



UNIVERSITY  
OF WOLLONGONG  
IN DUBAI

CSCI323: Modern Artificial Intelligence

# LeadVA

By: Maryam Zia (6854631), Hiba Gohar (7090997), Fatima Aya Tchantchane (5560251), Mohamed Nihal (7061158), Dani Ternani (7108771), Aravind Nair (7307883), Mohammad Ejaz (7305849)



## Assignment Cover Sheet

**Subject Code:** CSCI323

**Subject Name:** Modern Artificial Intelligence

**Submission Type:** Moodle Submission

**Assignment Title:** Project Report

**Student Name:** Hiba Gohar, Maryam Zia, Fatima Aya Tchanchane, Mohamed Nihal Kattakath, Dani Terrnani, Aravind Pradeep Nair, Mohammad Ejaz Chowdhary

**Student Number:** 7090997, 6854631, 5560251, 7061158, 7108771, 7307883, 7305849

**Student Phone/Mobile No.:** +971 568533412, +971561516727, +971521969079, +971556269471, +971501373840, +971553860914, +971 54 717 6189

**Student E-mail:** [Hg453@uowmail.edu.au](mailto:Hg453@uowmail.edu.au), [Mz948@uowmail.edu.au](mailto:Mz948@uowmail.edu.au), [Fat120@uowmail.edu.au](mailto:Fat120@uowmail.edu.au), [mnk665@uowmail.edu.au](mailto:mnk665@uowmail.edu.au), [dt104@uowmail.edu.au](mailto:dt104@uowmail.edu.au), [apn712@uowmail.edu.au](mailto:apn712@uowmail.edu.au), [merc682@uowmail.edu.au](mailto:merc682@uowmail.edu.au)

**Lecturer Name:** Dr. Patrick Mukala

**Due Date:** 4th Dec 2022

**Date Submitted:** 4th Dec 2022

### PLAGIARISM:

The penalty for deliberate plagiarism is FAILURE in the subject. Plagiarism is cheating by using the written ideas or submitted work of someone else. UOWD has a strong policy against plagiarism. The University of Wollongong in Dubai also endorses a policy of non-discriminatory language practice and presentation.

**PLEASE NOTE:** STUDENTS MUST RETAIN A COPY OF ANY WORK SUBMITTED

### DECLARATION:

I/We certify that this is entirely my/our own work, except where I/we have given fully-documented references to the work of others, and that the material contained in this document has not previously been submitted for assessment in any formal course of study. I/we understand the definition and consequences of plagiarism.

**Signature of Student:**

### Optional Marks:

**Comments:**

### Lecturer Assignment Receipt (To be filled in by student and retained by Lecturer upon return of assignment)

**Subject:**

**Student Name:**

**Due Date:**

**Signature of Student:**

**Assignment Title:**

**Student Number:**

**Date Submitted:**

### Student Assignment Receipt (To be filled in and retained by Student upon submission of assignment)

**Subject:**

**Student Name:**

**Due Date:**

**Signature of Lecturer**

**Assignment Title:**

**Student Number:**

**Date Submitted:**

## Table of Contents

<b>Executive Summary</b> .....	4
<b>Introduction</b> .....	4
Background information about the copy & project .....	4
Purpose & Overview .....	5
Project scope & business objective .....	5
Main problems and how we plan to address them (Motivation) .....	5
Business Value .....	6
Stakeholders .....	6
<b>Project Management</b> .....	6
Work breakdown structure .....	6
Gantt Chart.....	8
Project Management Tools .....	8
<b>Overall Design (Features, modules, functions)</b> .....	8
<b>System architecture</b> .....	9
<b>Technical requirements</b> .....	9
Platforms, Tools & Libraries.....	9
Code Structure.....	10
Code Documentation .....	17
Implementation (Test cases) .....	19
<b>Limitations &amp; Constraints</b> .....	20
<b>Future improvements</b> .....	20
<b>Conclusion</b> .....	21

## Executive Summary

The majority of companies today are attempting to integrate artificial intelligence into their businesses or industries. The University of Wollongong in Dubai was the company that the CSCI323 Project Team reached out to for developing an artificial intelligence-based solution the company. The University of Wollongong in Dubai (UOWD) is the first international Australian university in the United Arab Emirates. It has approximately 3,500 students from 108 countries and 152,000 alumni. UOWD still aspires to increase enrolment by introducing new initiatives that will attract students.

Upon meeting the UOWD stakeholders, it was determined that the Recruitment Team's recruitment periods, during which they attempt to secure potential leads via thousands of phone calls per month, posed a significant challenge. Following up on each lead was time-consuming for the contact center and could potentially be counterproductive if the lead turns out to be uninterested. In addition, it was found that the call centre typically spends a significant amount of time discussing general information rather than providing information that is more specific to the lead's interests. After completing the calls, the Recruitment team manually determines which prospective student is a lead. To solve this issue, the proposed solution is a system called "LeadVA" that consists of a virtual assistant & a Machine Learning model. The virtual assistant replaces the call centre by initiating phone calls to the list of prospective students. The virtual assistant will engage the prospective student in an interactive conversation, responding to their inquiries and then probing deeper with questions tailored to their particular areas of interest. On the basis of the conversation, a Machine Learning model will determine whether the prospect should be considered a lead based on a set of parameters, such as the emotional score, the length of the chat, etc.

Throughout the project, several AI modules were implemented, including speech-to-text, NLP, text-to-speech, and Machine Learning. However, there were limitations on the available resources. Access to real datasets from the university was restricted, as these datasets were deemed confidential. However, the project's implementation was still feasible. Certainly, there was room for future improvements so as the appropriate resources, particularly budget, were available. When the LeadVA solution was presented to the Recruitment Team, they immediately considered it and desired to implement it, as LeadVA will be able to initiate every call instead of the employees and determine whether or not the potential students are leads.

## Introduction

### Background information about the copy & project

The University of Wollongong in Dubai (UOWD) is the first international Australian university in the UAE, there are about 3,500 students from 108 different countries, and around 152,000 alumni. UOWD aims to increase their number of students by introducing innovative marketing strategies that will attract prospective students. Dr. Patrick Mukala, Professor of CSCI323 - Modern Artificial Intelligence course at UOWD, presented his students with an assignment that entailed contacting a local company, conducting an interview about an issue the firm were facing, and proposing an AI-based solution.

The project members had decided on UOWD as the client for this assignment, specifically the recruitment department. We conducted multiple interviews with a member of the recruiting team in the project's early stages to learn about the challenges they encounter and what they wish would be an alternative to it. The member of the recruiting team began by explaining how they utilize the CRM (Customer relationship management) system, which is a platform that holds information received from marketing events about each obtained lead. Leads are prospective students who may or may not have an interest in enrolling at the university. The call center receives a list of leads and related information from the recruitment team and contact about 60 of them to gauge interest. If the lead is converted to a



contact, it indicates that the student is interested in applying (this is only inferred from the phone call between the student and advisor) Once a student becomes a contact, the recruiting officer will contact them to confirm their interest, guide them through the application process, and notify them about early bird discounts/bursaries.

The team pointed out that the main difficulty of the CRM process was the extensive time and effort required to go through leads and convert them into contacts. They also has mentioned that the prospect might be preoccupied and end the conversation, leading the caller to assume that they are no longer interested. Since this is the case, the caller may draw the wrong conclusion. It's also possible that the lead is not interested, but the customer service advisor infers otherwise, in which case the lead is sent to marketing for a series of drip emails with pertinent information. This would result in the prospects irritation can raise customer complaints and decrease satisfaction with customer service.

### Purpose & Overview

After meeting the recruitment department, we concluded they want to find a solution, since they manually identify the leads and manually call the students. The purpose of our project to solve their complications by eliminating and eradicating the need of the Human Resources to make these calls automatically. The Virtual Assistant that we are implementing to them will produce the calls and the machine learning will generate the lead scores. Our implementation will be able to analyze the behavior and the interaction of the student with the V.A to output the lead scores.

In addition, we would want to optimize the system and reduce resources. The optimization of the system makes it simpler for the users and make the life of the employees easier. As well as, the advancement of the CRM can reduce time in repetitive tasks and allocate the employees time in other activities. This will increase conversion and application rate which is a main goal for every recruitment department.

### Project scope & business objective

The business objective here is to implement an AI powered system that replaces the calling center and interacts with potential students in an effort to boost productivity and efficiency.

Our scope is to obtain details on what sort of questions are answered by the call center and a dataset that is a used by the recruitment team to sort lead details and manage their interests. This dataset will be used to train the Machine Learning model we built that is based on Logistic Regression to predict whether the student is a future lead or not.

The details acquired from call center will be applied towards implementing a Vistual Assistant that will interact with the prospect more efficiently as compared to a call center worker. After that, the details acquired from the virtual assistant such as Length of call, Questions Asked, Prior Knowledge etc will be added to the provided dataset, which will then be sent to the Machine Learning model to give as an output of interested and uninterested leads.

### Main problems and how we plan to address them (Motivation)

The primary challenges of this project are to streamline the lead generation process and diminish the amount of time spent by HR professionals on it. Currently, the UOWD recruitment team provides the UOWD call center with a database of leads. Using that database, the call center employees contact 50–60 leads every day identifying which leads are interested and which are not. Promoting interest in enrolling at the university with this approach is a highly ineffective use of resources. For the contact center, following up on each lead is time-consuming and might be counterproductive if the lead turns out to be uninterested. Furthermore, the call center typically spends a substantial share of their time discussing general information as opposed to providing information which is more specialized to the lead's interests. Decrease in productivity is also a consequence of this tedious approach.

To solve this problem, the proposed solution is a system called “LeadVA” that comprises of a virtual assistant that initiates phone calls to the list of prospective students, replacing the call center. The virtual assistant will engage in an interactive conversation with the prospective student, responding to their queries and then probing deeper with questions tailored to their specific areas of interest. It will then compile the information gathered for each student and forward it to the recruitment department. The recruitment department will next forward the file to the deployed Machine Learning model. The ML model will then provide an output indicating whether the prospect should be deemed a lead based on a set of parameters, such as the emotional score and the length of the chat.

### Business Value

This solution is designed to cut down significantly on the time spent each month making outbound calls to prospective students by the recruiting team and other staff members. The success of our project will enable the recruitment department to have more time and resources to focus on other areas, which will ultimately lead to a boost efficiency and results. In addition to these benefits, implementing LeadVA solution, more students may be drawn to the university due to the improved quality of their communication experience. Therefore, this would not only enhance enrolment, but may also improve the rankings of the university.

### Stakeholders

#### Internal stakeholders

The UOWD recruitment department employees

The UOWD call center employees

Project team members

UOWD executives

CSCI323 Professor Dr. Patrik Mukala

#### External stakeholders

Prospective students

### Project Management

#### Work breakdown structure

The Work breakdown structure below describes the entire timeline of building the project and the duration of each sub task. The project was initiated on 17<sup>th</sup> October 2022 and closed on 5<sup>th</sup> December 2022, at a total of 52 days.

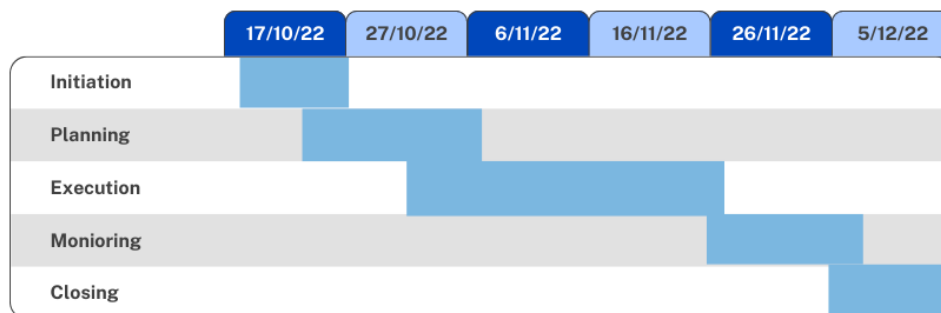


Project Processes	Sub Tasks	Start Date	End Date	Duration - Days
Initiation - I1	Interview with Recruiting officer	17/10/2022	19/10/2022	2
I2	Discussion of needs	19/10/2022	20/10/2022	1
I3	Get project approval from senior management	20/10/2022	22/10/2022	2
I4	Research about Client company background and history	22/10/2022	25/10/2022	3
I5	Client sign-off	25/10/2022	26/10/2022	1
Planning -P1	Meet and break down work into parts	26/10/2022	30/10/2022	4
P2	Assign Roles to team members	30/10/2022	31/10/2022	1
P3	Create Gantt Chart to show the estimated timeframe.	31/10/2022	2/11/2022	2
P4	Conduct analysis on deciding coding languages, software, and cloud API	2/11/2022	8/11/2022	6
Execution - E1	Meet project sponsor, client, and stakeholders	8/11/2022	11/11/2022	3
E2	Meet the IITS department representative	11/11/2022	12/11/2022	1
E3	Build a virtual assistant	12/11/2022	19/11/2022	7
E4	Sentimental analysis	19/11/2022	21/11/2022	2
E5	Making a Predictive Lead Scoring Model.	21/11/2022	27/11/2022	6
E6	Create Data sets to test the model	27/11/2022	29/11/2022	2
E7	Implement solution	29/11/2022	30/11/2022	1
E8	Perform tests to test the precision of the model	30/11/2022	1/12/2022	1
Monitoring & Controlling - MC1	Hold meetings and discuss project progress, and budgets and resolve any issues	1/12/2022	2/12/2022	1
MC2	Review change requests and monitor the implementation of said changes.	2/12/2022	3/12/2022	1
Closing - C1	Quality test on project result. Analysis of project implementation.	3/12/2022	4/12/2022	1
C2	Produce all deliverables and hand them over to Professor Patrick.	3/12/2022	4/12/2022	1
C3	Present our model to Prof. Patrick	4/12/2022	6/12/2022	2
C4	Closing meeting.	5/12/2022	7/12/2022	1
			<b>Total:</b>	<b>52</b>

## Gantt Chart

# LeadVA

## Project Timeline



## Project Management Tools

Numerous project management tools have been employed to help CSCI323 Project Team plan, track, and manage the project in order to achieve the defined project objectives within the allotted timeframe. Some of the helpful tools included:

### *Webex*

Webex by Cisco was used for our video conferences, where we held regular meetings to discuss project details, collaborate on project components or presentations, and interview stakeholders.

### *OneDrive*

Microsoft operates OneDrive, a service for file hosting and file sharing. Up to 2TB of storage space is available on the UOWD OneDrive account. Team members were able to easily upload, share, and download project files by utilizing OneDrive. Using a Word document uploaded to OneDrive in the cloud, team members could collaborate in real-time to edit the report, thereby reducing the workload and increasing productivity.

### *Trello*

Trello is a web-based application for creating visual to-do lists. The team was able to list tasks, keep them organized, create a to-do list, and create a custom workflow that was continuously updated to help them track their progress and complete the workload within the allotted time.

## Overall Design (Features, modules, functions)

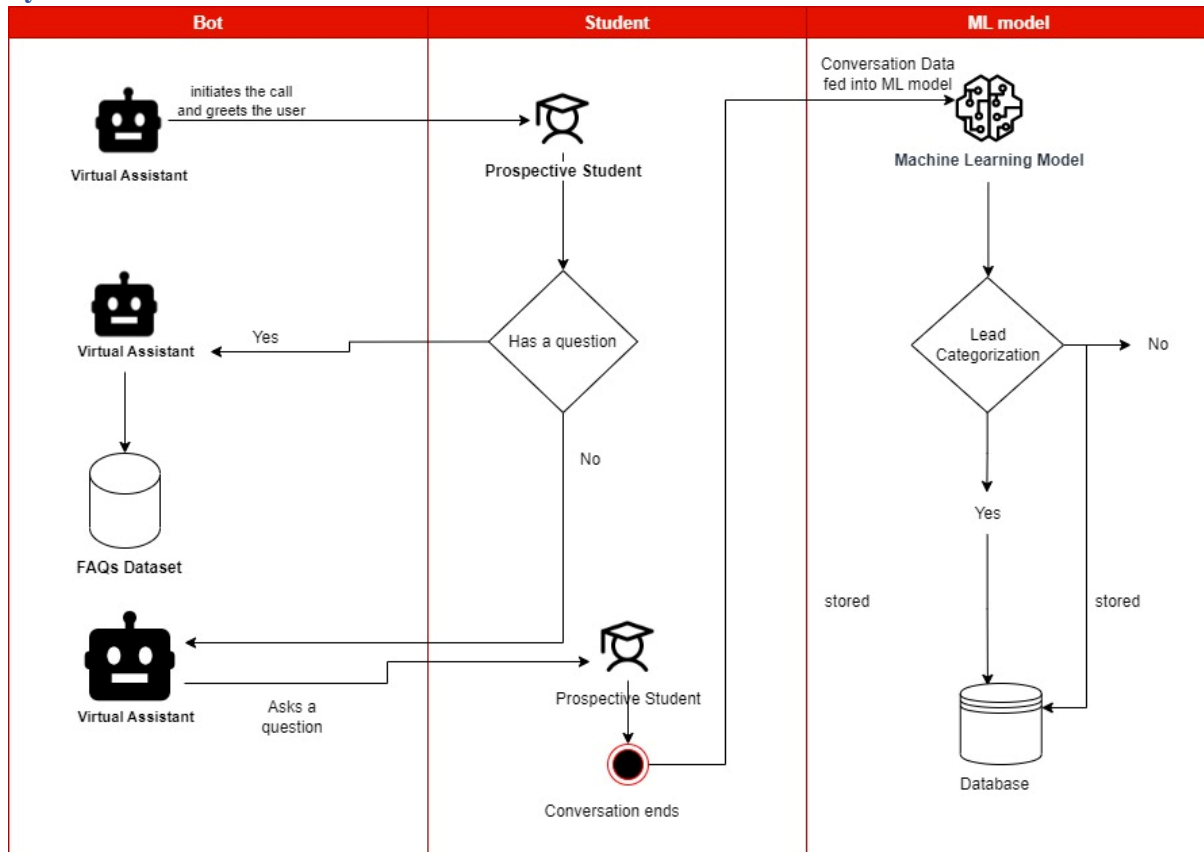
LeadVA is comprised of a virtual assistant trained using AI and a machine learning model that aims to assist UOWD's Recruitment & Admissions team with tasks that are currently performed by human resources and require significant time. It has several characteristics that help it achieve its objective. The following are the primary features:

1. Utilization of NLP to respond to user inquiries
2. Utilization of Machine Learning to generate lead scores automatically



To facilitate its implementation, this project was divided into two major modules. The first module addressed the development of the virtual assistant using Natural Language Processing. The second module focused on developing and training a machine-learning model that generates an accurate lead score.

### System architecture



### Description

The virtual assistant initiates the call and greets the prospective student (user) and inquires about their familiarity with UOWD. The user is then asked if they have any questions that will assist them in determining if UOWD is the best option for them. If the user has a question, the virtual assistant will utilize a FAQs Dataset to provide the best response possible. If the user does not have a question, the Virtual Assistant will attempt to engage in conversation. After the conversation concludes, the generated information is fed into the Machine Learning model for lead categorization.

### Technical requirements

Since Artificial intelligence provides a vast array of technologies that facilitate the automation of numerous tasks. Among those utilized for this project are the following:

- Natural Language Processing
- Sentimental Analysis
- Machine Learning (The model used was Logistic Regression)

### Platforms, Tools & Libraries

#### Platforms

Visual Studio and Jupyter Notebook were the primary testing environments for both project components.

### *Tools*

The speech-to-text module of IBM Watson was utilized. For model manipulation, pickle and glob were used.

### *Libraries*

This system was primarily programmed using the programming language Python. This was mainly due to the language's precision and user-friendliness. The following libraries were utilized:

#### Libraries for the creation of Virtual Assistant

- Numpy
- Pandas
- Lingualytics
- Textthero
- sentence\_transformers
- os
- sys
- speech\_recognition
- pyttsx3

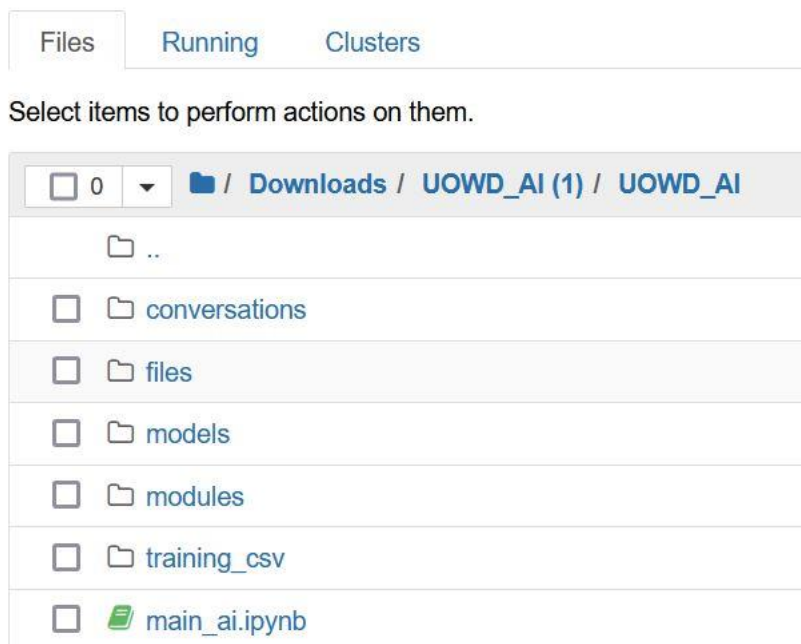
#### Libraries for the training of Machine Learning model

- sklearn
- numpy
- pandas
- matplotlib.pyplot
- seaborn

### **Code Structure**

#### *Virtual Assistant*

To facilitate organization, comprehension, and access to the technical aspects of the system, code files were systematically organized. Below is a screenshot illustrating how files and folders were organized:



### *Folders description*

**Conversations:** This folder contained text files with the student number as the filename and the entire conversation from the student's side. This file is primarily utilized for sentiment analysis and sentiment score generation. If the file already exists (indicating that the student has called before and their information already exists), it will be appended; otherwise, a new file will be created.

**Files:** This contained all the student information that the Recruitment team already has. It will allow the virtual assistant to access information such as the student's name, number, and school.

**Models:** Since there are already predefined questions and answers, it is necessary to determine if the user's question to the virtual assistant is similar to any of the predefined questions. To accomplish this, the Python library Sentence Transformers was used to generate sentence embeddings for each question. This provided a vectorized representation of each question and the questions that are like it. This was then written to a pickle file in binary format. A pickle file is simply a binary file used frequently in Python.

**Modules:** This contained the 3 python scripts which are as follows:

- FAQ\_Script.py
- Wrapper.py
- Stt.py

**Training.csv:** This contained the FAQs that are used by the virtual assistant to answer user queries.

**Main.AI:** This ran the main function that combines all the modules.



### Machine Learning model

The steps that were utilized to train the Machine Learning model were as follows:

- Import the necessary libraries

### Necessary Imports

```
1 # Importing Libraries
2 import numpy as np
3 import pandas as pd
4 import matplotlib.pyplot as plt
5 import seaborn as sns
6
7 # visualisation
8 from matplotlib.pyplot import xticks
9 %matplotlib inline
10
```

- Data Preparation

### Data Preparation

Reading in the advertising.csv file and setting it to a data frame called leads\_df.

```
: 1 leads_df = pd.read_csv('Sample DataSet.csv')
```

```
: 1 leads_df.head(20)
```

	STUDENT NO.	Staff	LAST NAME	FIRST NAME	GENDER	DATE OF BIRTH	COUNTRY ORIGIN	COUNTRY	MOBILE	EMAIL	...	END DATE	SESS START DATE	SE
0	1	Tobias	SMITH	Emma	F	22-06-08	Turkey	United Arab Emirates	9.715640e+11	hwestiii@hotmail.com	...	07-02-25	Autumn	Spr
1	2	Edgar	JOHNSON	Olivia	F	18-03-08	Turkmenistan	United Arab Emirates	9.715250e+11	retoh@outlook.com	...	07-02-25	Autumn	Spr
2	3	Jane	WILLIAMS	Noah	M	22-02-08	Uganda	United Arab Emirates	9.715230e+11	jaarnial@outlook.com	...	07-02-25	Autumn	Spr
3	4	Lyric	BROWN	Liam	M	19-02-08	Ukraine	United Arab Emirates	9.715890e+11	rogerspl@yahoo.ca	...	07-02-25	Autumn	Spr
4	5	Angel	JONES	Sophia	F	24-11-07	United Arab Emirates	United Arab	9.715070e+11	novanet@sbcglobal.net	...	07-02-25	Autumn	Spr

```
: 1 #checking duplicates
```

```
: 2 sum(leads_df.duplicated(subset = 'STUDENT NO. ')) == 0
```

```
: True
```

- Data inspection



```

19 Length of talk (Mins)      20 non-null    int64
20 Tone or sentimental score  20 non-null    int64
21 No. communications        20 non-null    int64
22 Prior awareness           20 non-null    int64
23 No. questions student     20 non-null    int64
24 No. questions VA          20 non-null    int64
25 Lead Score                 20 non-null    int64
dtypes: float64(1), int64(9), object(16)
memory usage: 4.2+ KB

```

```

]: 1 #statistical information of the data frame
   2 leads_df.describe()

```

```

]:

```

	STUDENT NO.	MOBILE	YEAR	Length of talk (Mins)	Tone or sentimental score	No. communications	Prior awareness	No. questions student	No. questions VA	Lead Score
count	20.00000	2.000000e+01	20.0	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000
mean	10.50000	9.715469e+11	2023.0	19.450000	41.250000	1.550000	0.500000	4.400000	3.050000	0.600000
std	5.91608	3.021136e+07	0.0	11.785249	23.552573	0.686333	0.512989	3.201973	1.431782	0.502625
min	1.00000	9.715050e+11	2023.0	5.000000	10.000000	1.000000	0.000000	1.000000	1.000000	0.000000
25%	5.75000	9.715238e+11	2023.0	11.000000	24.250000	1.000000	0.000000	2.750000	2.000000	0.000000
50%	10.50000	9.715495e+11	2023.0	15.500000	37.000000	1.000000	0.500000	3.000000	3.000000	1.000000
75%	15.25000	9.715692e+11	2023.0	26.250000	59.500000	2.000000	1.000000	5.000000	4.000000	1.000000
max	20.00000	9.715890e+11	2023.0	45.000000	85.000000	3.000000	1.000000	12.000000	6.000000	1.000000

```
]: 1 leads_df.columns
```

```

]: Index(['STUDENT NO.', 'Staff', 'LAST NAME', 'FIRST NAME', 'GENDER',
        'DATE OF BIRTH', 'COUNTRY ORIGIN', 'COUNTRY', 'MOBILE', 'EMAIL', 'YEAR',
        'STUDY LEVEL', 'POST GRAD.', 'FACULTY', 'COURSE OF INTEREST',
        'START DATE', 'END DATE', 'SESS START DATE', 'SESS END DATE',
        'Length of talk (Mins)', 'Tone or sentimental score',
        'No. communications', 'Prior awareness', 'No. questions student',
        'No. questions VA', 'Lead Score'],
        dtype='object')

```

```

19 Length of talk (Mins)      20 non-null    int64
20 Tone or sentimental score  20 non-null    int64
21 No. communications        20 non-null    int64
22 Prior awareness           20 non-null    int64
23 No. questions student     20 non-null    int64
24 No. questions VA          20 non-null    int64
25 Lead Score                 20 non-null    int64
dtypes: float64(1), int64(9), object(16)
memory usage: 4.2+ KB

```

```

]: 1 #statistical information of the data frame
   2 leads_df.describe()

```

```

]:

```

	STUDENT NO.	MOBILE	YEAR	Length of talk (Mins)	Tone or sentimental score	No. communications	Prior awareness	No. questions student	No. questions VA	Lead Score
count	20.00000	2.000000e+01	20.0	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000
mean	10.50000	9.715469e+11	2023.0	19.450000	41.250000	1.550000	0.500000	4.400000	3.050000	0.600000
std	5.91608	3.021136e+07	0.0	11.785249	23.552573	0.686333	0.512989	3.201973	1.431782	0.502625
min	1.00000	9.715050e+11	2023.0	5.000000	10.000000	1.000000	0.000000	1.000000	1.000000	0.000000
25%	5.75000	9.715238e+11	2023.0	11.000000	24.250000	1.000000	0.000000	2.750000	2.000000	0.000000
50%	10.50000	9.715495e+11	2023.0	15.500000	37.000000	1.000000	0.500000	3.000000	3.000000	1.000000
75%	15.25000	9.715692e+11	2023.0	26.250000	59.500000	2.000000	1.000000	5.000000	4.000000	1.000000
max	20.00000	9.715890e+11	2023.0	45.000000	85.000000	3.000000	1.000000	12.000000	6.000000	1.000000

```
]: 1 leads_df.columns
```

```

]: Index(['STUDENT NO.', 'Staff', 'LAST NAME', 'FIRST NAME', 'GENDER',
        'DATE OF BIRTH', 'COUNTRY ORIGIN', 'COUNTRY', 'MOBILE', 'EMAIL', 'YEAR',
        'STUDY LEVEL', 'POST GRAD.', 'FACULTY', 'COURSE OF INTEREST',
        'START DATE', 'END DATE', 'SESS START DATE', 'SESS END DATE',
        'Length of talk (Mins)', 'Tone or sentimental score',
        'No. communications', 'Prior awareness', 'No. questions student',
        'No. questions VA', 'Lead Score'],
        dtype='object')

```

- Exploratory data analysis

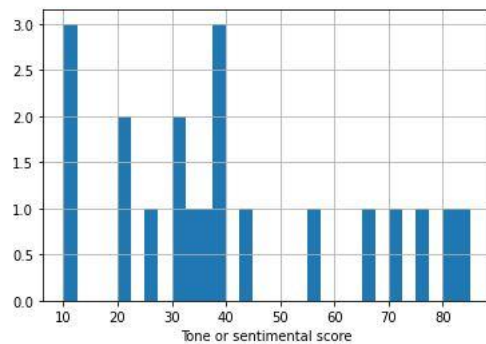


## Exploratory Data Analysis¶

1 Creating a Histogram of the Sentimental Score

```
1 leads_df['Tone or sentimental score'].hist(bins=30)
2 plt.xlabel('Tone or sentimental score')
```

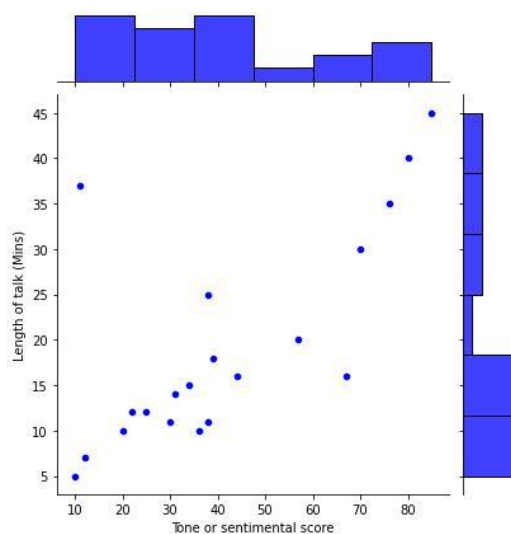
Text(0.5, 0, 'Tone or sentimental score')



Jointplot showing Sentimental Score versus Length of talk

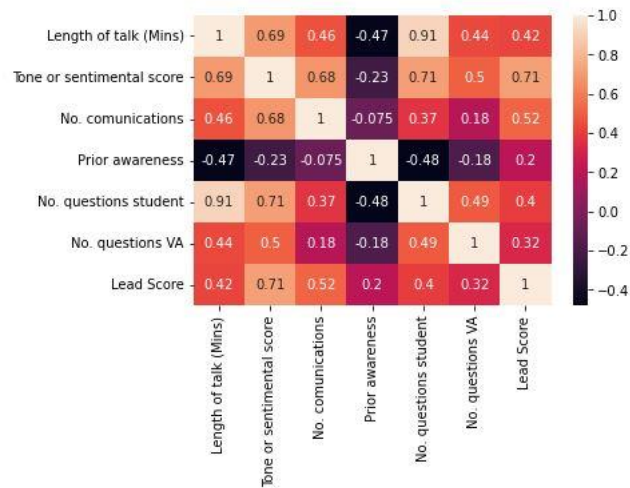
```
1 sns.jointplot(x = 'Tone or sentimental score', y = 'Length of talk (Mins)', data = leads_df, color = 'blue')
```

<seaborn.axisgrid.JointGrid at 0x23f01cc0970>



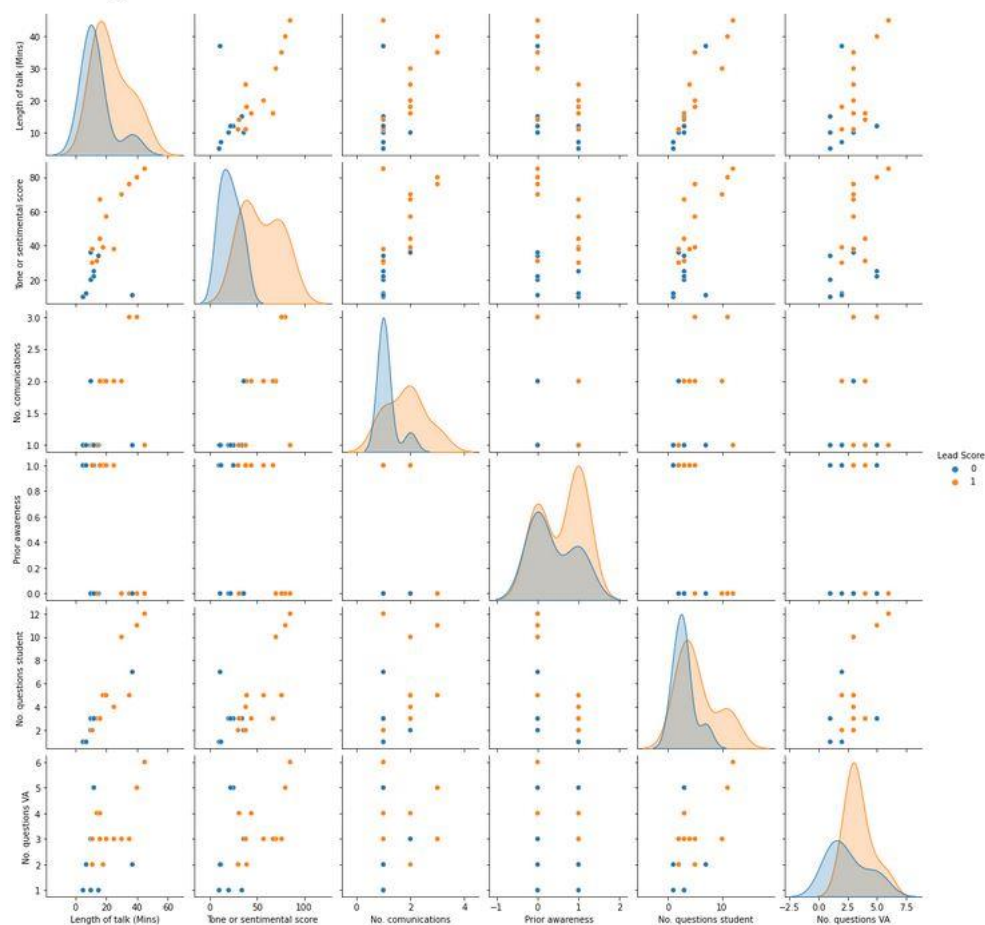
```
1 sns.heatmap(leads_df[['Length of talk (Mins)', 'Tone or sentimental score',
2                 'No. communications', 'Prior awareness', 'No. questions student',
3                 'No. questions VA', 'Lead Score']].corr(), annot = True)
```

<AxesSubplot:>



```
1 sns.pairplot(leads_df[['Length of talk (Mins)', 'Tone or sentimental score',
2                 'No. communications', 'Prior awareness', 'No. questions student',
3                 'No. questions VA', 'Lead Score']], hue = 'Lead Score')
```

<seaborn.axisgrid.PairGrid at 0x23f034c4f10>



- Training the model

## Train Test Split

Splitting our data into a training set and a testing set

```
1 from sklearn.model_selection import train_test_split

1 X = leads_df[['Length of talk (Mins)', 'Tone or sentimental score',
2             'No. communications', 'Prior awareness', 'No. questions student',
3             'No. questions VA']]

1 y = leads_df['Lead Score']

1 # Out of 20 samples of data, we use 30% for testing and remaining 70% for training
2 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

## Train the Logistic Regression Model

Train and fit a logistic regression model on the training set

```
1 # Importing LogisticRegression which is a module in linear_model library.
2 from sklearn.linear_model import LogisticRegression

1 # Declaring an instance of a logistic regression model
2 logm = LogisticRegression()

1 # After the model has been initialized, the fit() function is used to train the algo. on the training data.
2 logm.fit(X_train, y_train)
```

LogisticRegression()

- Model evaluation

## Model Evaluation ¶

Get prediction values from the model for the testing data and create a confusion matrix and a classification report.

```
1 # Predicting on the test set
2 y_predict = logm.predict(X_test)
3 y_predict
```

array([1, 1, 1, 0, 0, 1], dtype=int64)

```
1 # Importing the classification report & confusion matrix from the metrics library
2 from sklearn.metrics import confusion_matrix, classification_report
```

```
1 print(confusion_matrix(y_test, y_predict))
2
3 # 2 True positives
4 # False positives
5 # False negatives
6 # True negatives
```

```
[[2 1]
 [0 3]]
```



```

1 # PRECISION measures how many positives are predicted correctly from the total amount of positive predictions. TP/(TP+FP)
2 # RECALL measures how many positives are predicted correct from the real amount of positives. TP/(TP+FN)
3 # F1-SCORE combines precision and recall into a single measure.
4 # SUPPORT is the number of actual occurrences of the class in the dataset.
5
6 print(classification_report(y_test, y_predict))

```

	precision	recall	f1-score	support
0	1.00	0.67	0.80	3
1	0.75	1.00	0.86	3
accuracy			0.83	6
macro avg	0.88	0.83	0.83	6
weighted avg	0.88	0.83	0.83	6

## Code Documentation

### Speech To Text

Description: This module contains the code to transcribe user audio into text.

Functions:

- `transcribeAudio(self, api :str="houndify") -> str`  
 Description: Transcribe user audio into text using a specified API.  
 API List: IBM, Houndify  
 If invalid API picked, raises TRANSCRIPT ERROR  
 Returns the transcribed text if transcription successful.
- `YesOrNo(self, stt_api) -> bool`  
 Description: Gets the user input on a Yes or No question when asked a Yes or No Question

Exceptions:

ConversationAttempt

If attempts at conversation failed twice

TranscriptError

If invalid transcript API select or API cannot transcribe user audio

HangUp

If user says "bye", hangs up the call

### *FAQ Prediction*

Description: Module to predict an answer to a user query based on the question set it's been trained on

Functions:

- `train(self, csv_holder="training_csv")`  
 Using a specified parent folder containing Q&A files as CSV  
 Automatically imports them into the dataframe  
 and writes the embeddings for the question set  
 to a .pickle file.  
 This trains the model on the question set it is given
- `pred_ans(self, usr_query: str)`  
 Predicts answer based on user query  
 Question set converted to embeds in `.train()` which is loaded here  
 Convert user query to embeds and compare to question set  
 to find best match.  
 Returns a list with the answer and a success rating.  
 If question match was not enough, return 0 + "Repeat Question"  
 If question was match was satisfactory, return 1 + answer

### *AI*

Description: Combines the two modules into one and initiates calls using the provided student details.  
 Appends call data to the CSV which can be used by the ML Model to perform Lead Scoring Analysis.

Functions:

- `initConversation(self, phone_number, prior_know_bool)`  
 Initiates call using the supplied phone number. This will ask the user a set of preliminary questions to gauge if they knew about the university or not and if they have any questions they need answers for.  
 Returns call data along with a transcription of successful user communication
- `answerUserQuestions(self)`  
 Answers the user questions using STT and FAQ\_Predict
- `VAQuestionOpportunity(self, user_query)`  
 If there is an opportunity for a follow-up question to gauge if the student is interested or not, it will ask the question and record student answer
- `recordConversation(self, user_conversation, phone_number)`  
 Records the transcribed successful user communication to a .text file uniquely named by their phone number. This file is further used for sentimental analysis
- `getSentimentPolarity(self, text)`  
 Returns sentiment polarity on the given text
- `SentimentalAnalysis(self, phone_number)`  
 Opens the text file using the supplied phone number if the file exists and performs sentimental analysis to judge whether they are interested in enrolling in the university or not.
- `run_app(self)`  
 Reads the provided Student Details CSV and initiates calls and records call data to the CSV which is then used for Lead Scoring Analysis.





## Implementation (Test cases)

There are 3 test cases and the transcribed call for them can found below:

- Test 1 (User is not interested in UOWD)

---

Calling 971547000000.0  
Initiating conversation...

VA: Hello. I am calling from University of Wollongong in Dubai.  
VA: Do you have some spare time to talk to me?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
VA: Did you know about University of Wollongong In Dubai prior to this conversation?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
VA: That's alright. I can tell you a brief description about UOWD. <insert the description here im lazy>  
VA: Do you have any questions that'll help you decide if UOWD is the best choice for you?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
VA: Well, what do you have in mind?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
user query: does your university of your scholarships  
ans: Im not sure I quite understand. Can you repeat your question?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
user query: yes  
ans: Im not sure I quite understand. Can you repeat your question?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
user query: do you have scholarships  
ans: UOWD provides a range of scholarships offering financial support to new and continuing students.  
Scholarships offer a fixed percentage discount of tuition fee costs, but do not cover living and other study-related expenses.  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
VA: Bye Bye!  
VA: Thank you for your time. Have a good day!

- Test 2 (User is interested)

Calling 973000000.0  
Initiating conversation...

VA: Hello. I am calling from University of Wollongong in Dubai.  
VA: Do you have some spare time to talk to me?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
VA: Do you have any questions that'll help you decide if UOWD is the best choice for you?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
VA: Well, what do you have in mind?  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
user query: how can i sign up for clubs after i'm enrolled  
ans: If you would like to sign up, visit the Clubs/Societies section in MyUOWD, email StudentServices@uowdubai.ac.ae or call +971 4 278 1756.  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
user query: what are the fees for your transportation  
ans: To know more about fees, You may call +971 4 278 1758 or Email: transportation@uowdubai.ac.ae.  
SYSTEM: Listening...  
SYSTEM: Recognizing using: ibm  
VA: Bye Bye!  
VA: Thank you for your time. Have a good day!

- Test 3 (User may or may not be interested)



Calling 9720000000000.0  
Initiating conversation...

VA: Hello. I am calling from University of Wollongong in Dubai.  
VA: Do you have some spare time to talk to me?  
SYSTEM: Listening...  
SYSTEM: Reocognizing using: ibm  
VA: Do you have any questions that'll help you decide if UOWD is the best choice for you?  
SYSTEM: Listening...  
SYSTEM: Reocognizing using: ibm  
VA: Well, what do you have in mind?  
SYSTEM: Listening...  
SYSTEM: Reocognizing using: ibm  
user query: do you offer marketing  
ans: yes  
SYSTEM: Listening...  
SYSTEM: Reocognizing using: ibm  
VA: Bye Bye!  
VA: Thank you for your time. Have a good day!

Finally, after call ends the data gets written on a csv file. This data will then be used by the trained ML model for the categorization of Leads.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
STUDENT	Staff	LAST NAM	FIRST NAM	GENDER	DATE OF B	COUNTRY	COUNTRY	MOBILE	EMAIL	YEAR	STUDY LEV	POST GRA	FACULTY	COURSE	START DATE	END DATE	SESS STAR	SESS END	Length of	Tone or se	No. comu	Prior awareness	No. questions	No. question	Lead Score	
1	Tobias	SMITH	Emma	F	22/06/20C	Turkey	United Arc	9.72E+11	hwestil@	2023	High schor	No	Business, I	Accountar	#####	7/2/2025	Autumn	Spring	1.049852	0.051667	1	0	1	0	0	
2	Edgar	JOHNSON	Olivia	F	18/03/20C	Turkmenis	United Arc	9.73E+08	retoh@ou	2023	High schor	No	Business, I	Accountar	#####	7/2/2025	Autumn	Spring	4.625275	0.005556	3	0	13	4	0	
3	Jane	WILLIAMS	Noah	M	22/02/20C	Uganda	United Arc	9.72E+12	jaarnial@c	2023	High schor	No	Business, I	Accountar	#####	7/2/2025	Autumn	Spring	3.700528	0.089286	2	0	6	1	0	

### Limitations & Constraints

During the implementation of any system, complications are inevitable. Among those encountered during the LeadVA execution are the following:

#### *Limited choices for choosing a speech-to-text model*

This was due to restricted funding & resources, and the choices for choosing a speech-to-text model were minimum. Hence this wasn't giving the accuracy that was aimed for.

#### *Low training dataset for the virtual assistant*

The dataset for the training of the virtual assistant was minimum, hence why the existing model may not be able to answer all user queries.

#### *Unavailability of real training data for lead score generation*

Another constraint faced during the implementation was the unavailability of real training data for the Machine Learning (ML) model that was used to generate lead scores. As company data is supposed to be confidential, a fake dataset was made by using resources such as Kaggle to train the ML model.

#### *Virtual assistant only understands the English language*

UOWD hosts 108 nationalities as current students therefore, it would be inevitable that prospective students could come from various other nationalities. A limitation of the current system is that it caters to only English-speaking users and does not support other languages.

### Future improvements

In order to improve the current LeadVA system, the following future enhancements are recommended.

*Request for funding*

To increase the accuracy of LeadVA's Virtual Agent, the appropriate funding will be requested by the relevant parties so that the speech-to-text model's accuracy-maximizing options can be expanded.

*Increase in training dataset for the virtual dataset*

The UOWD website will be scraped in order to store its data in a.csv file. This data will serve as a library from which the virtual assistant can learn and respond to user queries accordingly. The data will include all pertinent information regarding academics, student life, sports, transportation, visas, and careers, among others.

*Training Virtual assistant with different nationalities language*

This will expand the scope of our project by enabling additional users to interact with LeadVA. This will benefit both the users and the UOWD Recruitment & Admissions Team.

*Expansion of LeadVA to current student queries*

Since UOWD also receives numerous inquiries from current students on a daily basis, a large number of personnel are employed to respond to these. Consequently, the CSCI323 Project Team acknowledged that expanding their solution would also benefit other university departments.

**Conclusion**

Overall, the implementation of the project was a success despite the limitations and difficulties that were faced which can be fixed through proper funding as well as access to the datasets needed for training; and of course, a larger span of time can allow more training, thus more accuracy and improvements to LeadVA's performance. Furthermore, the proposal has intrigued the client I.e UOWD's Recruitment Team, and in fact solved their main issue when trying to secure potential students. One of the Recruitment Team members had said the following, "The success of your project would actually save us a lot of time". The fact that only a few employees were initiating thousands of calls a month with potential students was time consuming. Not only that, but also the manual selection of every student as a lead or not consumed time as well. Thus, when LeadVA solution was presented to them, they immediately took it into consideration, and wanted to deploy it, as LeadVA will be able to initiate every call instead of the employees and determine itself whether the potential students are a lead or not.