Hansard Historical Sentiment Analysis and Comparison

Final Report for CS39440 Major Project

Author: Adam Neaves (adn2@aber.ac.uk)
Supervisor: Dr. Amanda Clare (afc@aber.ac.uk)

4th March 2018
Version: 1.0 (Draft)

This report was submitted as partial fulfilment of a BSc degree in Artificial Intelligence and Robotics (GH7P)

Department of Computer Science
Aberystwyth University
Aberystwyth
Ceredigion
SY23 3DB
Wales, UK

Declaration of originality

I confirm that:

- This submission is my own work, except where clearly indicated.
- I understand that there are severe penalties for Unacceptable Academic Practice, which can lead to loss of marks or even the withholding of a degree.
- I have read the regulations on Unacceptable Academic Practice from the University's Academic Quality and Records Office (AQRO) and the relevant sections of the current Student Handbook of the Department of Computer Science.
- In submitting this work I understand and agree to abide by the University's regulations governing these issues.

Name ....

Date ....

Consent to share this work

By including my name below, I hereby agree to this dissertation being made available to other students and academic staff of the Aberystwyth Computer Science Department.

Name ....

Date ....

Acknowledgements

I am grateful to...

I’d like to thank...

Abstract

Politics affects all aspects of a persons life. The results of debates in Parliament may have a profound influence on an individuals life, and knowing how your local MP speaks and the opinions they express during these debates could prove useful. Websites, such as https://www.theyworkforyou.com, show a user how their local MP votes, and how often they attend debates, ask questions, and other information that may allow the user to make informed decisions when voting during elections.

The aim of the Hansard Sentiment Analysis Tool is to provide a tool to detect the sentiment of statements made during political debates, and to compare it with sentiment expressed about the topic previously, or compare it with sentiment expressed by the same MP. The sentiment analysis will use a machine learning approach, where a model will be trained on labelled datasets generated from the source data.

Being able to view how sentiment changed over time, expecially about certain topics, could prove useful for a user who wants to not only know how an MP votes, but also how they represent themselves and their constituency in parliament. If an MP is seen to change opinion on a topic as time passes, this information could be used to keep voters informed.

This report will document the process of designing and developing the Hansard Sentiment Analysis Tool, highlighting any challenges encountered during development, and also the results of the technical work.

CONTENTS

1 Background & Objectives **1**

1.1 Background 1

1.1.1 Hansard Dataset 1

1.1.2 Natural language Processing 2

1.2 Analysis 2

1.3 Process 3

2 Design **4**

2.1 Overall Architecture 5

2.2 Some detailed design 5

2.2.1 Even more detail 5

2.3 User Interface 5

2.4 Other relevant sections 5

3 Implementation **6**

4 Testing **7**

4.1 Overall Approach to Testing 8

4.2 Automated Testing 8

4.2.1 Unit Tests 8

4.2.2 User Interface Testing 8

4.2.3 Stress Testing 8

4.2.4 Other types of testing 8

4.3 Integration Testing 8

4.4 User Testing 8

5 Evaluation **9**

Appendices **10**

A Third-Party Code and Libraries **11**

B Ethics Submission **12**

C Code Examples **13**

Annotated Bibliography **14**

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Background & Objectives

1.1 Background

1.1.1 Hansard Dataset

The Hansard Dataset is a set of documents produced by the British Parliament, which began in the 18th and 19th century. These documents contain reports and details of debates in the House of Commons, going back to the year 1803. Eventually, in 1907, these reports were made official and started being produced by Parliament itself, becoming The Official Report, though still unofficially known as Hansard. Along with becoming official, a report was officially defined as being one:

> which, though not strictly verbatim, is substantially the verbatim report, with repetitions and redundancies omitted and with obvious mistakes corrected, but which on the other hand leaves out nothing that adds to the meaning of the speech or illustrates the argument" (PUT REF HERE)

Hansard is available in a variety of versions. The most commonly used and best known version is the Daily Hansard, which appears each morning and reports of the previous days proceedings. However, access to this is via an API that only provides the most recent 7 days. For this project, most training and processing will be done on the Historical Hansard dataset, which is a dataset containing all 6 series of Hansard, between 1803 to 2004, though it is expected that some of the older documents will be less useful for this project due to the likelihood of them using outdated speech that would no longer be relevant.

The Historical Hansard Dataset is available online in an XML format. Multiple documents per series are available, each document covering a few days debates at most. It is a large dataset, reaching around 10Gb in size in total. Most of the documents available are scanned from hard copies, rather than typed up directly, meaning there is a possibility of small errors from the scanning process that may have to be dealt with. Additionally, from preliminary looks, the data itself appears to be only loosely formatted, and each of the six series appear to be formatted in a slightly different way, so any system designed to read this data will have to be capable of dealing with any changes to formatting.

1.1.2 Natural language Processing

To prepare for the project, there were two major areas needing research; Natural Language Processing, and the Hansard Dataset.

Research began on the Hansard Dataset, where the first task was to download the actual dataset, a 10Gb verbatum record of everything said in parliament between 1803 and 2004, but ...with repetitions and redundancies omitted and with obvious mistakes corrected" (https://www.commonwealth-hansard.org/about-hansard.html) This is sourced from hard copies, and as such there are some minor errors from when it was scanned, assumably by some form of OCR. However, these errors are minor, and uncommon enough that they prove no issue for the project as a whole.

One aspect of the Hansard dataset that needed to be planned for when designing the project was its lack of proper XML formatting. Though the data is presented in an Xml format, there were also HTML style tags within the bodies of text. In addition, each of the six series of data appeared to have slightly different formatting, and so anything used to read the data would have to be able to handle these differences.

In researching Natural language Processing, the Natural language Toolkit was discovered. This python module is a commonly used for Natural langage Processing, and provides methods and classes for many things, such as Named Entity Regongition, sentence splitting, and even some basic machine learning tactics for things like text classification, which could be used for the project.

1.2 Analysis

Taking into account the problem and what you learned from the background work, what was your analysis of the problem? How did your analysis help to decompose the problem into the main tasks that you would undertake? Were there alternative approaches? Why did you choose one approach compared to the alternatives?

There should be a clear statement of the objectives of the work, which you will evaluate at the end of the work.

In most cases, the agreed objectives or requirements will be the result of a compromise between what would ideally have been produced and what was determined to be possible in the time available. A discussion of the process of arriving at the final list is usually appropriate.

As mentioned in the lectures, think about possible security issues for the project topic. Whilst these might not be relevant for all projects, do consider if there are relevant for your project. Where there are relevant security issues, discuss how they will this affect the work that you are doing. Carry forward this discussion into relevant areas for design, implementation and testing.

1.3 Process

You need to describe briefly the life cycle model or research method that you used. You do not need to write about all of the different process models that you are aware of. Focus on the process model that you have used. It is possible that you needed to adapt an existing process model to suit your project; clearly identify what you used and how you adapted it for your needs.

Chapter 2

Design

You should concentrate on the more important aspects of the design. It is essential that an overview is presented before going into detail. As well as describing the design adopted it must also explain what other designs were considered and why they were rejected.The design should describe what you expected to do, and might also explain areas that you had to revise after some investigation.Typically, for an object-oriented design, the discussion will focus on the choice of objects and classes and the allocation of methods to classes. The use made of reusable components should be described and their source referenced. Particularly important decisions concerning data structures usually affect the architecture of a system and so should be described here.How much material you include on detailed design and implementation will depend very much on the nature of the project. It should not be padded out. Think about the significant aspects of your system. For example, describe the design of the user interface if it is a critical aspect of your system, or provide detail about methods and data structures that are not trivial. Do not spend time on long lists of trivial items and repetitive descriptions. If in doubt about what is appropriate, speak to your supervisor. You should also identify any support tools that you used. You should discuss your choice of implementation tools - programming language, compilers, database management system, program development environment, etc.Some example sub-sections may be as follows, but the specific sections are for you to define.

2.1 Overall Architecture

2.2 Some detailed design

2.2.1 Even more detail

2.3 User Interface

2.4 Other relevant sections

Chapter 3

Implementation

The implementation should look at any issues you encountered as you tried to implement your design. During the work, you might have found that elements of your design were unnecessary or overly complex; perhaps third party libraries were available that simplified some of the functions that you intended to implement. If things were easier in some areas, then how did you adapt your project to take account of your findings?

It is more likely that things were more complex than you first thought. In particular, were there any problems or difficulties that you found during implementation that you had to address? Did such problems simply delay you or were they more significant?

You can conclude this section by reviewing the end of the implementation stage against the planned requirements.

Chapter 4

Testing

Detailed descriptions of every test case are definitely not what is required here. What is important is to show that you adopted a sensible strategy that was, in principle, capable of testing the system adequately even if you did not have the time to test the system fully.

Provide information in the body of your report and the appendix to explain the testing that has been performed. How does this testing address the requirements and design for the project?

How comprehensive is the testing within the constraints of the project? Are you testing the normal working behaviour? Are you testing the exceptional behaviour, e.g. error conditions? Are you testing security issues if they are relevant for your project?

Have you tested your system on "real users"? For example, if your system is supposed to solve a problem for a business, then it would be appropriate to present your approach to involve the users in the testing process and to record the results that you obtained. Depending on the level of detail, it is likely that you would put any detailed results in an appendix.

The following sections indicate some areas you might include. Other sections may be more appropriate to your project.

4.1 Overall Approach to Testing

4.2 Automated Testing

4.2.1 Unit Tests

4.2.2 User Interface Testing

4.2.3 Stress Testing

4.2.4 Other types of testing

4.3 Integration Testing

4.4 User Testing

Chapter 5

Evaluation

Examiners expect to find in your dissertation a section addressing such questions as:

- Were the requirements correctly identified?
- Were the design decisions correct?
- Could a more suitable set of tools have been chosen?
- How well did the software meet the needs of those who were expecting to use it?
- How well were any other project aims achieved?
- If you were starting again, what would you do differently?

Other questions can be addressed as appropriate for a project.

Such material is regarded as an important part of the dissertation; it should demonstrate that you are capable not only of carrying out a piece of work but also of thinking critically about how you did it and how you might have done it better. This is seen as an important part of an honours degree.

There will be good things and room for improvement with any project. As you write this section, identify and discuss the parts of the work that went well and also consider ways in which the work could be improved.

In the latter stages of the module, we will discuss the evaluation. That will probably be around week 9, although that differs each year.

Appendices

The appendices are for additional content that is useful to support the discussion in the report. It is material that is not necessarily needed in the body of the report, but its inclusion in the appendices makes it easy to access.

For example, if you have developed a Design Specification document as part of a plan-driven approach for the project, then it would be appropriate to include that document as an appendix. In the body of your report you would highlight the most interesting aspects of the design, referring your reader to the full specification for further detail.

If you have taken an agile approach to developing the project, then you may be less likely to have developed a full requirements specification. Perhaps you use stories to keep track of the functionality and the 'future conversations'. It might not be relevant to include all of those in the body of your report. Instead, you might include those in an appendix.

There is a balance to be struck between what is relevant to include in the body of your report and whether additional supporting evidence is appropriate in the appendices. Speak to your supervisor or the module coordinator if you have questions about this.

Appendix A

Third-Party Code and Libraries

If you have made use of any third party code or software libraries, i.e. any code that you have not designed and written yourself, then you must include this appendix.

As has been said in lectures, it is acceptable and likely that you will make use of third-party code and software libraries. If third party code or libraries are used, your work will build on that to produce notable new work. The key requirement is that we understand what is your original work and what work is based on that of other people.

Therefore, you need to clearly state what you have used and where the original material can be found. Also, if you have made any changes to the original versions, you must explain what you have changed.

As an example, you might include a definition such as:

Apache POI library - The project has been used to read and write Microsoft Excel files (XLS) as part of the interaction with the client's existing system for processing data. Version 3.10-FINAL was used. The library is open source and it is available from the Apache Software Foundation [**?**]. The library is released using the Apache License [**?**]. This library was used without modification.

Appendix B

Ethics Submission

This appendix includes a copy of the ethics submission for the project. After you have completed your Ethics submission, you will receive a PDF with a summary of the comments. That document should be embedded in this report, either as images, an embedded PDF or as copied text. The content should also include the Ethics Application Number that you receive.

Appendix C

Code Examples

For some projects, it might be relevant to include some code extracts in an appendix. You are not expected to put all of your code here - the correct place for all of your code is in the technical submission that is made in addition to the Final Report. However, if there are some notable aspects of the code that you discuss, including that in an appendix might be useful to make it easier for your readers to access.

As a general guide, if you are discussing short extracts of code then you are advised to include such code in the body of the report. If there is a longer extract that is relevant, then you might include it as shown in the following section.

Only include code in the appendix if that code is discussed and referred to in the body of the report.

Annotated Bibliography

[1] "Sentiment Analysis in Python with TextBlob and VADER Sentiment (also Dash p.6) - YouTube." [Online]. Available: https://www.youtube.com/watch?v=qTyj2R-wcks

[2] B. Riedel, I. Augenstein, G. P. Spithourakis, and S. Riedel, "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task," jul 2017. [Online]. Available: http://arxiv.org/abs/1707.03264

[3] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, "Stance Detection with Bidirectional Conditional Encoding," jun 2016. [Online]. Available: http://arxiv.org/abs/1606.05464

[4] S. Dori-Hacohen, "Controversy Detection and Stance Analysis," *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1057–1057, 2015.

[5] W. Ferreira and A. Vlachos, "Emergent: a novel data-set for stance classification," pp. 1163–1168. [Online]. Available: http://aclweb.org/anthology/N/N16/N16-1138.pdf

[6] A. Aker, L. Derczynski, and K. Bontcheva, "Simple Open Stance Classification for Rumour Analysis," aug 2017. [Online]. Available: http://arxiv.org/abs/1708.05286

[7] P. Nagy, "Python NLTK Sentiment Analysis," 2017. [Online]. Available: https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis/notebook

[8] O. Onyimadu, K. Nakata, T. Wilson, D. Macken, and K. Liu, "Towards sentiment analysis on parliamentary debates in Hansard," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8388 LNCS, 2014, pp. 48–50.

BLOODY CANT READ AT THE MOMENT CAUSE PAYWALLS

[9] O. Onyimadu, K. Nakata, Y. Wang, T. Wilson, and K. Liu, "Entity-Based Semantic Search on Conversational Transcripts Semantic." Springer, Berlin, Heidelberg, 2013, pp. 344–349. [Online]. Available: http://link.springer.com/10.1007/978-3-642-37996-3{_}27

[10] L. Richardson, "Beautiful Soup Documentation." [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Documentation for BeautifulSoup4, a HTML/XML parser for Python, which could be useful for dealing with the large badly formatted XML documents.

[11] J. Perkins, “Text Classification For Sentiment Analysis - Naive Bayes Classifier,” 2010. [Online]. Available: https://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/

An online article that describes how the author trained a Naive Bayes classifier on movie reviews using the NLTK for Python. It should serve as a good reference when looking to do something similar.

[12] H. M. Noble, *Natural Language Processing*, 1st ed., T. Addis, B. DuBoulay, and A. Tate, Eds. Oxford: Blackwell Scientific Publications, 1988, 0632015020.

This book provides a decent overview on Natural Language Processing for computer systems. It does not, however, mention any specifics for Sentiment Analysis.

[13] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, 1st ed., 2009, vol. 43, p. 479. [Online]. Available: http://www.amazon.com/dp/0596516495 9780596516499.

A guide to the Natural Language Toolkit for Python, a specific package that may be used for the project.