Interim Project Report

# Ethical Consideration

No human subjects were or will be involved in the production of this project, and there is no requirement for Ethical Approval.

# Completed work (up to 600 words)

My aim for this project started off as simply to build a system to help my colleagues access information, however, to find a way to do that which was realistic and relevant proved challenging. After initially considering an application based on a search engine, I considered what existing systems could be complemented with new technology and hit upon the idea of constructing an email ‘conversational agent’ that could automatically respond to and process technical support requests. The system underpinning this would be an ‘intention classifier’ – a model designed to classify incoming email according to categories of user intent. The research and construction of this intention classifier is the focus of this project. Additional secondary systems could easily be built around this to execute and action the responses, perhaps using dialogue trees and SQL queries.

Once this idea crystallised, I had to consider which technologies were most appropriate to the task. Artificial Intelligence today is a diverse field, with many sub-divisions and methodologies to choose from. On offer were a variety of technologies ranging from (now obsolete) rules-based implementations, through to shallow learning methods like K-Nearest Neighbours (KNN), and on into more current Neural Network models such as Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRU). I wanted to assess the viability of these technologies in practice, in the context of a typical work environment. As industrial SME that does not specialise in software or data, my employer (Origin Frames ltd) is just such an example. A system implemented in this context would need not only to deliver high quality results, but also require little maintenance and have a low Total Cost of Ownership. By this definition, a rules-based approach was instantly dismissed due to its high level of set up and maintenance costs combined with poor results. However, a more intriguing question existed over which of the more modern methods produced better results. Neural Networks seemed more likely to produce better results, but if a simpler and more maintainable KNN classifier could deliver similar quality then this could also be a desirable option. To answer this question, I needed to build a number of these models in order to test and evaluate them.

Before I could begin working on any classification models, I first had to obtain a data set. Initially this was a simple matter of obtaining my employer’s permission and exporting the data. However, this quickly became a weighty task of cleaning, formatting, anonymising and classifying over 700 records. The data was cleaned by considering deterministic tests of emails which should not be responded too – for instance emails from outside the company’s domain name, or from within the support team (who often email the support mailbox to inform their colleagues of relevant but non-actionable information). I formatted the data by combining the subject and body of the email, removing email signatures, and anything but the latest message in the thread. Anonymisation was trickier, involving removal of any personally identifiable or contractual information. I was able to retain meaning by generalising to generic tags, for instance <NAME>, <PRODUCT> or <COMPANY>.

Up to this point, I was able to use automated processes and data from the email export itself to do most of the processing automatically. This was a positive result not just for my labour time, but also because the final system will need to follow the same processes perfectly if it is to be meaningfully implemented. Additionally, generalising the data may prove useful as with a relatively small dataset of only ~700 emails, the appearance of unique or sparse names will prove difficult to train on. However, I needed to go over the data personally at least three times to assure myself of complete anonymisation before commencing work. Additionally, the classification process required significant manual handling, although on reflection I could have made my work easier by implementing a K-means clustering algorithm to group similar emails.

Building models themselves has proved comparatively straightforward. I have been buidling KNN models using the Sci-Kit Learn library within only a few lines of code, likewise, building RNN models is only a little more difficult, using the Keras library.

# Work to be done (up to 600 words)

Currently I find myself optimising my hyperparameters to drive better results for these algorithms to build better models. I have been using cross validation to measure results and drive improvements, as my data set is too small to meaningfully break into training and test sets. As part of this work I am considering how best to score predictions. A multi-class confusion matrix seems ideal to analyse how the models perform in individual classifications, however, I believe this could be better aggregated into a single numeric score.

Two common approaches to this include ‘Precision’ and ‘Recall’, which involve measurements based on different functions of four categories of results: True Positive, False Positive, True Negative, False Negative. In this instance a False Positive (sending a response which should have been ignored, or isn’t correct) is much more costly than a false negative (ignoring an email when an automated response or action would have been appropriate) as a false positive causes confusion, may introduce errors into production data, and requires extra work to fix and follow up with a correct response. Meanwhile a false negative merely requires the support team to action the email correctly. Therefore, I am intending to score my models on Precision, rather than Recall.

Additionally, the results should be normalised so that only model predictions with a high degree of certainty result in actions taken by the secondary systems. Grouping the results into the four categories listed above should take this into account. The four categories can thus be defined based on how a secondary system would have used the classification results. This will require consideration for which classifications can neatly be handled in an automated fashion and therefore should be actioned automatically, and which are too complex or physical and must be handled manually by the support team. Once complete, this scoring work will be useful not just in cross-validation but also evaluation of performance between different architectures.

Using the cross-validation process will enable to me optimise hyperparameters more effectively, by enabling me to make objective design decisions. For instance, the hyperparameter *k* of a KNN model can be optimised by iteratively building KNN models with increasing values of *k* and then plotting k

In order to accomplish my goals, I will need to build more models, based on LSTM and GRU algorithms. It is my intention to use the same Keras library that I used to build my RNN models, with which to build LSTM and GRU models. I anticipate that the implementation of these algorithms in code will be similarly complex.

Finally as the project draws to a close I will approach my employer for a second data export, with which to validate my results. Having then built a number of models using these technologies, I will be able to assess their suitability for use in a commercial environment, by considering ease of deployment and maintenance as well as performance in the form of precision.

Marks will be awarded for:  
• Quality and amount of practical work done on the project itself, as judged from the report (3 marks).  
• Discussion of issues relating to work done and planned to be done on the project (2 marks).  
• Description of problems you have faced and overcome, and other challenges that remain to be resolved (2 marks).  
• Presenting the issues in a logical, well ordered way (2 marks).  
• Readability, grammar and spelling (1 mark).