Methodology of narrative analysis sub-task for SMART project

Andrzej Nowak,
Magda Roszczyńska-Kurasińska,
Wojciech Borkowski,
Karolina Ziembowicz,
Mikołaj Biesaga,
Szymon Talaga
The Robert Zajonc Institute for Social Studies
University of Warsaw

Abstract

This white paper provides a self-contained, high-level overview of the methodology of narrative analysis that is the subject of Workpackage 2 / Sub-Task 2.6.2. The goal is to detect and describe most prevalent narratives concerning social and policy issues, especially in relation to EU, as perceived through discourse in leading weekly opinion magazines in various European countries. The analysis will combine modern quantitative approaches (collection of big data from social networks and unsupervised machine learning) with qualitative insights in order to produce in-depth and actionable results.

Narratives can be defined as prevalent ways of thinking and telling about the world, especially in the context of important social and policy issues, which have a broad but not necessarily unequivocal social support. As such, they have a power to shape public opinion and decision making processes. They can be used to mobilize social support both for good and bad. This seems to be especially important in this historical moment when the new media landscape has fundamentally reshaped our collective means of communication facilitating local and global cooperation but also providing entirely new tools of propaganda and misinformation, which are often popularly referred to as *fake news*. Also it is quite apparent that many European societies are rapidly changing and entangled in often heated and very polarizing public debates and conflicts. This social rupture seems to be at least partially resulting from a significant decrease of the pool of shared social norms and values, and this in turn can be thought of as a result of coexistence of competitive and outrightly incompatible narratives concerning very foundations of social and political life. Hence, understanding of what are the narratives that are shaping our current historical moment in Europe seems to be of crucial importance. Such a knowledge would help identify not only the points of the most ferocious discoursive conflicts but also and perhaps more importantly the remaining issues and values that are still common for everyone.

The aim of this research is to develop and apply an empirical method of measuring and describing social narratives. The main objective is to describe the current most prevalent narratives that shape social life in various European countries as well as identify most conflictual and shared regions of discoursive space.

Due to preliminary and pioneering character of the project the scope of the analysis had to be carefully chosen and somewhat limited. Only narratives as presented in the discourse of leading weekly opinion magazines will be considered. This approach is favorable as it can provide a clear-cut definition and enumeration of data sources. At the same time opinion magazines are important actors that both influence and reflect the state of ongoing public debates and as such can be considered good proxies for estimating general public narrative dynamics.

Therefore, the analysis will use content collected from social media profiles of selected publishers together with content of the linked articles. This data will form a comprehensive corpus of text that will be used to discover the most important narratives in an empirical, evidence-based fashion using specialized unsupervised machine learning algorithms from the topic modeling family such as *Latent Dirichlet Allocation* (Blei et al., 2003). Subsequently, the discovered narratives will be classified and interpreted using a combination of quantitative methods typical for natural language processing and qualitative analysis based on expert reading of random samples of texts typical for specific narratives.

Finally, narratives information will be combined with sentiment data (emotional profile of texts) in order to provide synthetic description of discoursive space in various countries. Here discoursive space is defined as a joint representation of how different narratives are used by different actors (publishers) and what is the dominant emotion associated with them. This approach will enable identification of both conflictual and shared narratives and parts of the discoursive space.

Data collection

For each country[1] a list of leading weekly opinion magazines (actors) will be established based on industry rankings of circulation (up to ten biggest titles will be included). Then, for each title a year-long history of *Facebook* posts will be extracted. This will include both posts content, numbers of reactions and shares as well as `URLs` of any linked articles. *Facebook* data extraction will be carried through by our business partner *Sotrender*. Gathered article links will be subsequently used to scrape actual content of articles from publishers websites. Hence, the final corpus of documents for a given country will consist of a set of *Facebook* posts and content of linked articles as well as basic information on social media reactions. This will enable narratives discovery as well as analysis of discourse of specific actors and assessment of similarity between various actors.

[1] The exact set of countries that will be considered is being worked out at the time of writing of this paper.

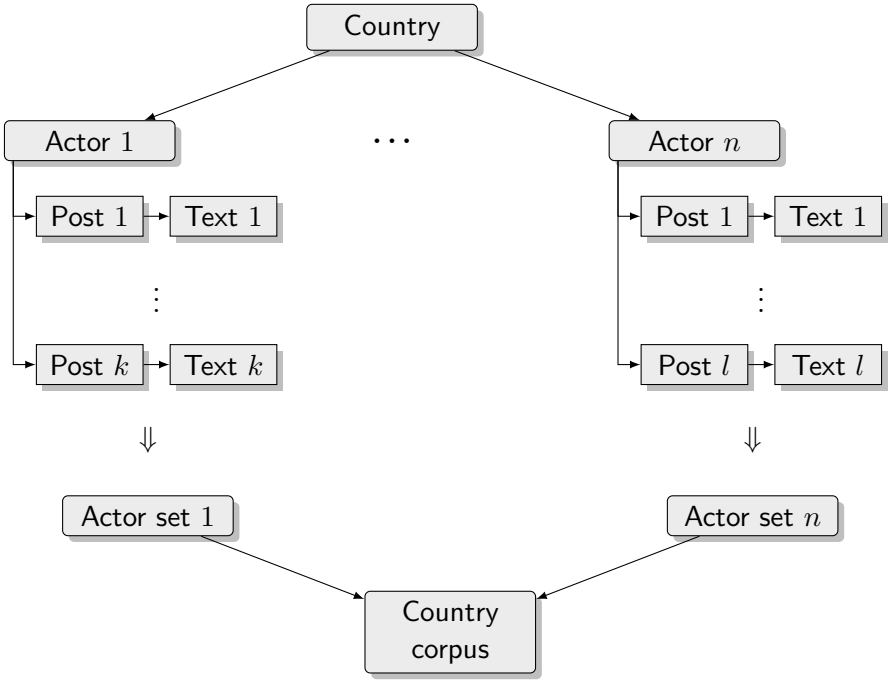

Figure 1. Data collection process for a single country. A country corpus is composed of several *actor sets* which consist of post and articles published by a specific actor.

Analysis

Data preparation

First, the data will be prepared for the main analysis. For each country corpus a subset of documents related to politics, economy and social issues (as opposed to articles on topics such as science and technology, culture or fashion) will be determined.

Sentiment analysis

Each document will have a sentiment score assigned to it. Bipolar positive-negative scoring will be used, so every document will be described not only in regard to what topics it discusses but also in regard to emotional valence of its content. It is most likely that dictionary-based sentiment analysis techniques will be used.

Topic modeling and narrative discovery

Document corpuses will be used as input data for topic modeling procedures. We plan to use *Latent Dirichlet Allocation* (LDA) as our main algorithm for topic discovery. This is a modern and industry tested technique that can cope with most of the technical difficulties typical for topic modeling problems (such as word polysemy) (Blei et al., 2003; Kosinski, Wang, Lakkaraju, & Leskovec, 2016).

It is quite likely that the first step of the analysis based on LDA will lead to discovery of wide arrays (tens, hundreds or even thousands) of particular topics that are too specific to be considered narratives. In such a case an additional step will be introduced and data

dimensionality will be reduced using the PCA algorithm (Murphy, 1991). As a result much smaller sets of higher-order topics will be discovered and these topics could justifiably be considered narratives.

Narratives interpretation

Discovered narratives will be subsequently interpreted. This part of the analysis will be based on mixed methods approach. First, quantitative summary reports will be prepared for every narrative. The reports will focus on basic high-level summarizations such as distributions of words and phrases, also presented graphically via word clouds and/or semantic networks. Additionally informational measures such as `tf-idf` (Anand & Jeffrey, 2011) will be used to identify words and phrases that are most specific and descriptive for a given narrative.

The final stage will be dedicated to qualitative analysis of random samples of texts typical for specific narrative by experts with a sound knowledge of politics, social issues and media landscape of a given country. Their work will be facilitated by quantitative reports which are supposed to provide useful hints for interpreting narrative data.

Estimating shared and conflictual parts of discoursive spaces

Description of narratives will serve as a starting point for additional analysis that will focus on identifying the most conflictual and shared narratives. Every actor (publisher) will be described in regard to every narrative in terms of importance of this narrative (i.e. how often the actor discusses it, its popularity in social media etc.) and sentiment towards it (what is the emotional attitude to it). This will enable identification of clusters of actors that tend to think similarly and differently about a given issue/narrative.

Summary

The analysis will hopefully not only provide useful information on prevalent discourses that shape public opinion in Europe, but will also shed some light on the question of which issues can be used to bring more people together. This seems to be an important problem in the context of contemporary media landscape which often tend to unnecessarily amplify social divisions.

Moreover, it is hoped that this pilot study will constitute a good starting point for developing a framework for systemic studying of public discourse.

References

Anand, R. & Jeffrey, D. U. (2011). Mining of massive datasets. *2011-01-03. Http://Infolab. Stanford*. doi:10.1017/CBO9781139924801. arXiv: arXiv:1011.1669v3

Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., & Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022. doi:10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining Big Data to Extract Patterns and Predict Real-Life Outcomes. *21*(4), 493–506. doi:10.1037/met0000105. arXiv: arXiv:1011.1669v3

Murphy, K. (1991). *Machine Learning: A Probabilistic Perspective*. doi:10 . 1007 / SpringerReference_35834. arXiv: 0-387-31073-8