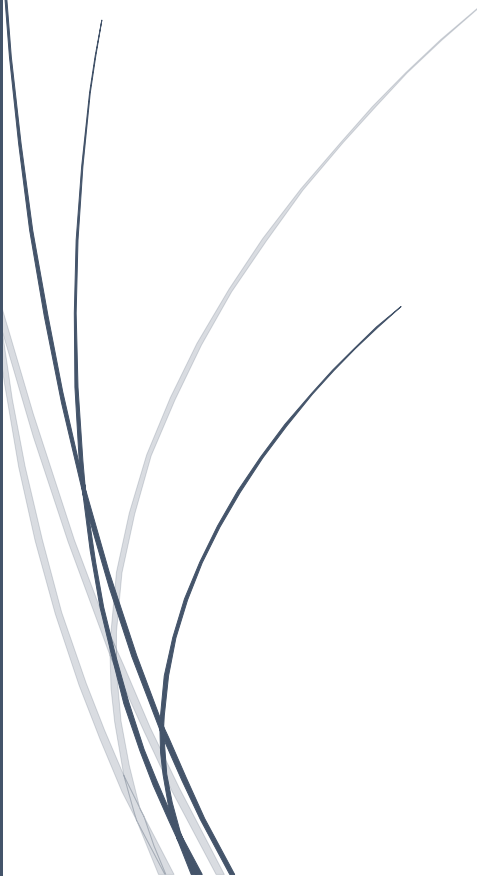


A dark blue vertical bar on the left side of the page. A blue arrow points to the right from the bar, containing the date 4/7/2021.

4/7/2021

# 2000- FINAL PROJECT

NOOPA JAGADEESH

Several thin, curved lines in dark blue and light grey originate from the bottom left corner and curve upwards and to the right.

Sagunesh Grover – 100800447  
Roshandeep Singh Saini – 100766638  
Parth Chotai - 100802846

## Contents

|                                 |   |
|---------------------------------|---|
| 1. Problem Statement .....      | 2 |
| 2. Use Case / Requirements..... | 2 |
| 3. Data Acquisition.....        | 2 |
| 2.1 Dataset Description.....    | 3 |
| 2.2 Data Preparation .....      | 3 |
| 3. Model Building.....          | 4 |
| 4. Output and Conclusion .....  | 4 |
| 5. Work Distribution.....       | 5 |

## 1. Problem Statement

The Problem Statement is “To develop an AI chatbot to answer to answer user queries regarding **Durham College COVID-19 Awareness Program**”.

The project aims at building a chatbot that can efficiently take in user query inputs about the **Durham College COVID-19 Awareness Program**, and it able to answer the questions with a standard answer. In action it an intent classification problem where the question coming in from the user can be classified with ac “intent” and the chatbot displays the answer corresponding to that intent. It is an NLP task.

## 2. Use Case / Requirements

From the intelligence viewpoint, there are 2 types of Chatbots:

- Rule Based Chatbots
- AI based conversational bots.

The problem statement demand a Rule based chatbot because With rule-based bots, we need to pick answers/ Intents on their best prediction at the keywords you used in your inquiry question.

The Technologies that can be used for the above task are:

- CHATFUEL
- Dialogflow CX (Google Dialogflow)
- RASA
- Python Chatterbot Lib

All the above are good, serve the purpose but are meant for full-fledged conversational chatbots. Our scope is figuring out what the user intent is given a user utterance regarding **Durham College COVID-19 Awareness Program**. An intent classifier fulfill our requirement to classify user queries to the intent of the user proving them with a standard answer.

## 3. Data Acquisition

The dataset used for the chatbot was synthetically created, specifically tailored to serve the aim of the chatbot. It contains he intents / intention of the question, The ways in which a single question with the same intent can be asked and the answer to the

questions with the same intent. All of these are specifically designed to serve the aim of the chatbot.

## 2.1 Dataset Description

The data set is manually created. It consists of 1 Independent Feature “Questions” and 1 Dependent Feature “Intents”. The Feature “Answer” maps directly to the Intent and is indirectly the Intent classified by the model. There are 15 unique Intents namely:

```
{'greetings': 1,  
'cost': 2,  
'latest_measures': 3,  
'student_infected': 4,  
'student_data_collection': 5,  
'set_up': 6,  
'customize': 7,  
'data_security': 8,  
'use': 9,  
'professor_infected': 10,  
'contact_tracing': 11,  
'support': 12,  
'consent': 13,  
'processing_time': 14,  
'cancel': 15}
```

## 2.2 Data Preparation

Data preparation involved the following steps:

**Removing Punctuation** – The questions used for training were cleaned by removing punctuation with the use of Regular Expression.

**Tokenization** – The sentences were tokenized to list of words and the entire set of sentences was tokenized to a list of sentences. This was done to get the total unique word count and further encoding the list of words.

**Conversion to Lower Case** – All the words were converted to lower case to remove redundancy.

**Encoding Sentences** – The list of words was encoded to a sequence of integers for the model to train on. The words were encoded based on the words extracted from document (BoW).

**Padding sentences** – The converted sequence of integers representing a sentence was padded to match the length of the longest sentence. This needs to be done so the size of input is the same.

**One-Hot-Encoding** – The unique intents were one hot encoded. one hot encoding is used for the output variable, to get a more nuanced set of predictions than a single label.

### 3. Model Building

The model used for the chatbot is a neural network. A sequential model allows us to create models layer by layer in a step by step fashion.

```
def create_model(vocab_size, max_length):  
    # Creating a Sequential model  
    model = Sequential()  
    # Embedding layer with no. of neuron=vocab size  
    model.add(Embedding(vocab_size, 128, input_length = max_length, trainable = False))  
    # Bidirectional LSTM for better performance and understanding context  
    model.add(Bidirectional(LSTM(128)))  
    # Fully connected Dense layer  
    model.add(Dense(32, activation = "relu"))  
    # Dropout Layer to prevent Overfitting  
    model.add(Dropout(0.5))  
    # Dense output Layer with no. of Neurons=no. of Intents  
    model.add(Dense(15, activation = "softmax"))  
  
    return model
```

It is Sequential model with an Embedding layer(input size = vocabulary size and length of input = max length of sentences).

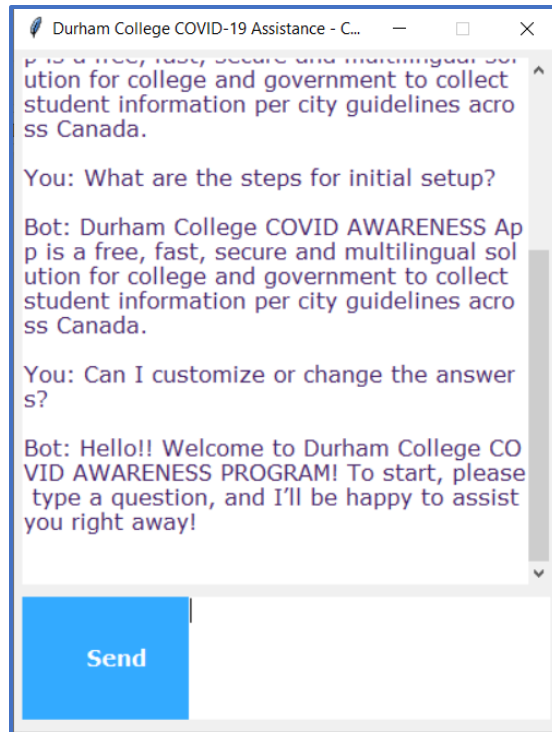
A layer of Bi-LSTM, a fully connected Dense layer, a Dropout Layer(to prevent overfitting) and a Dense layer as the output layer of size 15 for the 15 unique intents.

### 4. Output and Conclusion

The user Input is tokenized, and the intents are predicted with the probabilities. The answer associated with the intent of the highest probability is displayed on the UI.

```
text = "I would like to get some technical assistance."  
pred = predictions(text)  
answer = get_final_output(pred, unique_intent)  
print(answer)  
  
['i', 'would', 'like', 'to', 'get', 'some', 'technical', 'assistance']  
  
C:\Users\rosha\AppData\Roaming\Python\Python37\site-packages\tensorflow\python\keras\engine\sequential.py:425: UserWarning: `model.predict_proba()` is deprecated and will be removed after 2021-01-01. Please use `model.predict()` instead.  
warnings.warn("`model.predict_proba()` is deprecated and "
```

Contact us at covidawareness@durhamcollege.ca



A detail video presentation is attached with the code files.

The assumption and constraints of the classifier are:

The data is synthetically generated so the amount of data is very for the attaining high accuracy. As it can be seen that the validation loss does not decrease beyond 2.41108 which high value. The model is still underfitting.

## 5. Work Distribution

| Category                | Tasks              | Participant                       |
|-------------------------|--------------------|-----------------------------------|
| <b>Individual Tasks</b> | Setting up Project | Sagunesh Grover (1.5 hrs.)        |
|                         | Dataset Creation   | Parth Chotai (1.5 hrs.)           |
|                         | Coding             | Sagunesh Grover (3 hrs.)          |
|                         | UI Code            | Roshandeep Singh Saini (3 hrs.)   |
|                         | Testing            | Parth Chotai (3 hrs.)             |
| <b>Team Tasks</b>       | Code Merge         | Roshandeep Singh Saini (1.5 hrs.) |
|                         | Data Preparation   | Team (2 hrs.)                     |
|                         |                    | Team (2 hrs.)                     |

|  |                                      |               |
|--|--------------------------------------|---------------|
|  | Final Presentation and Documentation | Team (2 hrs.) |
|--|--------------------------------------|---------------|