Term project - Phase 1:
Pandemic's media coverage in UK

Alina Anisimova (3325121), Benedetta Ghedi (7005571),
Alexander Kern (5711088) & Amalia Musters (6633668)

May 7, 2020

1 Problem Definition

This research is aimed at analysing the current pandemic's media coverage. Our goal is to study the development of agenda setting of the British newspaper The Guardian and the identify different framing categories. We also aim at identifying how big events are covered by the news by devoting attention to the days following these events, such as the announcement of new measures taken by the government or the hospitalisation of Boris Johnson.
There is little research on the development of crisis news coverage and its framing. The understanding of a crisis, as of any other phenomenon, is directly determined by how it is communicated. Newspapers contribute greatly to the social construction of an event such as the current pandemic, and it is therefore of interest to understand and analyse which topics are covered most and what kind of framing is used: communication during a crisis is crucial in determining people's view on the situation and their subsequent behaviour.

2 Literature review

Social scientists have been using content analysis to study patterns in communications for decades. With the rise of the World Wide Web and the increasing amount of content available, scientists have developed and used computational methods to carry out content analysis which was, in the past, carried out by humans. Although computational content analysis allows for larger bodies of data to be studied, it was often noted that they lack the ability of human coders to identify latent features and therefore can result in less precise measurements [2].
Framing analysis applied to news broadcasting involves the definition of categories of perspectives from which news is delivered and the identification of each piece of content's category. Framing research can be divided in deductive and inductive: the former requires researchers to first define the categories to which divide the content into. This approach has sometimes been criticized due to the subjective choice of frames that might convey bias results (van Gorp, 2010). These remarks led to the development of inductive approaches in which categories are extracted from the analysis of the corpus of data through methods such as hierarchical clustering: frame elements are first coded manually and then the clustering is carried out automatically. [6]. In more recent times, machine learning was introduced to tackle the task. Topic modelling is an unsupervised machine learning method that defines topics based on frequencies of words and co-occurrences. This method was used to analyse news coverage by Di Maggio [3]. Network analysis can then be applied to the topics to identify themes, following a hybrid approach described by Walter & Ophir [9].
As for the subject of our research, the current global pandemic, there is little systematic understanding of how crisis communication operates and how framing is used in the news [7]. Research on media coverage of crisis and epidemics has adopted both classical thematic frames [5] and developed specific frames for crisis [1] - human interest, economy, morality and responsibility- and diseases [8] - consequence, uncertainty, action, reassurance, conflict and new evidence.

3 Research & Data Plan

Throughout the research question development, particularly in the analysis phase, we are going to revise the methods and tools we apply accordingly. Our initial goal is to apply Natural Language Processing (NLP) techniques [4] whereby non-verbal conversations can be easier interpreted and analysed. This involves finding the numeric representations for words from our dataset corpus. This can be done using count vectors tf-idf, and word embeddings. Once we build these feature vector representations of text, we may use this in many useful machine learning techniques such as classification of the media coverage contents by topics, categories and so on. Another useful technique that is relevant to our research question is topic modelling. By this we mean specific implementations such as non-negative matrix factorization, latent Dirichlet allocation, or LDA, or latent semantic analysis. Topic modelling will serve us the purpose of answering the question, what is this topic about? Topic modelling is the natural precursor to clustering media coverage content where we group contents based on topics or maybe keywords. The objective of clustering is to find groups of similar contents. Keywords extraction will be used for answering the question, what is important in the news coverage content in relation with specific measure announcement.
For this project we plan on combining data from two different sources. The sources and the datasets are described below.

3.1 Guardian API – News Articles

First off, this project will make use of The Open Platform by the Guardian, an Application Programming Interface (API) to get access to all the content the Guardian creates, categorised by tags and section. The Open Platform offers access to students for non-commercial usage of the content. This means that one can get access to over 1,900,000,000 articles. Beside the plain text within the articles the API also offers more information about the content. Each article contains information about the date published, the wordcount, the language used and relevance score, indicating its relevance when a search query was used. The guardian also manually categorizes their articles using tags. There are over 50,000 different tags and one article can be categorized using multiple tags. The guardian also uses sections to logically group their articles. Each article is featured in one of the many sections present. We will use the API to collect all the articles published on the United Kingdom's edition of the Guardian webpage since the 1st of December 2019. We also will query for COVID 19, Coronavirus and other terms relevant to the current pandemic to get all the articles that contain any mention of these terms and are therefore relevant to our project. This will result in a dataset containing the text of the articles and other information like tags, wordcount and relevance score. We will look into specifying the query a bit further, for example filtering out certain sections like 'Sports' because these sections will most likely not contain any relevant articles. At this point the dataset contains a vast amount of information about the media coverage of COVID19 by UK's leading online newspaper. This data can help us analyse how the Guardian handles the announcements of new measures in this crisis and other important events, like the hospitalisation of Boris Johnson. However, it cannot be used to make a statement about the entire media coverage in UK on COVID 19. We would need to add articles from other news sources to have a more complete view of the media coverage.

3.2 ACAPS – Government Measures

Secondly, this project will use the COVID19 Government Measures Dataset created by ACAPS. This dataset puts together all the measures implemented by governments worldwide in response to the Coronavirus pandemic. ACAPS collected this data by consulting governments, media organisations, the United Nations and other organisations sources. They created this dataset to help to create a structured view of measures taken, extended and lifted by governments around the world. The measures in the dataset fall into five categories: Social distancing, Movement restrictions, Public health measures, Social and economic measures and Lockdowns. Each category is broken down again into several types of measures. This dataset is going to be used to create a detailed timeline of events concerning COVID19 in the UK. However, because this dataset contains only measures to get a more complete timeline we would need to add other events, either manually or by adding other datasets. By having a structured and very accurate timeline of events concerning COVID19 in the UK we can analyse how each of these events were covered by the Guardian.

4 Ethical Discussion

Due to the nature of the current project, no specific issues concerning data collection were expected. More specifically, the Guardian API, which is a public data source, contains all the articles published on the United Kingdom's edition of the Guardian webpage since the 1st of December 2019. This means that no personal information that might potentially be harmful, is included in this data set. The information used from the Guardian API consists for example of the text of the articles, their date published and other information like wordcount. Potential harm due to the processing of the data is not likely, as again, no personal information was included and only information on articles is used. In other words, due to the nature of the data, it is very unlikely that any harm would occur due to the processing of the data, for example in terms of privacy or harm to groups.

In the same manner for the ACAPS data set on COVID19 Government Measures, no personal information is included in this public data set. The data includes information on several categories, e.g. Social distancing, and Public health measures. The data consists of information such as the local government in which COVID-19 measures were taken, the category (e.g. Social distancing) and its source. Therefore, as was also the case for the Guardian API, potential harm due to the processing of the data is not likely: again, no personal information was included, and the data is on local governmental level, which means that it is very unlikely that any harm would occur due to the processing of the data, for example in terms of privacy or harm to groups.

The DEDA-tool was used to assess the project's ethical position. Several topics were covered to assess how ethically responsible the current project is. One of these topics was Algorithms. It is necessary for a research team to have at least one person who can explain how the used algorithms or models work. Indeed, this is the case for our project, as all team members have sufficient background knowledge of the methods used, and at least one person has investigated how the algorithms work in detail.

In addition, the quality and source of the data is important to assess. ACAPS is an organization of independent specialists in humanitarian needs analysis and assessment, that is not affiliated to any other organization, which helps guarantee that the ACAPS analysis is objective and evidence-based. In the same manner, the Open Platform, containing the Guardian API, is a public web service for accessing all the content the Guardian creates. It might be the case that the data has a best before date, as the data uses information on measures taken by countries, as well as articles. It might be relevant to use articles that are not too old, and in line with the measures taken against COVID-19.

Another topic covered in the DEDA-tool is Anonymization. Data that are not profoundly anonymized might bring for example the ethical danger of exposing one particular individual, his or her interests and preferences. Anonymization is not necessary for our project, as no personal information is used, and the data only includes information on articles and local governmental-level variables. In terms of visualization, it might be possible to visualize the ACAPS data in terms of a timeline with measures taken for each local government by date.

According to the DEDA-tool, access is another relevant topic, as not every dataset should be freely accessible to protect confidentiality of citizens. A second common issue with access is that third commercial partners might be interested in datasets, which might cause further ethical challenges that should be investigated with care. Both data sets used in the current project are public and do not contain any information that might harm confidentiality of citizens. Therefore, no issues in terms of access are expected.

According to the tool, sometimes data could be reused in another context than your specific project. This should be handled with care because the data may lose its validity. In the current case, the 'best before date' should be taken into account. This means that it is only relevant to combine both data sets (Guardian API and ACAPS) when they align well in terms of dates, as it is only possible to assess the analyze the current pandemic's media coverage in different phases if the dates align.

Data projects often have an impact on the livelihood of citizens. Political parties, citizens, lawyers, or activists might use their rights to inquire about data projects. However, as the data used in the current project is textual (Guardian API) and country-level based (ACAPS), it is not expected that any harm to the livelihood of citizens will occur. Regulations and laws related to the current project are therefore not necessarily relevant.

In addition, to be transparent in data projects means to be able to explain the data set and its origins. Transparency also means to be able to explain the models and algorithms used to turn data into actionable

information. For the current project, transparency can be guaranteed regarding the data sets and their origins, as was explained in the above. In addition, the models and algorithms are known in detail by at least one member of the project team. All measures taken by the countries involved are publicly announced and the articles of the Guardian are publicly available. No harm in terms of personal livelihoods or civil rights for example is expected, as no sensitive information is included.
In terms of privacy, no personal or sensitive data is used in the current project, therefore, it is not expected that any violations of privacy will occur.
Furthermore, biases are severe issues in data analysis. A biased dataset, model or algorithm produces results that differ from the reality they aim to describe. In the case of the current project, the sample pf the ACAPS data used can be considered a truthful representation of the population, as the data contains all information on measures taken by countries of all continents concerning COVID-19. However, as was already mentioned before, for the Guardian API, only articles published by The Guardian are included. Therefore, the current project does not include full media coverage.
Informed consent involves the approval of a person to provide information or to participate in a research project. Informed consent, which is the final important topic in data projects, is not applicable to the current project, as already existing data is used that is not collected from individuals. Rather, it is based on articles and local governmental-level variables.

References

[1] Seon-Kyoung An and Karla K. Gower. "How Do the News Media Frame Crises? A Content Analysis of Crisis News Coverage." In: *Public Relations Review* 35.2 (2009), pp. 107–112. DOI: 10.1016/j.pubrev.2009.01.010.

[2] Mike Conway. "The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis". In: *Journalism & Mass Communication Quarterly* 83.1 (2006), pp. 186–200. DOI: 10.1177/107769900608300112.

[3] Paul DiMaggio. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." In: *Poetics* 41.6 (2013), pp. 570–606. DOI: 10.1016/j.poetic.2013.08.004.

[4] James Lappeman & Robyn Clark & Jordan Evans & Lara Sierra-Rubia & Patrick Gordon. "Studying social media sentiment using human validated analysis". In: *MethodsX* 7 (2020). DOI: 100867.

[5] Shanto Iyengar. "Framing Responsibility for Political Issues." In: *The ANNALS of the American Academy of Political and Social Science* 546.1 (1996), pp. 147–157. DOI: 10.1177/0002716296546001006.

[6] J. Matthes and M. Kohring. "The Content Analysis of Media Frames: Toward Improving Reliability and Validity". In: *Journal of Communication* 83.1 (2008), pp. 258–279. DOI: 10.1111/j.1460-2466.2008.00384.x.

[7] Yotam Ophir. "Coverage of Epidemics in American Newspapers Through the Lens of the Crisis and Emergency Risk Communication Framework." In: *Health Security* 16.3 (2018), pp. 147–157. DOI: 10.1089/hs.2017.0106.

[8] Tsung-Jen Shih. "Media Coverage of Public Health Epidemics: Linking Framing and Issue Attention Cycle Toward an Integrated Theory of Print News Coverage of Epidemics." In: *Mass Communication and Society* 11.2 (2008), pp. 141–160. DOI: 10.1080/15205430701668121.

[9] Dror Walter and Yotam Ophir. "News Frame Analysis: An Inductive Mixed-Method Computational Approach." In: *Communication Methods and Measures* 13.4 (2019), pp. 248–266. DOI: 10.1080/19312458.2019.1639145.