MA5851 – Assessment 3: WebCrawler and NLP System

Does political bias in the media impact public opinion? This is a central question for public opinion researchers.[1] Since the 2010s, researchers have been developing 'automated public opinion' analysis, which involves using machine learning for detecting the distinct topics present in public discourse and how different segments of the public feel about those topics.[2][3] Several researchers have discovered interesting connections between public discourse and economic phenomena, for example the connection between positive sentiment on Brexit in 2019 UK and the Pound Sterling (GBP) exchange rate.[4] Other research has used automated public opinion analysis to analyse what segments of society have which political views, such as vaccine hesitancy in the USA following the COVID-19 pandemic.[5]

There is rich research into using machine learning to quantify public opinion. Some research has also applied these same techniques to examining how certain topics are treated in the media.[6][7]

The underlying principle to this analysis approach is in two parts:

1) Given a corpus of media (for example, text), determine what topic(s) the material is related to.
2) Determine how the subject of this media feels about that topic (ie: do they feel positive or negatively towards that topic).

This report details a prototype which will take this basic approach and apply it to media articles with the aim of reaching 'sentiments per topic' metrics, which show when particular topics are treated with particular sentiments to a statistically significant degree.

In future, this prototype could be extended to research connections between sentiments-per-topics in new articles, to the public opinions of their readership. This would provide insight into the extent of the impact which media has on public opinion formation.

In this report, we detail an extensible prototype which accomplishes a basic version of the use case. The intention is for the prototype to be built on in later iteration. This version only proves the concept of using web crawling and NLP to generate 'sentiments-per-topic' metrics.

[1] Yakunin et al., "Mass Media Evaluation Using Topic Modelling."
[2] Sokolova et al., "Topic Modelling and Event Identification from Twitter Textual Data."
[3] Rabitz, Telešienė, and Zolubienė, "Topic Modelling the News Media Representation of Climate Change."
[4] Ilyas et al., "Analyzing Brexit’s Impact Using Sentiment Analysis and Topic Modeling on Twitter Discussion."
[5] Nemes and Kiss, "Social Media Sentiment Analysis Based on COVID-19."
[6] Heidenreich et al., "Media Framing Dynamics of the 'European Refugee Crisis.'"
[7] Yakunin et al., "Mass Media Evaluation Using Topic Modelling."

The architecture of this protype application is (figure 1):

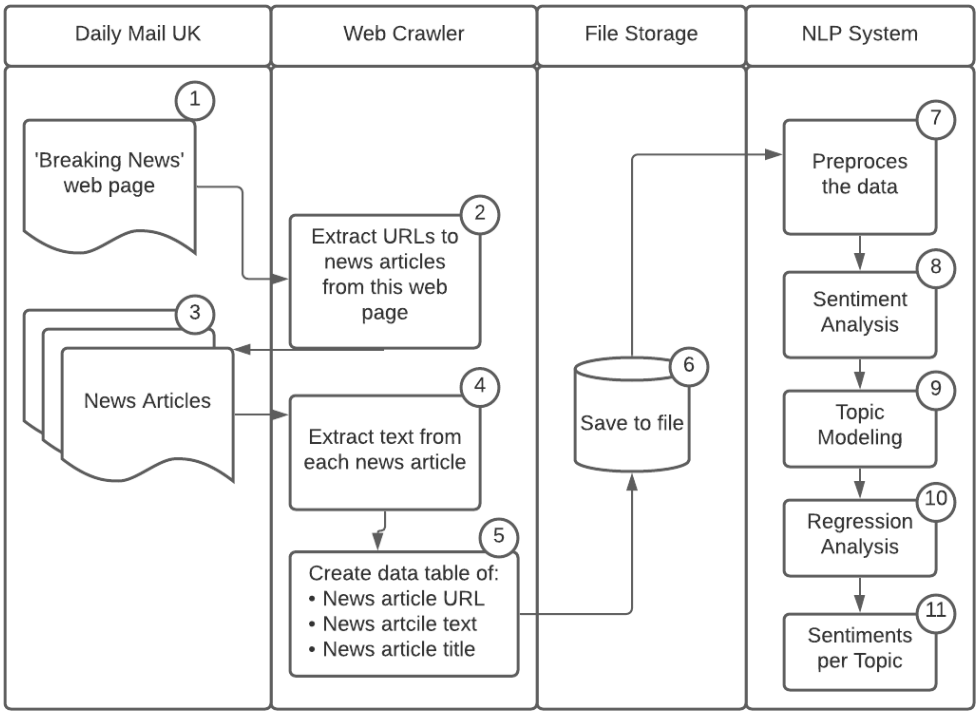

Figure 1

1) The input to the pipeline is the 'Breaking News' webpage of the Daily Mail UK news outlet. This pipeline can be reconfigured to work with other, or multiple, news outlets in future iterations.
2) The Web Crawler pipeline will extract the news article URLs from the 'Breaking News' header URL.
3) From each of these sub-URLs, the web crawler will access each article.
4) The web crawler pipeline will extract the text body from each article.
5) The web crawler pipeline will output a data table featuring the raw text of each news article.
6) The output from the web crawler pipeline is written to a file to be consumed by the subsequent pipeline.
7) The NLP System pipeline consumes the file and pre-processes the text data for the subsequent NLP operations.
8) Each news article is analysed for sentiment and given an overall sentiment score.
9) The whole corpus of articles is analysed for topic modelling and given a score for each article's resemblance to the identified topics.
10) Regression analysis is conducted between topic scores and sentiment scores.
11) The final output of the application is a report of any significant correlations between topics and sentiment.

Please proceed to read file 'Jack_Collins_A3_MA5851_Part2'.

The repository for this project can be found at:

https://github.com/JackCollins1991/Topic-Modelling-and-Sentiment-Analysis-on-News-Media

References

Heidenreich, Tobias, Fabienne Lind, Jakob-Moritz Eberl, and Hajo G Boomgaarden. “Media Framing Dynamics of the ‘European Refugee Crisis’: A Comparative Topic Modelling Approach.” *Journal of Refugee Studies* 32, no. Special_Issue_1 (December 1, 2019): i172–82. https://doi.org/10.1093/jrs/fez025.

Ilyas, Sardar Haider Waseem, Zainab Tariq Soomro, Ahmed Anwar, Hamza Shahzad, and Ussama Yaqub. “Analyzing Brexit’s Impact Using Sentiment Analysis and Topic Modeling on Twitter Discussion.” In *The 21st Annual International Conference on Digital Government Research*, 1–6. Dg.o ’20. New York, NY, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3396956.3396973.

Nemes, László, and Attila Kiss. “Social Media Sentiment Analysis Based on COVID-19.” *Journal of Information and Telecommunication* 5, no. 1 (January 2, 2021): 1–15. https://doi.org/10.1080/24751839.2020.1790793.

Rabitz, Florian, Audronė Telešienė, and Eimantė Zolubienė. “Topic Modelling the News Media Representation of Climate Change.” *Environmental Sociology* 7, no. 3 (July 3, 2021): 214–24. https://doi.org/10.1080/23251042.2020.1866281.

Sokolova, Marina, Kanyi Huang, Stan Matwin, Joshua Ramisch, Renee Black, Chris Orwa, Sidney Ochieng, and Nanjira Sambuli. “Topic Modelling and Event Identification from Twitter Textual Data,” n.d., 17.

Yakunin, Kirill, Ravil Mukhamediev, Rustam Mussabayev, Timur Buldybayev, Yan Kuchin, Sanzhar Murzakhmetov, Rassul Yunussov, and Ulzhan Ospanova. “Mass Media Evaluation Using Topic Modelling.” In *Digital Transformation and Global Society*, edited by Daniel A. Alexandrov, Alexander V. Boukhanovsky, Andrei V. Chugunov, Yury Kabanov, Olessia Koltsova, and Ilya Musabirov, 165–78. Communications in Computer and Information Science. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-65218-0_13.