to integers used when vocabulary is very large.

BOW (Bag of words)

Bag of words is a Natural Language Processing technique of text modelling. In technical terms, we can say that it is a method of feature extraction with text data. This approach is a simple and flexible way of extracting features from documents. A bag of words is a representation of text that describes the occurrence of words within a document. We just keep track of word counts and disregard the grammatical details and the word order. It is called a “bag” of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document

4 NEURAL NETWORK ARCHITECTURE

ANN Architecture

ANN architecture is based on the structure and function of the biological neural network. Similar to neurons in the brain, ANN also consists of neurons which are arranged in various layers. ANN has an input, various hidden layers and output layer.

Training

A training model is training a neural network means finding the appropriate weights of the neural connections thanks to a feedback loop called gradient backward propagation. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model. This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model.

Testing

The purpose of testing is to compare the outputs from the neural network against targets in an independent set (the testing instances). we test the trained machine learning model using the test dataset quality assurance is required to make sure that the software system worksaccording to the requirements. Were all the features implemented as agreed? Doesthe program behave as expected? All the parameters that you test the program against should be stated in the technical specification document. Moreover, software testing has the power to point out all the defects and flaws during development. You don't want your clients to encounter bugs after the software isreleased and come to you waving their fists. Different kinds of testing allow us to catch bugs that are visible only during runtime.

5 ANN MODEL

ANN models are the extreme simplification of human neural systems. An ANN comprises of computational units analogous to that of the neurons of the biological nervous system known as artificial neurons. Mainly, the ANN model constitutes of three layers, viz., input, hidden, and output. Each neuron in the nth layer is interconnected with the neurons of the $(n + 1)$th layer by some signal. Each connection is assigned a weight. The output may be calculated after multiplying each input with its corresponding weight. The output passes through an activation function to get the final ANN output. The ANN may be useful in solving different engineering and science problems. As such, the ANN has image compression, function approximation, differential equations, stock market prediction, medical diagnosis, and signal processing.

6 MODEL ACCURACY

Accuracy is a metric that generally describes how the model performs across all classes. It is useful when all classes are of equal importance. It is calculated as the ratio between the number of correct predictions to the total number of predictions.

SYSTEM IMPLEMENTATION

In this chapter, the System Implementation for the chatbot application for FAQ's.

SYSTEM DESCRIPTION

Deep learning is a subset of machine learning, which is essentially a neural network with three or more layers. These neural networks attempt to simulate the behavior of the human brain—albeit far from matching its ability—allowing it to "learn" from large amounts of data. While a neural network with a single layer can still make approximate predictions, additional hidden layers can help to optimize and refine for accuracy.

A chatbot application using a Deep Learning approach is presented. The main focus of this study was to investigate deep learning based techniques with the best accuracy in predicting and explore its applicability with particular importance to the dataset. Deep Learning techniques were used to analyze the dataset to carry out data validation, data cleaning, and data visualization on the given dataset. The results of the different deep learning algorithms were compared to predict the results. The proposed system consists of data collection, data preprocessing, construction of a predictivemodel, dataset training, dataset testing. The aim of this study is to prove the effectiveness and accuracy of a deep learning algorithm for predicting user queries.

Artificial neural networks, usually simply called neural networks, are computing systems inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Deep learning is highly scalable due to its ability to process massive amounts of data and perform a lot of computations in a cost- and time-effective manner.

This directly impacts productivity (faster deployment/rollouts) and modularity

and portability (trained models can be used across a range of problems).

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.

DATASET

I have created a intent dataset for General FAQ's about the university.

Example:


```
"intents": [
    {
     "tag": "Campus",
     "patterns": ["Can I visit the campus", "Is it possible for me to come to the campus",
     "May I pay a visit to the campus", "Is campus visitation allowed?",
     "Am I permitted to tour the campus?","Can I explore the campus in person?","Is there an option for me to check out the campus?",
     "Would it be acceptable for me to visit the campus?","Am I allowed to physically go to the campus?",
     "Is there an open-door policy for campus visits?",
     "Is campus access available for visitors?"],
     "responses": [
      "Yes! Visit our Graduate Events page to learn more about visit opportunities."
     ],
     "context_set": ""
    },
    {
     "tag": "Online Degree",
     "patterns": ["Do you offer online degrees", "Do you provide online degree programs?",
     "Are online degrees offered at your institution?",
     "Can I pursue a degree online through your university?",
     "Is it possible to earn a degree via online courses?",
     "Do you have options for distance learning degrees?"],
     "responses": [
      "The University of New Haven offers graduate programs in business, healthcare, technology, and more. Visit our Online Degrees page for more information."
     ],
     "context_set": ""
    },
```

```
{
  "tag": "Placement",
  "questions": [],
  "patterns": ["Do your school offer job placement", "Does your school provide job placement assistance?",
  "Are there career placement services available to students?",
  "Can students access job placement support through the school?",
 "Is job placement assistance part of the student services?",
 "Are there resources to help graduates find employment?"],
  "responses": [
    "Our nationally-recognized Career Development Center provides the skills and connections to identify a meaningful career and an opportunity for students to pursue their passion. Services include career assessments, networking opportunities, interview preparation, and more."
  ],
  "context_set": ""
},
```


PREPROCESSING

Data pre-processing is a process of cleaning the raw data i.e. the data is collected is the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for the analysis. Therefore, certain steps are executed to convert the datainto a small clean data set, this part of the process is called as data pre-processing. The majority of the real-world datasets for machine learning are highly susceptible to be missing, inconsistent, and noisy due to their heterogeneous origin. Applying data mining algorithms on this noisy data would not give quality results as they would fail to identify patterns effectively. Data Processing is, therefore, important to improve the overall data quality. Duplicate or missing values may give an incorrect view of the overall statistics of data. Outliers and inconsistent data points often tend to disturb the model's overall learning, leading to false predictions.

As we know that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, we definitely need data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

Most of the real-world data is messy, some of these types of data are:

1. **Missing data:** Missing data can be found when it is not continuously created ordue to technical issues in the application (IOT system).
2. **Noisy data:** This type of data is also called outliners, this can occur due to humanerrors (human manually gathering the data) or some technical problem of the device at the time of collection of data.
3. **Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Three Types of Data

1. Numeric e.g. income, age
2. Categorical e.g. gender, nationality
3. Ordinal e.g. low/medium/high

How can data pre-processing be performed?

These are some of the basic pre processing techniques that can be used to convert raw data.

1. **Conversion of data:** As we know that Deep Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.
2. **Ignoring the missing values:** Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn't be performed if there are a lot ofmissing values in the dataset.
3. **Filling the missing values:** Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median orhighest frequency value is used.
4. **Deep learning:** If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.

5.2 TESTING

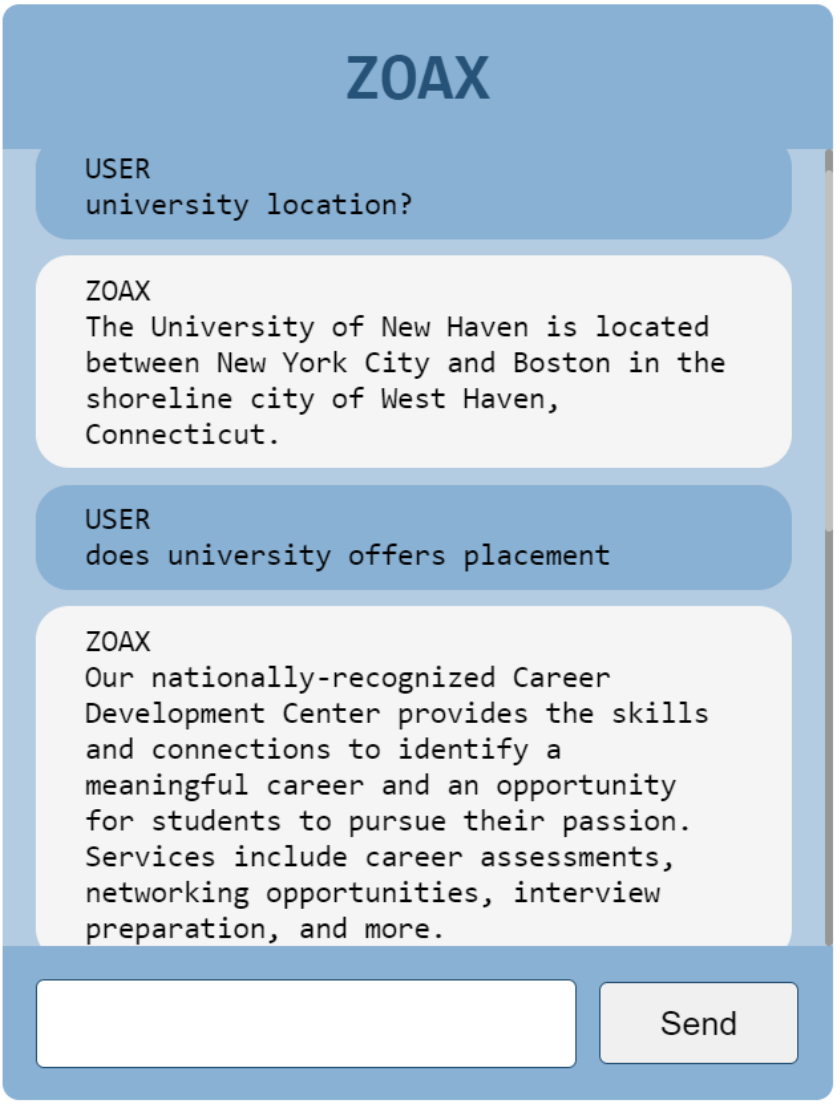
Once the data is divided into the 3 given segments we can start the training process. In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. The model uses any one of the models that we had chosen in step 3 point 3. Many practitioners may rely solely on machine learning model performance evaluation. However, evaluation is not the same as testing. It is important to identify their differences. Machine learning model evaluation focuses on the overall performance of the model. Such evaluations can consist of performance metrics and curves, and perhaps examples of incorrect predictions. Test after introducing a new component, model, or data, and after model retraining. Test before deployment and production. Write tests to avoid recognized bugs in the future. Once the model is trained we can usethe same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a confusion matrix, this tells us how well our model istrained. A confusion matrix has 4 parameters, which are '**True positives'**, **'True Negatives'**, **'False Positives'** and '**False Negative'**. We prefer that we get more values in the True negatives and true positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

- **True positives :** These are cases in which we predicted TRUE and our predicted output is correct.
- **True negatives :** We predicted FALSE and our predicted output is correct.
- **False positives :** We predicted TRUE, but the actual predicted output is FALSE.
- **False negatives :** We predicted FALSE, but the actual predicted output is TRUE.

We can also find out the accuracy of the model using the confusion matrix.

Accuracy = (True Positives +True Negatives) / (Total number of classes)

Output Screen Shots



To Run the Application the command is -

Python manage.py runserver

But need required packages before run the application.

I have uploaded the demo video.
https://drive.google.com/drive/folders/1YGlm9l6BghDqVFsqVerFCTyLRanrTySw?usp=sharing

The General Questions are,

1. Can I visit the campus?
   Yes! Visit our Graduate Events page to learn more about visit

opportunities.

2. Do you offer online degrees?
   Yes. The University of New Haven offers graduate programs in business, healthcare, technology, and more. Visit our Online Degrees page for more information.

3. Does your school offer job placement?
   Our nationally-recognized Career Development Center provides the skills and connections to identify a meaningful career and an opportunity for students to pursue their passion. Services include career assessments, networking opportunities, interview preparation, and more.

4. What is the university ranking?
   The University of New Haven is regularly recognized by national publications for academic quality and rigor. Visit our About page for more information.

5. Do you offer hybrid degree programs?
   The University of New Haven's online degrees are delivered 100 percent online in a flexible format ideal for adult learners.

6. How many students attend the University of New Haven?
   Nearly 7,000 undergraduate and graduate students attend the University of New Haven.

7. Is the diploma different for online students?
   No. Graduates of our online programs earn the same diploma as on-campus students.

8. Is the university regionally accredited?
   The University has been accredited by the New England Commission of Higher Education since 1948 and is chartered by the General Assembly of the State of Connecticut.

9. May I walk at graduation?
   Yes! All students may participate in commencement ceremonies.

10. What are the values of the University of New Haven?
    Our mission is to prepare our students to lead purposeful and fulfilling lives in a global society by providing the highest-quality education

through experiential, collaborative, and discovery-based learning.

11. What is the mission of the University of New Haven?
    The University of New Haven is a student-centered comprehensive university with an emphasis on excellence in liberal arts and professional education. Our mission is to prepare our students to lead purposeful and fulfilling lives in a global society by providing the highest-quality education through experiential, collaborative, and discovery-based learning.

12. What is the university's religious affiliation?
    The University of New Haven is an independent, nonsectarian institution.

13. What's the difference between regional and national accreditation?
    Regionally accredited higher education institutions are predominantly academically oriented, non-profit institutions. Nationally accredited schools are predominantly for-profit and offer vocational, career or technical programs.

14. When was the university founded?
    The University of New Haven was founded in 1920 on the Yale campus.

15. Where is the campus located?
    The University of New Haven is located between New York City and Boston in the shoreline city of West Haven, Connecticut.