TNA ACTIVITY REPORT

Developing an attribute-based sentiment analysis model for Romantic-period letters

Author: Cassandra Ulph

Current position: Digital Development Officer (Arts and Humanities)

Affiliation: University of Leeds

Host institution: University of Galway

Mentor(s): Justin Tonra

Period of stay: 8 Weeks

Introduction

This project was designed to explore sentiment in relation to place (as pre-defined geographical entities and as more dynamic concepts) in the letters of romantic-period British Author Hester Piozzi, in particular her former home of Streatham Park. This was a proof-of-concept project, to test the applicability of existing sentiment analysis models for the analysis of eighteenth-century private correspondence. The dataset for this project was The Piozzi Letters, a published edition of the collected letters of Hester Thrale Piozzi from the period 1783-1821.

Methodology

1) Pre-processing

The dataset was obtained from a digitized edition provided (via subscription access) at e-enlightenment.com. Plain text files were created for each letter with regular filename conventions. Metadata (author, recipient, date) was retained in text body. This was then transformed into structured data (csv) using linebreaks in regular expression to mark new sections. Individual letters would be divided into 'texts' (i.e. subunits for analysis) and those texts containing a reference to 'place' would form the 'place' corpus.

2) Placelist and creation of 'place' corpus.

Once the dataset was obtained and cleaned, 'places' would need to be identified in the corpus. Methods for georeferencing locations in the text would be investigated, including machine learning methods for identifying these.

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

3) Analysis and fine tuning

The 'place' corpus would be subjected to analysis by colleagues in Computational Linguistics using existing sentiment models trained on contemporary large language datasets. A sample of the results would then be error-checked to allow more precise tuning of the model for the domain of eighteenth-century familiar letters.

Description of the research visit and its outcomes

The early part of my research visit was spent in establishing the NLP pipeline for the project, and in pre-processing stage and the identification 'places' for the placelist, and in early conversations with Paul Buitelaar and Omnia Zayed (with whom I was collaborating in the Data Insight Centre) regarding the creation of the 'place' corpus. On their advice I transformed the unstructured data of the text files into structured data in .csv form, which allowed me to separate out the editorial metadata (author, recipient, date) and the letter text. One issue with this was the inconsistency of the original author's dateline practice: when included, the dateline was often between one and four newlines (as per regex format) and thus there was considerable cleaning work to be done to ensure all 'dateline' text was in a single column in CSV and could be identified as such. A key challenge was clearly communicating the expectations of collaborative working which in turn impacted decisions about what format would be most appropriate for providing pre-processed text for analysis. Much of the first few weeks was spent wrangling data into structured forms that ultimately weren't necessary for the approach ultimately taken due to lack of clarity regarding the type of analysis being performed and by whom. However, this did give me an opportunity to explore ways of transforming, cleaning and tokenising the dataset with NLP tools such as R, Python and NLTK that could be used in future projects.

Early discussions with colleagues in the Data Insight Centre led me to revise the scope of the 'placelist', particularly in the ambition to have parallel analyses of 'place-concept' terms as some of these were inherently sentiment-scored in existing models. Instead I narrowed focus to geographical places in the corpus for the purposes of this pilot study. I created a placelist through a combination of computer-assisted and manual means: first producing a concordance list of terms in the corpus using AntConc and from that extracting only geographical locations. In week 5 of the 8-week visit Gaurav Negi joined the Data Insight team for the project. Gaurav suggested that a named entity recognition model could automate this using Machine Learning with equal accuracy and less manual labour. Although I had already created a manual place-list, we decided to try this and compare the results. Some benefits to this approach was that the model picked out place-phrases of more than one token (e.g. 'Salt Hill') which were not obvious in from the concordance list produced in AntConc. However a large number of geolocations were missed by this method and many non-geographical places included. A challenge in both cases was the distinction between geographical places and individuals for whom place formed part of their title (e.g. Lady Jersey vs the state of Jersey). The creation of the placelist thus still required a high level of supervision and specialist contextual knowledge and ultimately I decided on the manual placelist to provide the aspect for the analysis of the letters, balancing the labour of hand-checking this against the improved accuracy it brought, which is an obvious challenge of working with historical texts using machine learning models trained on contemporary ones.

With the placelist established and the corpus prepared, Gaurav used a contemporary unsupervised large language model to predict sentiment at 'text' (ie subunit) level in relation to the specified 'places'. I then error checked a sample of 194 results from 50 letters (of a total of 4981 results from 1295 letters). This revealed a confidence of c. 84% which was considered reasonably high confidence for a model not trained on this type of data. No further fine-tuning was applied to the model at this point and no further analysis was done by colleagues in the Data Insight Centre, but discussions of how to address the implicit biases of the model will continue. The results returned a large number of 'neutral' sentiments, which was expected (for example when Piozzi states where she is writing from or describes a journey). When neutral returns were discounted, there was a slight tendency in the corpus towards positive over negative sentiment and the results should be understood in light of this. One slightly surprising result was that the expected shift in sentiment towards Streatham Park over the years didn't really take place within this dataset, instead vascillating across Piozzi's lifetime, however what was notable was that Streatham was one of the locations that proportionally elicited more frequent non-neutral sentiment, confirming that its emotional resonance continued for Piozzi throughout her life.

The geolocation of the 'placelist' required contextual expertise of historical placenames, anglo-welsh spellings, and classical geography. While historical gazetteers are available and these were considered, the international nature of Piozzi's data combined with her inconsistent naming practices and the tendency of APIs to default to American cities over British towns made use of a single gazzetteer impractical so in the interests of time I ultimately used a combination of the MS Excel 'geography' function and manual checking. Once geolocation of the 'placelist' was complete, I experimented with these results through a number of visualization platforms, including NodeGoat and QGIS (the latter with more success than the former). The final visualisations for the project presentation were produced by the Moore Institute's technical manager David Kelly, who produced a timelapse map-based visualization of sentiment data by date and geolocation based on data I provided.

Considerations for future work

A longer, better resourced study would allow for more thorough pre-processing, ideally through .xml markup of key elements in the text that might obviate some of the issues with named entity recognition, dateline identification and/or removal. Other possibilities discussed (but not explored due to time constraints) was the application of topic modelling to the corpus, and to further fine-tune the model for the specific domain of eighteenth-century letters. In terms of NER and place-based analysis, I would be keen to explore other models of named entity recognition including the 'surprising phrase detector' algorithm, and spend more time exploring methods of parsing historical names from the dataset using OpenRefine. As cross-disciplinary working presented some challenges, for future collaborations of this kind I would recommend establishing a workplan with clear task allocation agreed on by all parties before the commencement of the project.

Evaluation of NLP tools used

During pre-processing and scoping I investigated the use of NLTK as a method of tokenizing text. This could be used for a statistical supervised model of sentiment analysis which might be more accessible to researchers without a HPC specialism. However, the analysis was

ultimately undertaken by Gaurav Negi using an unsupervised Machine Learning model which would require more advanced computational methods of the researcher so was only possible through collaboration with colleagues with expertise in this area. Because the analysis was begun relatively late in the project, there was not time for further refinement of the ML model during the term of the fellowship, but the reasonably high predictive accuracy in an unfamiliar domain suggests that this model could be successfully adapted to eighteenth-century letters with a high degree of accuracy, thus establishing a convincing use-case for this methodology.