Part II Project Progress Report

Louis Max Cley Slater

February 1, 2018

Project Information

Title: Sentiment Analysis of Texts on the Iraq War
Email: lmcs2@cam.ac.uk
Supervisor: Dr Tamara Polajnar
Director of Studies: Dr Anil Madhavapeddy
Overseers: Dr Timothy Griffin & Professor Anuj Dawar

Project Progress

Initially, I planned to use a 2003 study by Robinson, Goddard, Brown and Taylor in which they manually assessed the biases of British Newspaper articles on the Iraq War. As stated in my project proposal, using this dataset required me to have access to the newspaper archives referenced in the study. At the time of submitting the project proposal, I had found multiple archives containing the information I needed, all of which the University had licences for. Shortly after submitting the project proposal, I discovered that these licences restricted access to the newspaper archives, meaning that I couldnt text mine the quantity of data I would require from them and subsequent analysis would be unlicensed. After enquiring about the cost of a license to access the data that would fit my needs, the cheapest quote I received was $20,000, so accepted that I had to find another dataset.

After some more research, I decided to use a set of transcripts from debates in the House of Commons as my corpus, in part due to its availability under the Open Parliament Licence meaning that I could acquire and analyse the data in great quantities. After extensive research, I found no other projects in which these transcripts had been used as a corpus for computational techniques, which motivated me further to pursue the use of this data.

Unfortunately, changing the dataset I was using meant that the project was delayed and I had to establish a new timeline for it. Part of the difficulty was in the fact that the parliamentary transcripts were not in an easy form to deal with, so parsing them to compile a database took longer than compiling the database using the original dataset would have. Despite this, I put extra work on to my project over the Christmas vacation and am now only about two weeks behind where I planned to be in my initial project timetable and hope to catch up through extra work this term.

So far, I have parsed the transcripts of all the House of Commons debates between September 11th 2001 and 18th March 2003 and collated that data with all House of Commons votes in the same period to compile a relational database. After this, I wrote an email spam detector using a bag-of-words model and a support vector machine, then adapted this program to detect the stance of MPs on the Iraq War, using the database I previously developed. Through doing this, I learned a lot about the techniques I was using and ensured that I had not hard-coded anything into the political stance detector, thereby ensuring I was maintaining good practices for implementing machine learning algorithms. Since completing this basic approach, I have worked on implementing other NLP techniques such as stop-word removal and stemming to improve upon the accuracy of the classifier.