doi.org/10.26398/asaproc.0029

Public opinion and textual data analysis: Towards an automatic framing approach

Alberto Maria De Mascellis[a], Michelangelo Misuraca[b]
[a] DiSS, University of Naples Federico II, Naples, Italy
[b] DiScAG, University of Calabria, Rende Italy

1. Introduction

Facing the ever-increasing need to process large amounts of information in an automated way (Matthes and Kohring, 2008), the attention of social and political researchers has turned to a large branch of communication which concerns the automatic identification and the analysis in a collection of texts of *frames*. In public opinion studies, one of the most incisive definitions of *framing* is the one provided by Riker (1986): "framing is the central process by which the government or the journalists exert political influence among each other or on the public". However, a more operational definition comes from Entman (1993), who speculated that newspapers operate through *selection* and *salience*. In other words, anyone who provides information could "select the aspects of a perceived reality and make them more salient within a text, thus promoting a particular definition of a problem, a causal interpretation, a moral evaluation and/or a recommendation for the object described". Entman refined his view in a study dated 2003, streamlining and expanding the definition of framing: "framing is selecting and highlighting facets of events or problems and creating connections between them to promote a particular interpretation, evaluation and/or solution".

The reference to the connections/relationships of elements is particularly interesting because it implies that the frames encompassed in a text can be seen as complex structures involving multiple dimensions. Furthermore, this interpretative paradigm seems to be coherent with the definitions of frames typically used in the qualitative literature of political studies, in which frames are seen as *closed field of meaning* (Schütz, 1972), *interpretative package* (Gamson and Modigliani, 1989), or *boundary of expectations* (Goffman, 1993). An analysis of the reference literature about framing theory suggests that texts contain *framing judgments* in the form of specific keywords, stock phrases, stereotyped images, and favourite sources. Fairhurst (2005) talks about frames in terms of "a choice of language aimed at framing people's actions as if they were framed by a telescope". In contrast, Iyengar (1994) defines them as a "subtle alteration in the statement or presentation of both a judgment and problematic choices". From a quantitative viewpoint, one of the research paths that seem to be more interesting relies on the possibility of treating the frames as communities of terms that occur in a text, obtained by text mining techniques based on network analysis (Misuraca and Spano, 2020). We can claim that most human activities can be modelled as networks (or graphs), represented as nodes connected by some criteria (Patgiri et al., 2023).

This work intends to pave the way to an automatic framing approach based on community detection. Several aspects have to be considered. First, there is a lack of a universal definition of community since network structures cannot be mapped *a priori* without some form of arbitrariness. It is then necessary to consider techniques capable of inferring distinct partitions without referring to preconceived structures or ground truths. Moreover, as the complexity of the task increases with the amount of data, it is important to consider computational costs (Nicholls and Culpepper, 2001). Finally, once the frames have been obtained, it is important to establish a procedure that allows for the unique identification and correct interpretation of the partitions.

To test the effectiveness of our proposal, we carried out a preliminary case study by investigating a topic broadly discussed in Italy in the last years, the so-called "reddito di cittadinanza" (RC, citizenship income). After collecting the articles published from 2018 to

2023 by five prominent Italian newspapers, we mapped the frames used to present this topic to public opinion and their temporal evolution.

2. Materials and methods

To understand in which way RC has been released to public opinion and explore which frames have been eventually used by different counterparts representing conflicting interest groups, we collected the articles published by five Italian newspapers (Il Corriere della Sera, Il Giorno, Il Resto del Carlino, La Nazione, and La Stampa) in the period that goes from May 2018 to June 2023. The articles were selected from the *Nexis Uni* repository[1] with a relevance criterion and an imposed limit of one thousand documents per search. The newspaper's name, publication date, the section containing the article, the title and the full-text body have been retained from each original document. After filtering out the articles with a length lower than 200 characters (associated with short comments on video articles or advertisements) and articles with missing meta-data, the total number of documents in the collection amounted to 3891.

Since data embodied in texts are unstructured, it was necessary to pre-process the collection with several procedures:

- removal of special characters, punctuation, numbers, extra whitespaces, and URLs;
- recoding of specific terms in unambiguous forms (e.g., names of politicians or parties' names);
- lemmatization of the text to bring each term back to its basic form;
- removal of stop words with null analytical value (e.g., conjunctions).

It was also decided to reduce the heterogeneity of the texts by keeping only the lexical POS (parts of speech), specifically nouns, adjectives, and proper nouns. After structuring the articles as vectors and juxtaposing them in a lexical table, we built a *term*×*term* co-occurrence matrix and depicted the knowledge structure of the collection as a network.

One of the most used approaches to study networks is their deconstruction into sub-units, the so-called *communities* (Fortunato and Castellano, 2007), viewed as groups of nodes that have a strong internal structure but weak external connections. Several computational methods have been developed over time to identify communities. Considering the nature of the data to be explored, extracted from texts written in natural language, and to implement an automatic approach to frame analysis, we decided to use an algorithm that relies on the concept of modularity (Girvan and Newman, 2002). Aiming at assigning each object/keyword to a single community, these algorithms operate very differently with respect to the divisive algorithms that remove inter-community links and the agglomerative ones that instead merge the most similar nodes (Radicchi et al., 2004). These methods try to maximize an objective function and return the modularity of the formed partitions, i.e. a scalar value between -1 and 1 that measures the density of the edges within communities compared to those formed between them (Newman, 2006). In particular, we referred to an approach known as *Leiden* algorithm (Traag et al., 2019), which starts from an already consolidated approach known as *Louvain* (Blondel et al., 2019) overcoming some of its well-known problems, such as the presence of bad internal connections between the elements of a community. To guarantee well-connected communities, the Leiden algorithm essentially operates by considering the local movement of nodes and refining the initial partition. An aggregated network is obtained starting from $P_{refined}$ instead of P (as happens in the Louvain approach) so that the algorithm can identify high-quality partitions more effectively. In more specific terms, it starts with a singleton partition, where each node is in its own community and then joins the nodes locally.

We applied the Leiden algorithm to our collection and obtained the communities of meaning to highlight the different frames used by the newspapers. The analysis has been carried out in the `R` environment with the `igraph` library after pre-processing the data with `quanteda`. In

[1] www.lexisnexis.com

the following, some early findings of the strategy are presented and discussed.

3. Preliminary results

In Tab. 1, for each newspaper, we reported the total number of articles included in the analysed collection and the average length (in characters) of each article.

Table 1: Main figures on the analysed collection.

Newspaper	N. of articles	Avg. length
Corriere della Sera	764	894
Il Giorno	887	2645
Il Resto del Carlino	779	427
La Nazione	762	2294
La Stampa	851	3536

We noted a heterogeneous average length, ranging from 894 characters (Corriere della Sera) to 3536 (La Stampa). It should be noted that for La Stampa all the 851 articles available in the repository are concentrated in 2018–2019.

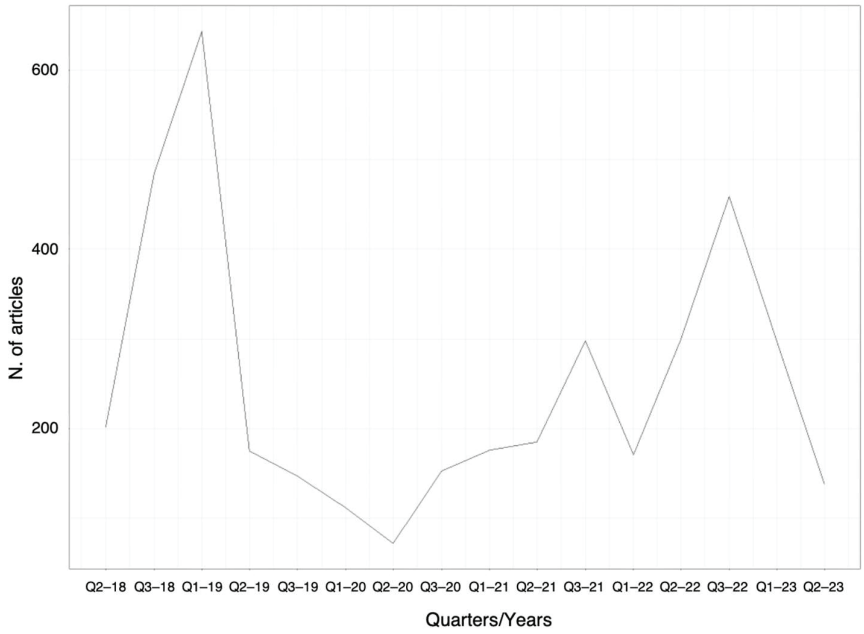

Figure 1: Year-wise distribution of articles about RC (2018–2023)

The chart reported in Fig. 1 is influenced by the latter anomaly, although it is still possible to note a peak in January 2019, the date of the Decree Law n. 4 that established the RC, and the declining curve of September-December 2022, after it has been revoked with the Budget Law (art. 1, subsec. 318). In Fig. 2, we can see the community detection results represented with a GEM force-directed graphic layout. We specify that the weights have been re-attributed based on membership for the sole purpose of an easier reading. The size of the nodes is scaled according to the modularity.

Figure 2: Graph of the main communities/frames about RC

We found multiple communities of meaning that can be structured as frames to guide the readers' debate. It is possible to group the communities in 4 different categories of frames:

- frames that could be referred to as the civil-social aspects of the RC, denoted by communities of terms such as *projects*, *inclusion* and *Puc* (projects useful to the community)
- which the beneficiaries of RC should be required to carry out), but also terms like *safeguard*, *cultural* and *artistic*;
- frames linked to the everyday-economic aspects, represented by the terms *family*, *child*, *age* and *pension*, but also *small-medium enterprises*;
- frames of political-economic nature, represented by all the names of politicians (*Di Maio*, *Salvini*, *Tria*) and by communities of a more complex economic nature, such as *Pil*, *deficit* and *manoeuver*, or *absolute*, *poverty* and *contrast*.
- frames of institutional-judicial type that refer to communities of terms such as *Inps*, *application* and *submission*, *controls*, *activities* and (job) *placement*, but also *substitutive* and *false* near *declaration*.

4. Final remarks and future developments

Bearing in mind that this is an embryonic proposal, the automatic approach to frame detection here introduced offers itself as a significant solution not only in terms of interpretation but also of the readability and immediacy of the outputs, even if they were to be provided to journalists and policy-makers, to broaden their communication and workflows. What we reported aligns with the results of other qualitative and quantitative studies on the surveyed subject. We highlighted, for example, how the contribution of the RC to *absolute family poverty* has been strongly debated, but also to RC as "a political problem, even before a policy one" (Vittoria, 2020). Other scholars described the problem as "individuals penalized in the access phase", in line with our outputs on the institutional dimension (Sgritta, 2020). In the interpretation phase, these data could also suggest communication gaps. For example, no relevant communities are referring to terms like *north* and *south*, while there are terms connected to the political-economic macro-category like *Italian* and *foreign* near *citizen*.

This work intends to be the first step towards an automatic approach to frame analysis, investigating in-depth the available alternatives and structuring a path that can account for a phenomenon as complex as framing. Further studies can certainly implement new data sources, new strategies in identifying communities and eventually a more effective way to connect the resulting frames more organically with the framing entities, such as newspapers or even the debate information in its entirety, taking into consideration other determining factors such as the temporal or spatial axis.

References

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, **2008**(10), P10008.

Entman, R.M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, **43**(4), pp. 51-58.

Entman, R.M. (2003). Cascading activation: Contesting the White House's frame after 9/11. *Political Communication*, **20**(4), pp. 415-432.

Fairhurst, G.T. (2005). Reframing the art of framing: Problems and prospects for leadership. *Leadership*, **1**(2), pp. 165-185.

Fortunato, S., Castellano, C. (2007). Scaling and universality in proportional elections. *Physical*

Review Letters, **99**(13), 138701.
Gamson, W.A., Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, **95**(1), pp. 1-37.
Girvan M., Newman M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, **99**(12), pp. 7821-7826.
Goffman, E. (1993). *Rahmen-Analyse: Ein Versuch über die Organisation von Alltagserfahrungen*. Suhrkamp Verlag.
Iyengar, S. (1994). *Is Anyone Responsible? How Television Frames Political Issues*. University of Chicago Press. Chicago (IL).
Matthes, J., Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, **58**(2), pp. 258-279
Misuraca, M., Spano, M. (2020). Unsupervised analytic strategies to explore large document collections, in *Text Analytics: Advances and Challenges*, eds. D.F. Iezzi, D. Mayaffre and M. Misuraca, Springer Nature, pp. 17-28
Newman, M.E.J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, **103**(23), pp. 8577-8582.
Nicholls, T., Culpepper, P.D. (2021). Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, **38**(1-2), pp. 159-181.
Patgiri, R., Deka, G.C., Biswas, A. (2023). *Principles of Big Graph: In-depth insight*. Academic Press.
Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, **101**(9), pp. 2658-2663.
Riker, W. (1986). *The Art of Political Manipulation*. Yale University Press.
Schütz, A. (1972). *Gesammelte Aufsätze: I Das Problem der sozialen Wirklichkeit*. Springer Science
Sgritta, G.B. (2020). Politiche e misure della povertà: il reddito di cittadinanza. *Social Policies*, **7**(1), pp. 39-56.
Traag, V.A., Waltman, L., Van Eck, N.J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, **9**(1), 5233.
Vittoria, A. (2020). La "scomparsa dei poveri". Una prima valutazione di policy sul Reddito di Cittadinanza. *Social Policies*, **7**(3), pp. 525-544.