Using transformers-based NLP to set and detect news frames on COVID-19 coverage

Keywords: transformers; nlp; topic modelling; news framing; covid-19

Extended Abstract

Several disciplines, such as psychology, political science and communication, use framing analysis to investigate how stories, events or subjects are presented to the public and how this process interferes with understanding these issues. This methodology is usually based on the classic notion that the information is framed in a central organizing idea or storyline [1]. When the framing process is carried out from journalistic instances, it is called news framing.

A common approach of studies that aim to identify news frames is to conduct an in-depth reading of the content produced by media outlets, attributing to each unit of analysis (e.g., news stories, reports, editorials, headlines) one or more frames. To overcome manual text analysis limitations, recent studies have sought to automate frame detection using natural language processing (NLP) and machine learning.

Some protocols suggest performing topic modelling in a textual database, using clustering techniques to verify the relationship between topics, and identifying frames based on this association [2]. But while many of these studies still rely on NLP models based on word co-occurrence, in this paper we test a technique of topic modelling that uses Bidirectional Encoder Representations from Transformers (BERTopic) [3] both to create a framework and to use it to automatically detect frames in new documents. The objective was to investigate the model's performance in generating contextually coherent topics and predicting new instances.

The case study was applied in the news coverage of COVID-19. The dataset contains all news stories about the pandemic published in 2020 by *The New York Times*, *The Guardian* and *China Daily*, totalling 71,128 documents. Following Liu et al. [4], we used only the headlines for the analysis, as we consider them the journalistic element that contains the centralizing idea – i.e., frame – which the newspaper chose to highlight.

First, we randomly split the dataset into train (80%) and test (20%). In the train dataset, we applied the standard five steps of the BERTopic algorithm for unsupervised topic modelling: (i) sentence-transformers for extracting embeddings; (ii) UMAP for dimensionality reduction; (iii) HDBSCAN for clustering; (iv) CountVectorizer for tokenization and stopwords exclusion; and (v) c-TF-IDF to select the words most related to the topics. The class-based TF-IDF is a variation of the traditional TF-IDF that compares the importance of words between documents, considering a cluster as a single document.

We conducted successive tests to reach a number of topics that would reduce the outliers and meaningless topics. The best result was achieved with 40 topics. Following Smith et al. [5], the second stage consisted of manually labelling the topics based on the analysis of the most significant words of each topic and its most representative documents (based on the cosine similarity between the c-TF-IDF representations).

The categorization process was done in three levels: (i) topic labelling – more descriptive and document-oriented (e.g., *outbreak in China*); (ii) thematic labelling – more generalist and cluster-oriented (e.g. *outbreak in China* and *outbreak in Italy* gathered into the theme *disease*

spreading); and (iii) frame labelling – broader and corpus-oriented (e.g. *disease spreading*, *vaccine* and *medical care* into the *medical/scientific/health* frame).

The 40 topics were grouped into four frames, adapted from previous studies [6]: (i) *medical/scientific/health* – disease spreading, tracking and tracing, vaccine developing, drug trials, health care, etc.; (ii) *social/economic disruptions* – non-health consequences of the pandemic, such as the impact on cultural events, sports, economy, etc.; (iii) *individual response* – information directed towards readers to take action or reduce risks, such as social distancing and wearing a mask; and (iv) *organizational response* – information about the actions taken by governmental or non-governmental organizations in combating the health crisis.

In the last step, we used the already-categorized train dataset as a model to detect the frames in the test dataset. The classification was made with an in-built function that calculates the probability that a document belongs to a topic (and, consequently, to a frame) based on the previous cluster representation performed by HDBSCAN.

We observed that the distribution of frames predicted in the test dataset followed the same proportion as in the train dataset, with a significant predominance of the *medical/scientific/health* frame, followed by the *social/economic disruptions* frame and, to a lesser extent, the *individual response* and *organizational response* frames (Figure 1a). A check of a random sample of 100 headlines on the test dataset (25 of each predicted frame) reached a range of precision (true positives) between 36% (*individual response*) and 88% (*social/economic disruptions*) (Figure 1b).

Our results show that the use of a transformer-based NLP technique was able to generate semantically coherent topics. Associated with an expert-driven approach to manual labelling, it was also possible to categorize them into significant news frames showing how important newspapers from three countries framed the COVID-19 pandemic, contributing to science and health communication studies.

Regarding the prediction of new instances, the algorithm presented different agreement rates, whose differences should be further explored. A possible explanation is that precision was higher in frames with better-delimited subjects (e.g., *social/economic disruptions* and *organizational response*), while the classification of *medical/scientific/health* and *individual response* may have been hampered because they contain interchangeable themes and topics.

References

[1] Gamson WA, Modigliani A. Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach. *American Journal of Sociology*. 95(1):1–37, 1989.
[2] Walter D, Ophir Y. News Frame Analysis: An Inductive Mixed-method Computational Approach. *Communication Methods and Measures*. 13(4):248–66, 2019
[3] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:220305794*, 2022.
[4] Liu S, Guo L, Mays K, Betke M, Wijaya DT. Detecting Frames in News Headlines and Its Application to Analyzing News Framing Trends Surrounding U.S. Gun Violence. *Proceedings of the 23rd Conference on Computational Natural Language Learning*, 2019.
[5] Smith A, Tofu DA, Jalal M, Halim EE, Sun Y, Akavoor V, et al. OpenFraming: We brought the ML; you bring the data. Interact with your data and discover its frames. *arXiv:200806974*, 2020.
[6] Kott A, Limaye RJ. Delivering risk information in a dynamic information environment: Framing and authoritative voice in Centers for Disease Control (CDC) and primetime broadcast news media communications during the 2014 Ebola outbreak. *Social Science & Medicine*. 169:42–9, 2016.

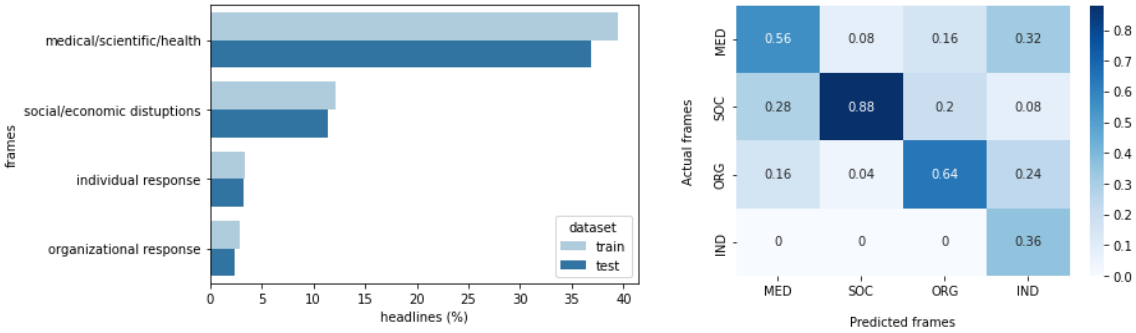

Figure 1a (left). Proportion of frames in the train and test datasets; Figure 1b (right). Agreement rate between actual and predicted frames