UCL
SCHOOL OF
MANAGEMENT

MSIN0221 Natural Language Processing

Group Assignment - Group 5

Part 1: Project Proposal

Problem Statement

Problem

UCL School of Management (SoM) currently appoints 12 students from undergraduate and postgraduate studies as ambassadors every academic year to answer questions from prospective applicants via a web chat. Ambassadors are encouraged to reply within 3 working days depending on their availability and UCL shall pay them an hourly compensation in return.

Proposed Solution

The current practice can be improved by building a Q&A chatbot, that would respond immediately to prospective applicants regarding generic questions.

Business Value

Over 30,000 enquiries have been made since March 2018, meaning each ambassador receives up to 2 questions per day (UCL School of Management, 2022a). After consulting one of the ambassadors, it is noted that most questions asked were regarding generic admission requirements. It is believed that the implementation of a Q&A chatbot can improve the application experience by providing immediate answers to prospective applicants whilst saving time and costs by reducing the workload of student ambassadors.

This project aims to explore, how accurately can a Q&A chatbot answer enquiries from prospective students regarding content on the degree pages of the UCL SoM website?

Literature Review

Chatbots are a smart solution that provides information for enquiries on demand (Adamopoulou and Moussiades, 2020). Having a chatbot on the SoM website will be more efficient when answering general enquiries, since chatbots are available 24/7 whilst being able to carry out multiple conversations simultaneously; it complements the shortcoming of human factors and enables student representatives to focus on answering interpersonal questions (Kulkarni, Bhavsar, Pingale and Kumbhar, 2017).

Mikic et al. examined the effects of an AIML-based chatterbox (Charlie) using natural language to communicate for educational purposes, where chatbot Charlie provides students with material and specific learning content whilst finding impactful contributions to overall performance.

The benefits of AI-based chatbots are multifaceted; overall availability is advantageous for international users who want immediate answers to their questions as well as cost reduction due to simplicity and efficient use of resources (Satu and Parvez, 2015).

Data Sources

The training data consists of 10 web pages on the SoM website, each containing details on a single degree (UCL School of Management, 2022b). This text is publicly accessible and available, although all contents are protected by intellectual property rights and owned by the SoM. The Acceptable Use Policy requires permission to be acquired when using the content for any purpose other than personal (UCL Extend, 2021). Therefore, the SoM has been contacted and permission has been granted. Furthermore, sample prospect questions will be gathered by an ambassador for training purpose.

Upon approval to use the website content, the text will be web scraped. Currently, no existing labels exist in the dataset; all data must be labeled manually. Labeling will be performed using a free tool called tagtog and distributed across the 7 team members as there is a relatively small amount of raw text (tagtog, 2022). To ensure entity labeling is consistent, internal documentation on rules will be created and followed; inter-rater reliability can be used on text samples to ensure labeling is consistent.

Our Approach

This project will go through the five major phases as detailed below:

1. Data Preprocessing

The spaCy library will be leveraged for the natural language processing to allow English as a human language to transform into the machine language. To avoid any unnecessary processing and word duplication that could reduce accuracy, the following tasks will be carried out - dealing with capitalization and special characters, stopword removal, stemming and lemmatization and tokenization.

2. Feature Engineering

To represent texts in a way that the machine learning model can interpret, three core techniques will be performed to transform the text data into numerical form.

- Bag-of-Words: it measures the occurrence of words in the dialogue
- TF-IDF: it evaluates the importance of a word based on its relevancy in the given dialogue
- Word Embeddings: it measures syntactic and semantic word similarities in the given dialogue

In addition, the dataset will be divided into 3 sets, i.e., train, development and test on the ratio of 8:1:1 for model training and validation purposes.

3. Model Development & Training

An existing model named ***BERT-tiny-5-finetuned-squadv2*** has been chosen as the baseline model for this project, mainly due to two reasons:

- It has the SQuAD dataset, a reading comprehension dataset based on a set of Wikipedia articles, trained already which is simple to start with and allows us to achieve solid performance in a short timeframe (Google Research, 2020).
- It is designed to be used in lower-processing environments, which is suitable to run on the faculty environment or personal computers (Google Research, 2020).

The model is open source and available in the AI community *Hugging Face*. It has built-in pre-processing capability, which can be run either via spaCy or other libraries depending on the specifications. When fitting in the dataset, a consistent library will be employed for pre-processing tasks.

4. Model Assessment

In this phase, the model performance will be evaluated by looking into the accuracy score and performing an error analysis.

Firstly, manual annotation for all questions and their answers will be carried out by the team. Based on the F1 and accuracy scores, the model will be further fine-tuned if necessary.

Secondly, an error analysis will be carried out to examine the wrong predictions and discover the patterns, which could help further understanding how the model performs and inform the direction of fine-tuning. Depending on the findings, there are several fine-tuning methods that can be adapted, e.g., including more data in the model training, adjusting the pre-defined hyperparameters in the model, and reviewing the syntactic processing methods.

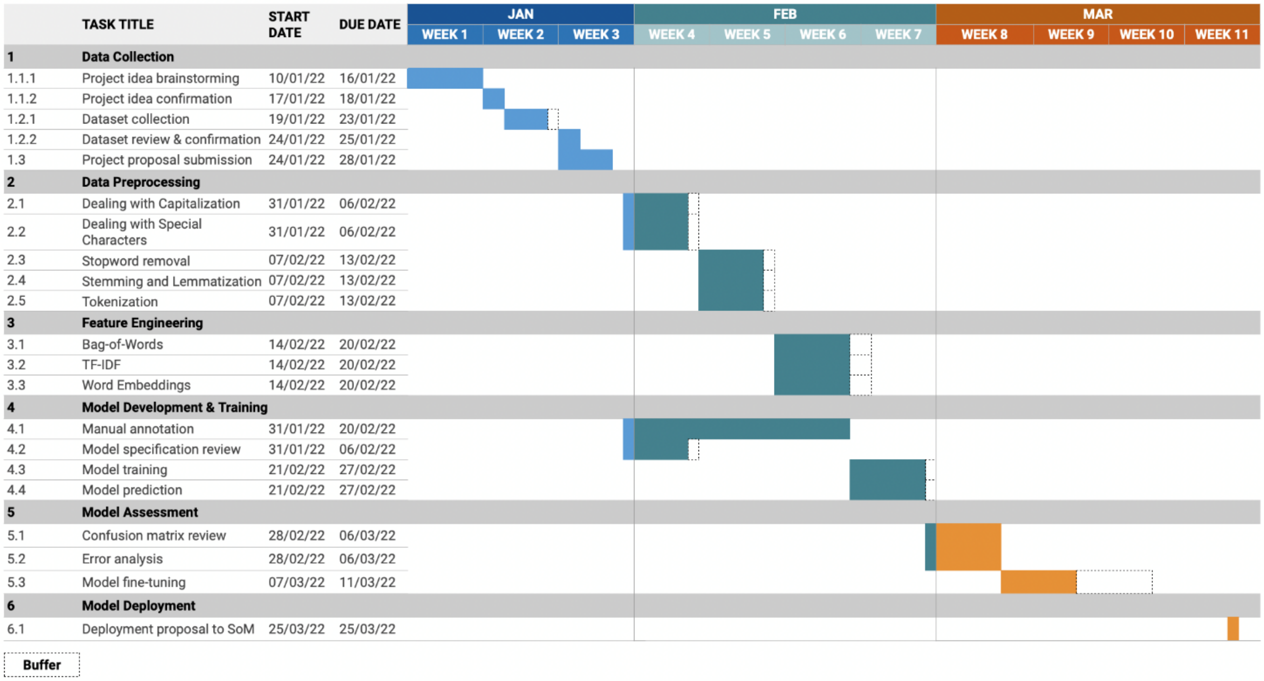
5. Model Deployment

A fine-tuned model based on BERT using a customized set of hyperparameters and pre-processing steps for the dataset will be integrated into the existing chatbot page. This ensures greater performance for this specific task. As no significant amount of new enquires are received daily, batch learning is proposed as it is simpler to maintain and more cost-efficient.

The first phase of deployment is to roll out the chatbot to a selection of SoM degrees to test the model's performance in the production environment. If there is room for improvement, the model will be revisited to go through additional fine-tuning. Otherwise, it will be deployed to all other degrees.

Project Roadmap

The following Gantt chart shows the progression of the work across the project's duration.



	TASK TITLE	START DATE	DUE DATE
1	**Data Collection**		
1.1.1	Project idea brainstorming	10/01/22	16/01/22
1.1.2	Project idea confirmation	17/01/22	18/01/22
1.2.1	Dataset collection	19/01/22	23/01/22
1.2.2	Dataset review & confirmation	24/01/22	25/01/22
1.3	Project proposal submission	24/01/22	28/01/22
2	**Data Preprocessing**		
2.1	Dealing with Capitalization	31/01/22	06/02/22
2.2	Dealing with Special Characters	31/01/22	06/02/22
2.3	Stopword removal	07/02/22	13/02/22
2.4	Stemming and Lemmatization	07/02/22	13/02/22
2.5	Tokenization	07/02/22	13/02/22
3	**Feature Engineering**		
3.1	Bag-of-Words	14/02/22	20/02/22
3.2	TF-IDF	14/02/22	20/02/22
3.3	Word Embeddings	14/02/22	20/02/22
4	**Model Development & Training**		
4.1	Manual annotation	31/01/22	20/02/22
4.2	Model specification review	31/01/22	06/02/22
4.3	Model training	21/02/22	27/02/22
4.4	Model prediction	21/02/22	27/02/22
5	**Model Assessment**		
5.1	Confusion matrix review	28/02/22	06/03/22
5.2	Error analysis	28/02/22	06/03/22
5.3	Model fine-tuning	07/03/22	11/03/22
6	**Model Deployment**		
6.1	Deployment proposal to SoM	25/03/22	25/03/22

References

Adamopoulou, E. & Moussiades, L., 2020. An Overview of Chatbot Technology. Artificial Intelligence Applications and Innovations, pp. 373-383.

Google Research (2020) BERT-Tiny (5) fine-tuned on SQuAD v2, Hugging Face. Available at: https://huggingface.co/mrm8488/bert-tiny-5-finetuned-squadv2 (Accessed: 24 January 2022).

Kulkarni, C. S., Bhavsar, A. U., Pingale, S. R. & Kumbhar, P. S. S., 2017. BANK CHAT BOT – An Intelligent Assistant System Using NLP and Machine Learning. International Research Journal of Engineering and Technology, pp. 2374-2377.

Mikic, F. A., Burguillo, J. C., Llamas, M., Rodríguez, D. A., & Rodríguez, E. (2009, June). Charlie: An aiml-based chatterbot which works as an interface among ines and humans. In 2009 EAEEIE Annual Conference (pp. 1-6). IEEE.

Satu, M. S., & Parvez, M. H. (2015, November). Review of integrated applications with aiml based chatbot. In 2015 International Conference on Computer and Information Engineering (ICCIE) (pp. 87-90). IEEE.

tagtog (2022) The Text Annotation Tool to Train AI. Available at: https://www.tagtog.net/ (Accessed: 24 January 2022).

UCL Extend (2021) UCLeXtend terms and conditions and website acceptable use policy. Available at: https://extend.ucl.ac.uk/admin/tool/policy/view.php?versionid=1&returnurl=https%3A%2F%2Fextend.ucl.ac.uk%2F (Accessed: 23 January 2022).

UCL School of Management (2022a) CHAT TO OUR STUDENTS. Available at: https://www.mgmt.ucl.ac.uk/chat-to-students (Accessed: 25 January 2022).

UCL School of Management (2022b) STUDY. Available at: https://www.mgmt.ucl.ac.uk/study (Accessed: 23 January 2022).