# Take-home exam

University of Amsterdam

Graduate School of Communication Research

Research Master

Course: Big Data and Automated Content Analysis

Lecturer: Prof. Damian Trilling

Student: Janice Butler

Student number: 12356093

22.April 2020

*Assignment 1: Essay (Word Count: 539)*

There is no question that the use of Big Data in both scientific research and in industry has led to diverse questions. Big Data has been described as a paradigm shift i.e. a fundamental change in the approach. The emergence of Big Data Analysis – enabled through advances in the field of computer science – has influenced many disciplines and provided previously inaccessible insights.

Big Data is, as Boyd and Crawford (2012) emphasize, imprecise, as it can be misleading; originally implying amounts of data so large, that they required supercomputers to process. Body and Crawford (2012) refer to Big Data as the capacity to "search, aggregate, and cross-reference large data sets". With this capability looms the threat of a "Big Brother" scenario with the attendant invasion of privacy and security. Recent scandals surrounding the Russian invasion in the US presidential election of 2012 and Cambridge Analytica in 2014 have not only affected the trust in governments but also academia (Wong, 2019).

Within the European Comission's BYTE[1] project, it was realised that European policy, confronted with Big Data efforts, may have inadequacies. Cuquet and Rensel (2018) explain that of all the sectors analysed, only the healthcare sector would be adequately equipped when data sources and propriety protection are addressed.

The ability to collect and process massive amounts of data has led to new opportunities for researchers, but ethical questions persist regarding which data is valid for collection and for which purpose it may be used. Often people who post on social media don't consciously realize that their data is being collected for scientific purposes and consent is difficult if not impossible to refuse. Even if consent were given, both Boyd and Crawford (2012) and Kitchin (2014) reference scientists and people in industry who argue that numbers – irrespective of theoretical considerations – speak for themselves. Both of the paper's authors write that this assumption is flawed and data can easily be taken out of context. Kitchin (2014) even goes a step further in quoting people that believe that there is no need for the grounding of theories anymore. This could lead to false inferences, though, as Boyd and Crawford (2012) argue, one may look for and find patterns that do not exist. This is due to the sheer amount of data where spurious correlations can show up. The need for theory is thus still present and is even more important in datafied times.

Boyd and Crawford (2012) even more crudely quote a scientist, saying that academics should leave the analysis of data to industry. The capability to analyse the data has evolved to an ingroup versus outgroup formation. The propensity of data being only analysed by one or a few groups of people (be they in academia or industry) can lead to an inherent bias in the analysis. The authors state that interpretation is at the heart of the data analysed. Knowing where data comes from is more important than ever. Boyd and Crawford (2012) explain that Twitter data is often generalized to explain all

[1] BYTE Project: Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities

(natural) “people”, but when Twitter data is extracted, as Lewis, Zamith and Hermida (2013) explain, only a small percentage is actually utilised, leading to questions about representativeness.

*Assignment 3: Planned analysis (Word Count: 739)*

It has been noticed that the data that was analysed previously in Assignment 2 (results can be found in a separate file) is compromised of the RSS feed of news and sports articles of BBC Northern Ireland. For a further analysis, a research question could be formulated as follows:

RQ: To what extent does the coverage of provincial politicians differ across the four countries of Great Britain from the perspective of English-based news articles versus those emanating from the provinces (Wales, Scotland and Northern Ireland) and vice versa?

Thus, the range of BBC articles to be analysed could be extended to also include other parts of Great Britain. It is feasible to surmise that an English journalist, for example, would describe Northern Irish politics and politicians with more distance and a different slant than the provincial (NI) journalists themselves do, but it would be even more interesting to measure whether the same bias exists equally for each of the provinces and whether a symmetry exists in the opposite direction regarding Scottish, Welsh and Northern Irish journalism referencing English politicians (Gibson, 2008).

A particular focus of the coverage could be the emphasis and tone of writing. A way of analysing the tone could be in evaluating (political) sentiment on the body of the news articles. Several methods would be applicable to do this from a simple dictionary-based approach to – assuming sufficient time and resources – a supervised machine-learning approach.

With the dictionary-based approach using the Sentistrength algorithm, as an improved bag of words approach, a basic list of positive and negative words would be read in extended to include specific political and economic terms with a negative or positive sentiment. With this approach the booster words for strengthening or weakening the following word would allow for a more realistic analysis of news articles.

A more tailored approach – and more complex in its implementation – could be to use supervised machine learning to reduce the amount of manually coded words and with sufficient training data being available to achieve a more nuanced estimation of political sentiment. Training and verification data could be gathered from BBC Northern Ireland, which then could be used to test news from the other BBC sites (i.e. England, Wales and Scotland). The lengthy process of manually grading sentences from the sample data would be one method of supplying target outputs for training. Alternatively, it may be adequate to utilize standard sources of training data from other researchers in the field of political sentiment research. In order to put the sentences from the news article body through the neural network, the words have to be converted to numerical form. This is typically achieved using word embeddings, which encode words as a high-dimensional vector. Conveniently, the GloVe method is utilised with 300 dimensional vectors (Pennington et al., 2014). Other algorithms are available.

A further idea might also be to cross-reference the content (and its metadata) with statistics from a web-tracking tool like Google Analytics. In this way the different BBC news sites could be comparted and ranked for popularity. The topics of news and subtopics containing references to the province in question could be picked out. However, analytics data is never freely accessible and would require a close cooperation with – in this case – the BBC.

Taking the supplied data as a basis for the analysis, it could be collected in dictionaries using the primary (_id) key to cross-reference any data that is there. In this way one could see for instance, which authors produce how many articles for the different subtopics and topics. A start to this approach was already used in creating a dictionary (authorDict) for the second Assignment, where for the dictionary items, the primary key of the data (the URL of the rss article) is used. It is thought that this approach could be extended to the other BBC sites and associated data. Using the created dictionary for cross-referencing, one could ascertain all manner of statistics for articles associated with different topics and subtopics, grouping or filtering by different dates or groups of dates. As an example, it would be possible to visualize which numbers of topics/subtopics are being published week-by-week over a period of potentially several months or years. Similarly, one could visualize how many articles each author publishes on each topic/subtopic.

An overview of all of the items of interest would ideally be displayed in the style of a dashboard, combining charts and tables using pandas.

*References*

boyd, danah, & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Cuquet, M., & Fensel, A. (2018). The societal impact of big data: A research roadmap for Europe. Technology in Society, 54, 74–86. https://doi.org/10.1016/j.techsoc.2018.03.005

Gibson, O. (2008, June 12). BBC journalists accused of London bias. The Guardian. https://www.theguardian.com/media/2008/jun/12/bbc.tvnews

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 205395171452848. https://doi.org/10.1177/2053951714528481

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. Journal of Broadcasting & Electronic Media, 57(1), 34–52. https://doi.org/10.1080/08838151.2012.761702

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Empirical Methods in Natural Language Processing (EMNLP), 1532–1543. http://www.aclweb.org/anthology/D14-1162

Wong, J. C. (2019, March 18). The Cambridge Analytica scandal changed the world – but it didn’t change Facebook. The Guardian. https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook