Interim Report

By: Bagus Maulana

Project Title (Project Plan): Using NLP and sentiment analysis to analyse media representation of people with disabilities.

Project Title (Current): Using natural language processing and sentiment analysis to develop a model to determine media representation of people with disabilities from web-based text articles.

Internal Supervisor(s): Cathy Holloway, Nicholas Firth

External Supervisor(s): -

Current Status:

In contrast to the vague, iterative approach described in my project plan, I have organised this project into several separate but inter-connected goals that I would need to achieve by the end of this project. Then, I re-organised the remaining work that I have yet to do into categories based on these goals. The advantage of this approach is that now the separate goals can be concurrently worked on (and prior work can be re-visited later),

The main goals that I have identified for this project are:

- Research and literature review (Review prior work relating to the techniques and the domain relevant to this project. The list of topics/keywords to research will grow as the project goes along and the specific natural language processing techniques and research questions are determined)
- Data collection (Determine relevant topics/keywords to the domain i.e. specific disabilities, determine relevant sources i.e. top N news sources in the UK, develop a script to crawl Web sources for these topics/keywords and parse relevant information e.g. text, date posted)
- Develop natural language processing pipeline (Based on a small sample (~1000 articles per topic and source) of collected articles from a selection of topics/keywords and news sources, develop a natural language processing pipeline to quantify relevant metrics, for example how positively or negatively does an article represent a specified disability)
- Experiments (Start formulating research questions relevant to the domain that would be answered by the data being collected and the metrics being parsed or analysed, e.g. 'how does sentiment with regards to people with ___ from source ___ change over time?' Once the data collection and natural language processing pipelines are ready, start data collection of a large sample based on the research question, and run analyses through the collected dataset to create meaningful conclusions to the research question)
- Writing the Final Report

For my research and literature review, I have compiled a list of topics or keywords that are relevant to my project, corresponding to the specific natural language processing techniques (e.g. sentiment analysis, part-of-speech tagging) that are relevant to my project, and the domain of the research (e.g. natural language processing of news articles, studies relating to representation of people with disabilities). I have also read up-to-date survey papers relating to natural language processing and the specific techniques relevant to my project, such as sentiment analysis. This will, in turn, make it

easier for me to find specific natural language processing papers for various needs later in the project (such as developing the pipeline and writing the report).

For data collection, I have determined the list of eight news sources that I am focusing on, and hope to implement web scrapers for. This list is based on quantitative data of the seven news sources with the highest monthly readership (of above 3 million) in the UK, published by a statistics company (Statista), plus the BBC. I have also determined the list of topics (specific disabilities) which I would collect relevant articles for, based on the officially-recognised list of disabilities by [the UK Government or UCL]. Furthermore, I applied a filter remove irrelevant articles (i.e. articles talking about something else with just a passing mention of the disability) using word occurrence. The filter checks for the rank of the keyword(s) in the article (i.e. how common the keyword(s) occur in the article relative to other words). To prove the effectiveness of the filter, I sampled the first 10 articles for each topic, manually check whether each sampled article is relevant as 'ground truth', and check whether the filter correctly detects relevant/irrelevant articles.

For the natural language processing pipeline, I have researched into various tools and Python libraries that are relevant to implement each component of the pipeline, such as SpaCy (for part-of-speech tagging, rule-based matching, etc..) and scikit-learn (for statistical / machine learning tools). I have also set up local environments with access to these libraries and implemented sample code relevant to components of my pipeline (e.g. a function to return semantically 'nearby' tokens, given a keyword and sentence as input). Additionally, I have also read natural language processing papers and planned a high-level design of the required pipeline; i.e. web crawling -> text collection -> topic filter -> token matcher (for keywords) -> part-of-speech / parse tree analysis -> sentiment analysis (of 'related' words to the keyword).

Remaining Work:

For my literature review, further work still needs to be done to gather research papers on each specific component of the natural language processing pipeline. This will go together with developing the pipeline, as I will use the information from these papers to understand better how to implement these components, and further work on the pipeline will allow a better understanding of the specific natural language processing techniques I would need to research. Additionally, I will do additional research on related domain-specific papers, to see the conclusions from prior work on this domain (i.e. analyses of media representation of disabilities). All prior research should be completed by mid-February (target deadline: 18th February), as it is crucial before developing the pipeline or formulating research questions.

For data collection, I will need to expand my web crawler script to be able to gather text articles from more data sources. Currently, my web crawler can gather text articles from two web sources (Daily Mail and Daily Express). To fulfil my stated requirements, I will need to expand this to include (hopefully) other sources with the seven highest monthly readerships in the UK, and the BBC (with the feasibility of each source to be determined, i.e. is it possible to retrieve historical articles?). This goal should be completed by the last week of February (target deadline: 4th March), as I would need to know the feasibility of data collection and filtering (i.e. how much data is available) with regards to certain topics and parameters (e.g. articles relating to x within time period y) before formulating research questions to address.

To develop the natural language processing pipeline, I will need to identify the components required, implement each specific component based on the format of the collected data, and

connect these components together. The components that I have identified to be useful for this project are: token matching, parse tree analysis, part-of-speech tagging, and sentiment analysis. I will also need to implement a final 'analysis' component to return relevant metrics (e.g. how positively or negatively does an article represent a specified disability) for each article. Then, I will analyse how well does my pipeline performs by analysing random samples, and see if I could improve these metrics by adding additional components or improving existing components, which will go together with further background research. This goal should be completed by the last week of February (target deadline: 4th March), as I would need to have completed the pipeline before conducting experiments, as well as gain a better understanding of what types of information can be understood from the textual content of articles.

To conduct experiments and create meaningful conclusions, I would need to formulate research questions based on prior domain-specific research, as well as the capabilities and limitations of the data collection and sentiment analysis pipelines. Then, I will conduct experiments to answer these research questions by collecting a large sample of relevant articles, performing natural language processing techniques on the dataset to gather relevant quantitative information. After the results have been gathered, I will develop data visualisations and generate meaningful conclusions (or lack thereof) based on the results. These experiments should be carried out throughout March (target deadline: 1st April).

The final Report is also due to be written, whereas the introductory, context, requirements and design sections should be written as I progress further towards this project and have a better understanding of the technical (i.e. data collection, pipeline) and experimental design. Once my data collection and sentiment analysis pipelines have been finalised and experiments have been conducted, I will be able to write down the results of the experiments, analyse the effectiveness of my approach, and do a thorough evaluation of the project to form an overall conclusion. The final report should be completed at least a week before the final deadline on 30th April 2018, with a rough draft being ready a week after completing experiments (target deadline: 8th April).

Supervisor's Signature: