**UG Project Plan**

**CSC-30014**

### Project Overview and Description

**Student Name:** Richard Jones

**Student Username:** w8a76

**Student Number:** 17011180

**Degree Title:** Computer Science

**Supervisor Name:** Dr Charles Day

**Project Title:** Detecting Political Bias in Text

**Please provide a brief Project Description:**

|  |
| --- |
| The project will focus around the design of a system that can process a . For each article, the system will use sentiment analysis (most likely through a Long-Short Term Memory NN) in an attempt to detect opinionated bias towards their contents, assigning a single value based on the article’s overall political bias.  The output of the system will be recorded in a database for later querying, and an interface will be constructed around the system for a user to interact with the Neural Network. |

**What are the aims and objectives of the Project?**

|  |
| --- |
| The project aims to design, develop and train a neural network system to recognise political bias in selections of text decided by an end-user.   1. Design a Deep Learning System capable of taking a sentence as an input and indicating one of several outputs. 2. The system created shall be trained using text-based data with its political ideology labelled as an output. The system will then be validated with similar data.    * Labelled data from the Hansard Parliamentary archives will be used to train these networks. 3. The project shall store key information on the system’s output in an efficient and easy to access manner. Specifically, the text inputted into the system, as well as its resulting outputs. 4. The project shall provide an Interface that allows a user to interact with the developed system.    * The interface shall allow the user to upload an article of their choice and store it in the system (see Aim 3)    * The interface shall pass the uploaded article into the System and retrieve the appropriate outputs (e.g. political ideology of each sentence, political ideology of the article as a whole). Only the data for the article as a whole will be written into the database.    * The interface shall then produce an analytics report on the political sentiment of the article for the user to read. |

**Please provide a brief overview of the key literature related to the Project:**

|  |
| --- |
| *See reference at bottom of the page for full Harvard reference.*  Political Ideology Detection Using Recursive Neural Networks (Iyyer, et al., 2014)  This paper investigated the use of Recursive Neural Network techniques to detect Political Ideology in text. The paper focused on American Political Ideology and trained it’s NN through a wide selection of datasets (Such as Convote & the Ideological Books Corpus).The fully trained RNN in question was able to outperform other more popular learning techniques (such as bag-of-words and word2vec baselines, though word2vec was also experimented with in the developed RNN). The RNN was able to achieve a 70.2% and 69.3% validation-accuracy on the Convote and IBC datasets respectively.  Political Bias Analysis (Misra & Basak, 2016)  Building on Iyyer’s Paper (Key Literature No. 1), Misra and Basak developed a Recursive Neural Network that implemented a more Specialised Long Short-Term Memory Network framework. This paper was once again performed on US Political data. However, instead of ranking the performance of the trained Network on its Test-Data Accuracy, Misra and Basak assessed their system using a F1 score, which could be argued to be a better classification of the task’s success.  Multi-view Models for Political Ideology Detection of News Articles (Kulkarni, et al., 2018)  This paper focuses on detecting political bias in news articles using a multi-view document assessment model approach, meaning that more then just the text of the news article was used to assess its ideology. The other content included the article’s title as well as the content of the hyperlinks it includes in its text. The report used labelled data from allsides.com, which is a user-labelled data set of news articles from 59 US-base sources. The MVDAM designed used a convoluted neural network to assess the title, a hierarchical document attention model (HDAM) to classify the contents and a feed-forward neural network to assess the article’s hyperlinks. The results of these engines are then processed through several layers of the MVDAM. As a result, the model significantly outperformed other state of the art solutions by approx. 10% on the F1-Score (which averaged out at about 79.67, compared to the 68.92 achieved by the CHANCE classification).  LSTM neural network for textual n-grams (D'Souza, 2018)  This paper aims to design a system using LSTMs and n-grams that can be used for language prediction (as well as generation). Using smoothed and processed data from the Penn Tree Corpus dataset (PTB), the developed system was trained using multiple layouts as well as switching out its activations functions (through sigmoid, softmax and hyperbolic-tangent). The final solution comprised of a MLP Network that predicted semantic similarity, which then was passed into an LSTM layer that generated the predicted word sentences. Using n-grams and the Continuous bag-of-words model, the system was successfully able to predict sentences from the PTB (Accuracy not specified). |

### Project Process and Method

**Please provide a brief overview of the Methodology to be used in the Project (inc. an overview of best practice within the Methodology):**

|  |
| --- |
| In order to ensure that every section of the project was completed in an organised and timely manner, an agile methodology shall be adopted. More specifically, the SCRUM framework shall be used as a main guideline. As I have had prior experience in this framework from previous projects, it will mean less time will be used up with coming to terms with a new methodology and more time will be used on the project. Furthermore, the implementations of sprints and product backlogs will allow the project to run in a more organised and well thought out manner.  In order to allow for some performance tuning, some sections of the project will require prototyping to be performed, making this project run on a more Quasi-Agile methodology. |

**Which Data Collection Methods will be employed (e.g card sorts, questionnaires, simulations, …)?**

|  |
| --- |
| The NN will be trained using labelled debate data from the Hansard Parliamentary Database:   * The debate data will be retrieved from the ParlParse XML data: <https://parser.theyworkforyou.com/hansard.html> * The debate will be labelled depending on the political party of the speaker. * This will then be organised in a way for the developed NN to process. |

**Briefly describe how you will ensure your project is in line with BCS Project Guidelines (BSc Computer Science Single Honours Students only)?**

|  |
| --- |
| 1. **Elucidation of the problem and the objectives of the project.**    1. This can be found on page 1 in the aims/objectives section. 2. **An in-depth investigation of the context and literature, and where appropriate, other similar products.**    1. This can be found on page 2 in the key literature section. 3. **Where appropriate, a clear description of the stages of the life cycle undertaken.**    1. A Gantt chart was created to identify all of the necessary stages of this project’s lifecycle. It can be found on page 6. 4. **Where appropriate, a description of how verification and validation were applied at these stages.**    1. Like Guideline 3, the Gantt chart shows the weeks in which each stage of the lifecycle will be fully completed, showing verification and validation. 5. **Where appropriate, a description of the use of tools to support the development process.**    1. This can be found in the Resource Planning section starting on page 4 6. **A critical appraisal of the project, indicating the rationale for any design/implementation decisions, lessons learnt during the course of the project, and evaluation (with hindsight) of the project outcome and the process of its production (including a review of the plan and any deviations from it).**    1. The appraisal will exist throughout the analysis, design and evaluation sections of the project report. 7. **A description of any research hypothesis**    1. The description of the research can be located on page 1 in the Project Description Section. 8. **References**    1. All references can be located in the references section of this plan (Page 6). |

### Time and Resource Planning

**Will Standard Departmental Hardware be used?** YES

**If NO please outline the Hardware/Materials to be used:**

|  |
| --- |
| The School’s Lab Computers will be used to develop, train and test the system. |

**Will Software which is already available in department be used?** YES

**If NO please outline the Software to be used including how any necessary licences will be obtained:**

|  |
| --- |
| Python 3.7 – Open-Source  Tensorflow 2.0 Library for Python – Open-Source  Keras for Python – Open-Source  PyCharm IDE – Free Educational License |

**Will the project require any Programming?** YES

**If YES please list the (potential) Programming Languages to be used (including any IDEs and Libraries you may make use of):**

|  |
| --- |
| Python 3.7 – Open-Source  Tensorflow 2.0 Library for Python – Open-Source  Keras for Python – Open-Source  PyCharm IDE – Free Educational License |

**Table of Risks (*if non Standard Hardware and/or Software to be used please include backup options/ contingency plans here)*:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Risk** | **Description** | **Impact (High-Low 1-5)** | **Countermeasures** | **How can the risk be overcome?** |
| Project is Lost | The drive containing the project data is lost or destroyed from accidental means | 1 | The Project will be kept on Cloud Storage (Google Drive), GitHub and backed up to the School drives every week. | When the Data is lost on the Primary Machine, the most up-to-date backup will be pulled |
| Project is not finished | Due to other commitments and time-management, certain objectives of the project may not be completed. | 1 | If unavoidable, Low priority objectives will be missed out and the solution will be completed with reduced functionality. | Use the Gantt Chart provided in this plan to adhere to a schedule. Also readjust the chart if some objectives take longer/shorter than expected. |
| Neural Network becomes overfitted | The Neural Network is trained for too long, where it will not generalise and begin learning the datasets. | 2 | During the training of the network, the generalisation error rate will be monitored. In the event that the rate increases, the training will be halted. | The network will be developed over several iterations. During this time, the number of epochs will be experimented with to ensure the generalisation error rate does not increase. |
| Initial Neural Network is not optimised for detecting Political bias | The Neural Network designed may not be able to | 3 | / | In order to ensure that the network is optimised to give the project its best chance of success. Different NN layouts will be tested and evaluated best on their effectiveness. |
| An optimum Learning Rate cannot be found. | An optimal or acceptable value cannot be found for the learning rate (too small – NN training won’t be effective, too large – Can lead to overfitting). This would lead to un-ideally trained neural network | 3 | Applying a decaying/scheduled learning rate will reduce the probability of this risk occurring. | Once again, the project will undergo several iterations, for which different NN layouts will be tested and evaluated, Meaning that the project will have several different chances at success. |

**Gantt Chart/ Pert Chart (must include milestones and deliverables):**

### References and Administration

**Please include a list of References used in this Plan:**

|  |
| --- |
| Bibliography D'Souza, S., 2018. *LSTM neural network for textual ngrams,* s.l.: PrePrints.  Iyyer, M., Enns, P., Boyd-Graber, J. & Resnik, P., 2014. *Political Ideology Detection Using Recursive Neural Networks,* Maryland: University of Maryland.  Kulkarni, V., Ye, J., Skiena , S. & Wang, W. Y., 2018. *Multi-view Models for Political Ideology Detection of News Articles,* Brussels: Association for Computational Linguistics.  Misra, A. & Basak, S., 2016. *Political Bias Analysis,* Stanford: Stanford University. |

**Submission Date:** 07/11/2019

**PLEASE NOTE THAT SHOULD YOUR PROJECT UNDERGO ANY MAJOR CHANGES FOLLOWING THE SUBMISSION OF THIS PLAN YOU ARE EXPECTED TO SUBMIT AN UPDATED PLAN WHICH ACCURATLEY REFLECTS YOUR PROJECT.**