LONDON'S GLOBAL UNIVERSITY

UCL

Using natural language processing to develop a pipeline to analyse media representation of people with disabilities in Web-based news articles

Collection and filtering of Web-based news articles, comparison of open-source sentiment models, and applications of the technology

Bagus Maulana[1]

MEng Computer Science

Catherine Holloway, Nicholas Firth

Submission date: 30th April 2018

[1]**Disclaimer:** This report is submitted as part requirement for the MEng Degree in Computer Science at UCL. It is substantially the result of my own work except where explicitly indicated in the text. *The report may be freely copied and distributed provided the source is explicitly acknowledged*

Abstract

Report Title: Using natural language processing to develop a pipeline to analyse media representation of people with disabilities in Web-based news articles: Collection and filtering of Web-based news articles, comparison of open-source sentiment models, and applications of the technology

Authors Name: Bagus Maulana

Supervisors Name: Catherine Holloway, Nicholas Firth

Date and Year of Submission: 30th April 2018

Contents

Chapter 1

Introduction

Natural language processing (NLP) encompasses a wide range of computational techniques for machine understanding of human (natural) language that are often used alongside each other. The review article [Cam14] defines natural language processing as 'a theory-motivated range of computational techniques for the automatic analysis and representation of human language.' The techniques that fall under the umbrella natural language processing include word tokenisation, probabilistic language modelling, translation, part-of-speech parsing, sentiment analysis (or opinion mining), text classification/categorisation, and topic modelling, among other things. The computational models used in natural language processing range from simple rule-based models (e.g. splitting a sentence on whitespace to tokenise words) to statistical machine learning and deep learning models. Natural language processing is now applied for various everyday technologies, for example, information retrieval for search engines such as Google and Bing, and categorisation and topic modelling for recommendation engines used to suggest 'similar' articles.

The obvious advantage of natural language processing is that machines can process vast bodies of human-created literature (books, articles, posts, e-mails, messages, etc.) much faster than humans can, processing thousands or millions of documents per second. This allows for high-level quantitative analyses of all documents in a vast corpora for a given domain to be feasible, which can uncover information previously inaccessible by only reading and generalising from a small sample of documents. For example, this level of quantitative analysis can uncover trends and patterns within a given domain (e.g. how does the popularity of the term 'mentally ill' increase or decrease year-on-year in English news media?).

Natural language processing covers three main 'curves' or areas: syntax, semantics, and pragmatics (narratives, understanding). Syntax specifies the way symbols (words, terms, tokens, or n-grams) and groups of symbols are arranged and whether they are well-formed in an expression, whereas semantics specifies what these expressions mean, and pragmatics specifies contextual information [Cam14]. Contemporary (or 'traditional') approaches to natural language processing mainly focus on syntactic analysis, due to the relative ease of extracting syntactic features of text such as term frequency, word co-occurrence, and part-of-speech tags, compared to extracting logical expressions and networks necessary for semantic analysis. However, syntactic analysis is much more limited as it often misses information such as the (semantic) context of a word (e.g. "one" in "there's no one there" (referring to a person) vs "we have only one car" (referring to a

quantity)). This paper will focus on mainly syntactic techniques and features, as these are more relevant to this domain of high-level topic matching and sentiment analysis that is feasible with current technology at this scale.

Various other studies have attempted to utilise natural language processing to perform high-level analyses in the domain of news media. The research done in [Lan17] assembled a vast corpus of regional newspapers in the United Kingdom spanning 150 years to detect long-term patterns of cultural change (e.g. increase of female representation in the news, or when trains overtook horses for transportation) by analysing n-gram trends and named entities. More specifically, in the domain of media representation of specified groups of people, studies such as [Zee16] has attempted to use features based on natural language processing (such as n-grams and part-of-speech tags) to classify racist and sexist posts in social media, although there is still a research gap in this area (especially for news articles, and/or relating to specific groups, such as people with (specific) disabilities or mental illnesses).

Applying natural language processing to perform meta-analyses over large text corpora has various interesting potential applications in improving our understanding of the human world - for example, to detect macroscopic cultural shifts as in [Lan17]. In particular, the representation of specific groups, such as people with disabilities, has been a popular research theme for psychologists, sociologists, and others. For example, the paper [Cov02] analysed a sample of 600 print articles relating to mental illnesses in New Zealand and categorised them to positive and negative depictions, and the predominant themes thereof (e.g. criminality (negative), educational accomplishments (positive)). Applying natural language processing to this area of research would allow the possibility of discovering higher-level trends, by computationally analysing a much larger sample of articles and identify trends by varying independent variables such as year of publication and publisher. In this research, a sample of 305,113 news articles (51,177 after filtering off-topic articles) from British online news sources are used. However, challenges remain as syntax-based statistical natural language processing approaches tend to be more limited in scope and is prone to false positives and negatives, and mitigating these factors is currently an open area of research.

The aim of this project is to utilise these natural language processing computational techniques in order to perform a high-level meta-analysis of literature available in the public news media available online. More specifically, to gather online news articles relating to people with disabilities in British media, and perform natural language processing analyses at scale to identify trends such as term popularity (e.g. 'suffer from ...' vs 'with ...') and variation in positive/negative sentiment.

The goal of this project is to develop a computational pipeline capable of performing this analysis of online media end-to-end. Given a list of topic consisting of keywords (or phrases) and query terms, this pipeline covers the task of web crawling and scraping, using public APIs if possible, to collect a dataset of news articles; filtering off-topic articles for the given keywords; matching relevant sentences referring to a keyword; performing sentiment analysis (using publicly available open-source libraries) on these sentences; and producing relevant plots to show applications of this technology. This pipeline will be available open source on GitHub (https://github.com/bmaulana/nlp-media)

This project was carried out in a step-by-step approach. The pipeline was developed as four individual components: a web scraper and crawler for data collection given a list of queries, a filter to remove irrelevant articles given a list of key terms, a parser to pattern-match relevant sentences

given key terms, a sentiment scorer (and results analysis/comparison of different open-source scorers on this domain), and a script to perform statistical analysis on the results and produce relevant plots. Each component's output is piped to the next component's input by saving its output to a JSON file and having the next component read the previous component's output file, which ensures computation can be 'resumed' without re-running the previous component. A main pipeline script connects these components together by calling them in order for each topic and supported Web source (Daily Mail, Daily Express, Guardian).

The body of this report is subdivided into four sections: context, requirements and analysis, design and implementation, and results evaluation.

Chapter 2

Context

2.1 Background

Analyses of news media in its various forms (print, online, etc.) has been a consistent research theme. The news media provides a quantifiable depiction of the prevailing society's popular conceptions or views regarding a theme or topic. This is especially applicable with regards to conceptions on particular groups of people, in which the language used in the media reflect on popular views, and has been shown to differ (with statistical significance) in different societies. For example, it was shown that the Canadian press was more likely to name individuals with disability and use appropriate labelling than the Israeli press in 1999 [Gol99]. Furthermore, there is evidence to suggest that news media sources contribute to shape and reinforce beliefs among the society, such as misconceptions and stigma [Wah92].

Public awareness of disability is also a popular research theme. While not related to news media, a review in 2011 [Sci11] found 75 articles and 68 studies that passed a selective inclusion criteria with regards to intellectual disabilities, published in English between 1990 and mid-2011. The topics brought up include the public's knowledge, attitudes, and beliefs about intellectual disability; and varying for socio-demographic characteristics, cross-cultural comparisons, and the effects of interventions.

Analyses of news or other media are primarily carried out by taking a small, statistically representative sample of documents (news articles) from a text corpora (for example, all news articles published in England for a certain period) and analysing them manually. There has also been various such studies within the domain of disability awareness in the media:

- In a 2002 study [Cov02], researchers analysed a sample of 600 print articles relating to mental health or mental illness that was collected by a commercial clipping bureau. The articles were then categorised into positive and negative depictions, then further into sub-samples such as danger to others, criminality, vulnerability, etc. The study found that at the time, in New Zealand, negative themes predominate about 3 to 1 (with 27% being positive). However, given the paper's scope, this cannot be generalised to learn trends, or how the conclusion varies given certain variables (e.g. time, location).
- A study conducted in 1998 [Gol99] were replicated in 2008 [Dev13] to assess change in representations of disability and persons with disability in the Canadian news media. This

study sampled 196 news articles in 1998 and 166 news articles in 2008. It found an increase in the usage of 'person-first' terminology (e.g. person with disabilities) and a decrease in 'disabling language' (e.g. disabled person). This is an attempt to identify trends with regards to media representation of disability, however only provides two data points (1998 and 2008) and has relatively small sample size.

- A study in 2005 [Jon09] analysed 1,515 articles relating to autism in Australian news media. All articles were read by two research assistants to ensure they are on-topic and then coded as either 'negative' or 'positive' in overall focus, and then coded into themes (e.g. funding, education, etc.). Key findings include a relatively limited amount of helpful information, and a 'dual stereotype' of people with autism labelled as either dangerous and uncontrollable, or unloved and poorly treated. Relative to other similar studies, this study appears to be focused more on qualitative discussion compared to quantitative results.

By applying natural language processing and computational techniques in analysing text articles, it is possible to develop a computational pipeline that could analyse and extract quantitative information from these articles at a much faster rate, enabling the analyses of a much larger scale of documents within a realistic time frame. While a sample of few hundred or thousand documents is usually enough to provide statistically significant conclusions, by providing an analysis of the full corpora (or a much larger sample), it is possible to uncover additional information from the data set. For example, higher-level trends (such as how the conclusion varies by year, location, publisher, etc.) can be discovered from a quantitative analysis of the larger dataset, by 'splitting' the result set into smaller subsets based on independent variables (e.g. year, location, publisher, etc.) and performing statistical comparisons of dependent variables (e.g. term frequency) between each subset.

Additionally, data collection is much more feasible in scale, cost and time using computational techniques. An automated script can be used to collect news articles published on the Internet at a rate of roughly one article per second, or thousands of articles per hour (varies on Web source, hardware, Internet connection, etc.), a vast improvement over contracting a commercial clipping bureau to provide 600 articles as in [Cov02].

Several studies has taken advantage of this approach to carry a more complete analysis of textual corpora. For example, [Lan17] assembled a corpus of 35.9 million news articles from 120 publishers in the United Kingdom between 1800 and 1950, representing 14% of all news articles published in the United Kingdom over that period. With this approach, the researchers were able to extract quantitative time-series information with regards to cultural trends as present in 35.9 million British news articles over the 150-year period. This amount of large time-series information (represented as n-grams and named entities) allowed them to discover macroscopic cultural trends. By analysing and comparing word (n-gram) trends across various topics, the researchers were able to identify trends that reflect cultural shifts e.g. 'train' increasing in popularity and overtaking 'horse' around 1900, or 'labour party' overtaking 'conservative party' and 'liberal party' in news coverage from the 1920s onward. Additionally, they also used entity recognition to extract named entities from articles and considered trends based on known information about these named entities, such as the proportion of female vs male entities, categories of entities (e.g. politicians, writers, etc.), and age of these entities. They also considered the geographical location of the publication to see how word usage trends (e.g. 'british' vs 'english') differ based on location.

[Lan17]'s study was based on prior discussions and studies on the potential of exploiting large text corpora to detect macroscopic, long-term cultural changes. [Mic11] was one of the first studies to suggest this approach. In this seminal study, a corpus 5 million digitised English-language books published over 200 years (or about 4% of all books ever published), provided by Google's effort to digitise books, were analysed to extract how often a given n-gram was used over time. (This data is available on http://www.culturomics.org/) This information is then used to analyse trends in language: the size of the English lexicon, regularisation of English verbs (from 'irregular' suffixes to '-ed'), or how quickly years (e.g. '1950') decline in use. Influenced by [Mic11], several other studies has been published adopting a similar approach:

- an analysis of 1.7 million Victorian-era books [Gib11]
- an analysis of 17,094 US Billboard Hot 100 songs between 1960 and 2010 [Mau15]
- an analysis of a 3.9 million news article sample from the Summary of World Broadcasts (SWB) collection [Lee11]
- an analysis of 2.5 million English-language million news articles from 498 online news outlets from 99 countries [Fla13]

However, there were also criticism of this approach, such that it ignored semantics and context, thus for example 'thirteen hundred words of gibberish and the Declaration of Independence are digitally equivalent' [Goo13], issues with OCR quality and duplicate editions [Goo13], or that the selection of digitised books are biased [Sch11].

There has been some progress in applying natural language processing techniques more specifically in the domain of how (specific) groups of people are represented in the media:

- A study has attempted to extract features (such as n-grams and part-of-speech tags) using natural language processing to classify racist and sexist posts in social media, providing annotations to 6,909 tweets [Zee16].
- Another study explored potential linguistic markers of schizophrenia in social media from a dataset of 174 users with self-reported schizophrenia and up to 3,200 tweets per user, and control users in a 50:50 split. They used a support vector machine (SVM) approach over features based on lexicon-based approaches (i.e. a list of mental health related keywords), latent dirichlet allocation (LDA), Brown clustering, character n-grams, and perplexity [Mit15].

However, there is still a visible research gap in this area, for using computational natural language processing specifically for news articles and/or with regards to specific groups, such as people with disabilities (or a specific disability).

To date, advances in natural language processing has made this approach much more feasible, efficient, and effective, even given limited time and hardware constraints. Various free and open-source tools for computational natural language processing and statistical analysis has been developed by the research community, such as NLTK [Bir09] and StanfordNLP [Man14] for natural language processing, scikit-learn [Ped11] for statistical analysis, and matplotlib [Hun07] for plotting. Depending on the type of data analytics performed and the hardware used, these computational tools are able to analyse news articles at a rate of multiple documents per second. The 'Technical Context' section contains a further listing and discussion of these tools, alongside the specific libraries and natural language processing techniques used for this research project.

2.2 Research Methodology and Sources

Background research were carried out by investigating papers from public sources (such as Google Scholar). Research articles were gathered from a list of important topics and query terms related to the technique (natural language processing) and/or the domain (disability and news media): for example, 'natural language processing review', 'news media' AND 'disability', 'natural language processing' AND ('news media' OR 'cultural trends'), and 'natural language processing' AND 'disability'. Highly-cited research articles are prioritised as examples, as they are deemed to be more 'important' papers or studies within its topic. Additionally, the author looked for highly-cited 'key' papers and review articles within each specific topic, and then looked at its list of citations (older papers), and research articles that cites the review article (newer papers), to expand the list of relevant research papers and examples.

Technical research, on the other hand, were carried out as necessary.

The sources used for this literature review are:

- Google Scholar, often used to find research articles to act as 'entry points' towards a research topic, or to find additional examples similar to another article within a given topic

2.3 Technical Context

As mentioned earlier, computational natural language processing techniques are utilised to extract features from collected news articles (text documents) at scale. For this project, the requirements for the computational pipeline can be subdivided to five main components: data collection, filtering, sentence matching, sentiment scoring, and statistical analysis and plotting. Among these components, natural language processing techniques are necessary for filtering, sentence matching, and sentiment scoring. On the other hand, collection is performed using established, general-purpose web-scraping tools. Similarly, statistical analysis and plotting is performed using general-purpose statistical tools and metrics. This section will cover the technical tools researched and used for all listed components regardless.

Python was chosen as the main programming language used for this project. The primary reason for this choice is the wide availability and range of existing tools for natural language processing, sentiment analysis, statistical analysis and plotting, web scraping and parsing, etc. in Python. A study in 2016 confirmed that Python is the most popular language for machine learning and data science [[TODO cite]], which correlates to the amount of available tools developers have created for the language. A GitHub search for the topic 'natural language processing' as of 18 April 2018 reported 991 Python and 338 Jupyter repositories with the tag 'natural language processing', compared to the second most popular programming language being Java with only 156 repositories tagged 'natural language processing' [[TODO cite]]. Furthermore, Python is also an ideal language for experimentation due to relatively high-level and low verbosity of the code, such that it is relatively easier to make small changes on the fly. Additionally, the Anaconda distribution of Python [[TODO cite]] is used for its suitability to set up and manage Python environments and packages for data science projects.

For data collection, general-purpose tools for sending HTTP requests (to 'open' web addresses and store HTML web pages programmatically) and parsing HTML code (to parse article text and

metadata from 'raw' HTML code) are sufficient. The Requests library [[TODO cite]] is a popular Python tool (with 400,000+ daily downloads) for sending HTTP/1.1 requests simply. A HTTP GET request will retrieve the HTML code (and other information) associated with a given URL from a web server, similar to opening the page on a web browser, stored as a Python object by Requests. Once the HTML code of a web page (given an article's URL) has been stored, the BeautifulSoup library [[TODO cite]] provides simple methods to navigate and search a parse tree (such as HTML code). Given that web pages from the same source/publisher tend to follow a similar structure, BeautifulSoup can be used to parse article text and relevant metadata (e.g. headline, date of publication, outgoing links in a search page) by searching for specific tags and attributes within the HTML code.

In this project, filtering off-topic articles is done somewhat crudely via ranking term frequency independently for each document. Term frequency is a simple and commonly-used metric for natural language processing, with various existing tools that can compute this metric for thousands of text documents within seconds. To put simply, the term frequency of a term (i.e. a word or token) in a document is the number of times that the term occurs in the document. The popular scikit-learn library [Ped11] provides a tool to measure term frequency of text documents, handling both tokenisation (converting a text document into a list of tokens (terms/words)) and counting word occurrence [[TODO cite]]. It also provides the option to ignore stop words (i.e. common, meaningless words in English, such as 'the' or 'a'), which can often skew the results due to its relative high frequency. Additionally, the NLTK library [Bir09] is used for stemming – to reduce all words in the document to its word stem (such that e.g. 'walk', 'walks', 'walked', and 'walking' are equivalent). Token ranking is implemented by sorting all terms (words) in the document based on their term frequency, and checking if the rank of the keyword in the document is lower (document is on-topic) or higher (document is off-topic) than a given threshold value. As multiple keywords per topic are used, all keywords in the document are first converted to a unique 'keyword token' before term frequency is measured (which will treat all instances of any keyword as a single term).

In literature related to natural language processing, various more sophisticated approaches has been proposed and used for the task of text classification and filtering off-topic articles. A conventional approach is by calculating term frequency – inverse document frequency (tf-idf) [[TODO cite]], a metric that builds on term frequency by taking into account the relative importance of each word. Inverse document frequency (idf) is calculated by counting the number of documents in a corpus where a term appears - if a term appears more frequently (e.g. common words such as 'the'), it is deemed to be less important and assigned a lower score. However, this approach is unsuitable for this project, given the selective nature of the dataset (where only articles with certain query terms are collected), thus the idf values of these query terms would be flawed (as they would appear in every document). Topic models, which compute the proportion, has also been proposed. Latent dirichlet allocation (LDA)

Chapter 3

Requirements and Analysis

3.1 Problem Statement

3.2 Requirements

3.2.1 Data Collection

3.2.2 Dataset Filtering

3.2.3 Rule-based Sentence Matching

3.2.4 Sentiment Scoring

3.2.5 Statistical Analysis and Plotting

3.3 Analysis of Requirements

Chapter 4

Design and Implementation

4.1 Overall Design

4.2 Dataset Description

4.2.1 Sources

4.2.2 Keywords, key phrases, and query terms

4.2.3 Dataset Size

4.2.4 Limitations

4.3 Components

4.3.1 Data Collection

4.3.2 Dataset Filtering

4.3.3 Rule-based Sentence Matching

4.3.4 Sentiment Scoring

4.3.4.1 Comparison of open-source sentiment models

4.3.4.2 Final implementation

4.3.5 Statistical Analysis and Plotting

Chapter 5

Results Evaluation

Chapter 6

Conclusions

6.1 Achievements

6.2 Evaluation

6.3 Future Work

Bibliography

[Bir09] S. Bird, E. Loper, and E. Klein. Natural Language Processing with Python. *O'Reilly Media Inc.*: 2009.

[Cam14] E. Cambria and B. White. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine*, 9(2): 48–57, 2014.

[Cov02] J. Coverdale, N. Raymond, and C. Donna. Depictions of Mental Illness in Print Media: A Prospective National Sample. *Australian & New Zealand Journal of Psychiatry*, 36(5): 697–700, 2002.

[Dev13] K. Devotta, R. Wilton, and N. Yiannakoulias. Representations of disability in the Canadian news media: a decade of change?. *Disability and Rehabilitation*, 35(22): 1859–1868, 2013.

[Fla13] I. Flaounas, et al. Research Methods in the Age of Digital Journalism: Massive-scale automated analysis of news-content – topics, style and gender. *Digital Journalism*, 1(1): 102–116, 2013.

[Gib11] F.W. Gibbs and D.J. Cohen. A Conversation with Data: Prospecting Victorian Words and Ideas. *Victorian Studies*, 54(1): 69–77, 2011.

[Gol99] N. Gold and G.K. Auslander. Media reports on disability: a binational comparison of types and causes of disability as reported in major newspapers. *Disability and Rehabilitation*, 21(9): 420–431, 1999.

[Goo13] P. Gooding. Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, 28(3): 425–431, 2013.

[Hun07] J.D. Hunter. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering*, 9(3): 90–95, 2007.

[Jon09] Sandra C. Jones and Valerie Harwood. Representations of autism in Australian print media. *Disability & Society*, 24(1): 5–18, 2009.

[Lan17] T. Lansdall-Welfare, et al. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences*, 114(4): 457–465, 2017.

[Lee11] K. Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9): 2011.

[Man14] C.D. Manning, et al. The stanford corenlp natural language processing toolkit. *ACL (System Demonstrations)*: 55–60, 2014.

[Mau15] M. Mauch, R.M. MacCallum, M. Levy, and A.M. Leroi. The evolution of popular music: USA 19602010. *R. Soc. opensci*, 2(5): 150081, 2015.

[Mic11] J.B. Michel, et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014): 176–182, 2011.

[Mit15] M. Mitchell, K. Hollingshead, and G. Coppersmith. Quantifying the Language of Schizophrenia in Social Media. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*: 11–20, 2015.

[Ped11] F. Pedregosa, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*: 2825–2830, 2011.

[Sch11] T. Schwartz. Culturomics: Periodicals Gauge Cultures Pulse *Science*, 332(6025): 35–36, 2011.

[Sci11] K. Scior. Public awareness attitudes and beliefs regarding intellectual disability: A systematic review. *Research in Developmental Disabilities*, 32(6): 2164–2182, 2011.

[Wah92] O.F. Wahl Mass media images of mental illness: A review of the literature. *Journal of Community Psychology*, 20(4): 343–352, 1992.

[Zee16] W. Zeerak. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*: 138–142, 2016.

Appendix A

Other appendices, e.g., code listing

Put your appendix sections here