A framework for the diachronic and linguistic analysis of online polarized discussions

Erica Cau[1], Virginia Morini[2,3], and Giulio Rossetti[3]

[1]Philology, Literature and Linguistics department, University of Pisa, Italy
e.cau@studenti.unipi.it
[2]Computer Science department, University of Pisa, Italy
virginia.morini@phd.unipi.it
[3]KDD Laboratory, ISTI, National Research Council, Pisa, Italy
{virginia.morini, giulio.rossetti}@isti.cnr.it

Social Network sites (SNSs) have disrupted and reshaped how personal communication is perceived and information is spread, in favor of a newer and faster way of sharing ideas and participating in public discussions. Despite the impressive benefits, social media platforms bring an equivalent amount of downsides and polluting phenomena that cannot be ignored. Among these issues, we want to shed light on *echo chambers* (ECs), i.e., polarized systems in which information, ideologies, and beliefs are repeated and amplified as the only truthful view of reality, without contemplating rebuttal or openness to different ideas.

Although many efforts were made over the years to detect and characterize echo chambers across SNSs, it is worth mentioning that the lack of a standard definition has led to a fragmentation of approaches, each tailored to a specific social platform and thus hard to be generalized. Moreover, current research generally ignores the diachronic evolution of these systems, constraining the analysis to static observations and - in the end - providing an overestimate of users' sociality. Lastly, another issue is that scarce to non-existent attention was given to characterizing the linguistic behaviors of the users therein.

In this preliminary work, we propose a framework to track ECs' temporal evolution as well as to characterize the linguistic productions of the users inside and outside ECs. The framework has been tested on a controversial topic, namely American politics during the first two years and a half of Donald Trump's presidency (i.e., January 2017-July 2019) in the specific context of Reddit discussion boards about Gun Control, Politics, and Minorities Discrimination. Particularly, the analysis is conducted on the interactions between people who sided with one of the two ideological extremes of the political spectrum (i.e., *Pro-Trump* and *Anti-Trump* users) across five semesters.

The proposed pipeline is built on the ECs detection framework defined in [3]. At first, we propose to model the temporal evolution of the interactions between users via node-attributed snapshot networks so that, for each timestamp, each user is characterized by an attribute describing their own political leaning. Then, we leverage a Labeled Community Detection algorithm (we suggest EVA [1]) to extract, for each snapshot, homogeneous communities both from a topological and ideological point of view. We assess the risk of them being echo chambers via two evaluation measures, namely Purity[1] and Conductance[2] to ensure that most interactions take place between like-minded users inside the very same community. Further, we propose to measure the ECs (dis-)similarity between adjacent snapshots in terms of users through the Jaccard index. As concerns the semantic dimension of ECs, we focus on text-specific features, emotions, and sentiments expressed through users' words, and finally, the subjects of their discussions, extracted through BERTopic [2]. This analysis is then enhanced by relating the temporal dimension to the topic to have a fine-grained overview of the users.

By applying the proposed framework to the already mentioned Reddit case study, we provide valuable clues in the formalization and understanding of the phenomenon. First of all, we find that ECs are systems that (when accounting for users' volatility) persist over a long time span as they remain so for up to a year and a half (Figure 1b). Additionally, as time (semesters) passes, ECs tend to keep their internal composition in terms of users.

Regarding the linguistic analysis, the most interesting results are obtained via topic modeling: we show

[1] The product of the frequencies of the most frequent labels carried by nodes of a community.

[2] The fraction of total edge volume that points outside the community.

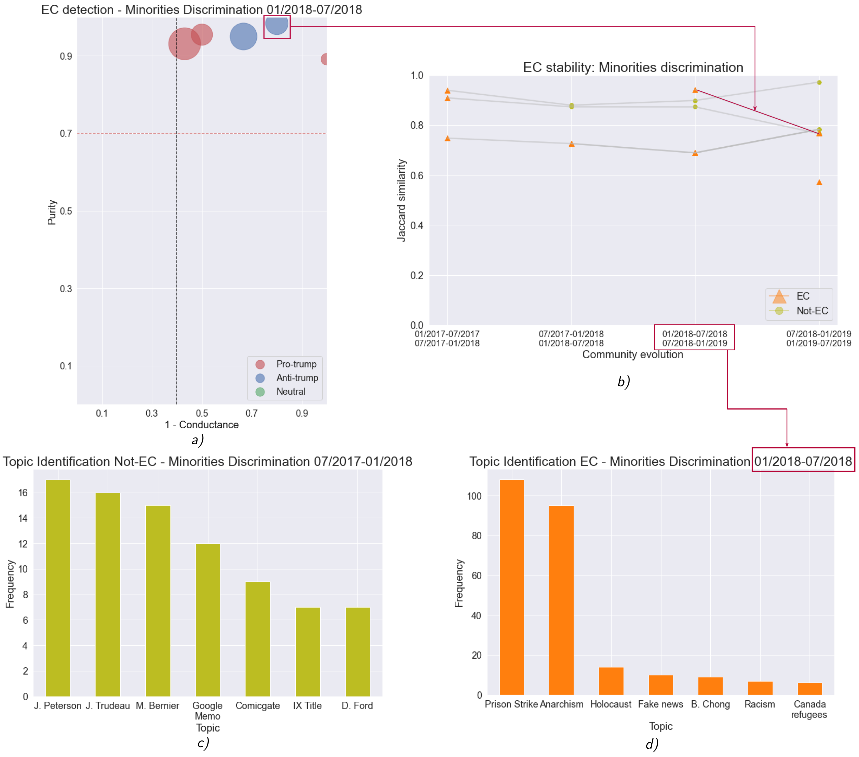

Figure 1: *a)* ECs detection, the community more at risk of being EC is highlighted in red *b)* ECs stability between adjacent semesters, *c)* example of topics discussed outside ECs, *d)* main issues debated inside the highlighted EC.

that the majority of users in communities more at risk of being ECs are prone to mainly debate one specific controversial matter or two different topics both related to a more generic polarizing issue. An example may be represented by an identified EC in which two topics were mainly discussed, the *U.S. 2018 Prison Strike* and the general ideology of *anarchism* (Figure 1d). These two subjects are strongly related since the first event was well-received and supported by anarchic movements. Outside ECs, instead, we notice a different behavior from the users that tend to discuss a variegated set of topics, e.g., Marvel's Comicsgate, Canadian and American politics (Figure 1c).
In conclusion, we propose a platform-independent framework to track ECs evolution and to analyze users' linguistic productions in a standard way. As an evolution of this work, we plan to further validate the obtained results and test the framework on other SNSs to have a wider overview of ECs behavior across different platforms.

Acknowledgements This work was supported by the scheme' INFRAIA-01-2018-2019: Research and Innovation action', Grant Agreement n. 871042 'SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics.

References

[1] Salvatore Citraro and Giulio Rossetti. "Identifying and exploiting homogeneous communities in labeled networks". In: *Applied Network Science* 5.1 (Aug. 2020). DOI: 10.1007/s41109-020-00302-1. URL: https://doi.org/10.1007/s41109-020-00302-1.

[2] Maarten R. Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *ArXiv* abs/2203.05794 (2022).

[3] Virginia Morini, Laura Pollacci, and Giulio Rossetti. "Toward a Standard Approach for Echo Chamber Detection: Reddit Case Study". In: *Applied Sciences* 11.12 (June 2021), p. 5390. DOI: 10.3390/app11125390. URL: https://doi.org/10.3390/app11125390.