UNIVERSITÀ DI PISA

DIPARTIMENTO DI FILOLOGIA, LETTERATURA E LINGUISTICA

Laurea in Informatica Umanistica

Echo chamber and Political Polarization: A time- and linguistic-aware analysis of online polarized discussions

Relatori:

Giulio Rossetti

Virginia Morini

Candidato:

Erica Cau

ANNO ACCADEMICO 2021/2022

Abstract

Social network sites (SNSs) have reshaped how information is spread, in favour of a faster way of sharing ideas and participating in public discussions. Despite the impressive number of benefits, SNSs bring with them an equivalent amount of polluting phenomena that cannot be ignored. Among these issues, we include *echo chambers*, i.e. polarized systems in which information, ideologies, and beliefs are amplified as the only truthful view of reality, without contemplating rebuttal or openness to different ideas. Although many efforts were made to study echo chambers, current research lacks a rigorous analytical framework to analyze EC's diachronic evolution. Moreover, scarce to non-existent attention was given to characterizing the individual behaviours of the users therein. In this thesis is presented a study on echo chamber detection on Reddit discussion boards revolving around the first two years of Trump's presidency. At first, for each chosen topic, we model the interaction network of the users via node-attributed graphs, so that each user is characterized by their own political leaning. Then, we extract communities and assess the risk of them being echo chambers by looking at the topological cohesion and the ideology homogeneity. Afterwards, we focus on their temporal evolution to gain new insights into their stability. The second part of the work is about users and the different ways of acting depending on whether they are trapped or not inside an echo chamber. We analyze their linguistic productions, looking at text-specific features, the emotion and sentiment vehicolated through their words, and finally, the topics they talked about. This analysis is then enhanced by relating the temporal dimension to the topic and sentiment, to have a fine-grained overview of the users. This provides insights into the stability of ECs through time and the tendency of their members to focus on a single controversial topic.

Contents

List of Figures

Chapter 1

Introduction

The birth of Social Network Sites (hereafter, SNSs) brought a revolution in terms of how interpersonal communication is perceived and in the way information spreads, which became immediate, given the absence of time-space barriers. SNSs give users very few limitations in terms of freedom of expression and possibility of building relations with unknown people around the world sharing the same interests.

Despite all their advantages, SNSs carry undeniable consequences that open the doors to polluting phenomena, due to the fact that they naturally inherit many inner biases that inhabit the offline world, to which were added many other biases born in online platforms.

The massive amount of heterogeneous information posted every day by users all around the world may trigger the so-called *cognitive dissonance*, a feeling of discomfort that arises when a person is exposed to information that is not aligned with their own. This feeling, as a consequence, brings users to avoid these potentially contrasting contents via *confirmation bias* and *selective exposure.* Accordingly, they tend to select and share only those contents aligned with their pre-existing beliefs, reinforcing their views even more. These kinds of human biases are further exacerbated by SNSs' algorithms which exploit users' behaviours in order to offer them content as similar as possible to their interests (i.e., *filter bubble effect*), with the aim of maximizing their interactions and engagement.

Several studies underline that online information systems are *polluted*, in a way that these online spaces are biased by the existence of polluted (i.e., polarizing) contents and behaviours that might interfere with the unfolding of public debates and in the opinion formation process.

Among the existent polluted realities, are included *echo chambers*, closed online systems where users' beliefs are amplified and insulated from rebuttal. The danger of echo chambers has often been underlined in the literature since they are usually at the roots of alarming events that might also outbreak in the offline world, leading to further – and more serious – consequences such as interference in elections or pseudo-science movements.

In recent years, a large body of work has addressed this issue, often with the aim to assess the presence of echo chambers in specific SNS platforms. There are still major gaps in this field, due to the fact that echo chambers are often detected as static entities, that is, without keeping into account their temporal evolution, thus leading to an undeniable loss of valuable knowledge. Additionally, to the best of my knowledge, only a few studies have attempted to characterize the behaviours of the users inside these polluted systems.

This thesis stems from the purpose to enrich the body of literature by offering a methodology to assess the diachronic evolution of echo chambers and for characterizing the way users discuss inside and outside them. The definition of the methodology will be followed by a case study centered on *American politics* since for its own nature is divided into two coalitions that foster the debates – and the divisions – in the United States. The timespan under analysis is between 2017 and 2019, so as to cover the first two years and a half of Donald Trump's presidency, which witnessed a clear increase in the division of the supporters of the two political sides. The data for the case study and the framework for the echo chamber detection were taken from [MPR21].

In Chapter 2 a literature review is introduced, to better assess the nature of the biases involved in the formation of *echo chambers*. We also discuss different approaches employed in the literature to address the problem and we introduce some basic notions about *Network Science* and *Natural Language Processing* that will be useful for fully understanding the designed framework.

Subsequently, Chapter 3 introduces the methodology employed to *i)* assess the echo chamber evolution over time and *ii)* characterize the users from a linguistic point of view. As for the first point, we leverage *snapshot graphs* to model the evolution of Reddit users' interactions through five different semesters. Each user is represented by a node in a graph, and carries as metadata their *political* leaning, as described in [MPR21]. After the creation of the dynamic network a Community Detection algorithm will be applied to extract the most likelihood partitions, evaluated by leveraging two measures to detect ideologically and topologically cohesive communities, to distinguish among the ones more at risk of being an echo chamber and the ones with a lower risk.

It is then studied the evolution of these partitions through the Jaccard index; in this way, we will also assess the stability of adjacent snapshots and it will be possible to check whether echo chambers maintain or losetheir strong polarization over time.

Then, we present various approaches to characterize users in different ways. At first, it is discussed to leverage *text specific features* extracted from the textual data; another approach presented is to analyze the sentiment underlining users' textual production via VADER and the emotion conveyed by applying EmoRoBERTa or Zero-Shot Learning. It is finally proposed to leverage BERTopic, a BERT-based model trained for topic modeling, to characterize communities with the core topic discussed.

This framework is then applied in Chapter 4 to a case study about American Politics. For each step of the methodology, we describe and analyze the results, providing possible keys for their interpretation.

Lastly, in Chapter 5 we will draw the conclusions of the work, focusing on the most relevant results, on the discussion of the limitations of this approach as well as the possible future developments of the work.

Chapter 2

State of the art

The purpose of this chapter is to break down – by conducting a literature review – the concept of *polluted information systems*, namely virtual environments dominated by biased contents and behaviours that can foster and ignite problematic phenomena such as *misinformation*, *hate speech* and *cyberbulling.* Although these phenomena are an offspring of the virtual realm, they are at the same time particularly dangerous because they can easily slip into the physical world, leading to concerning effects, i.e., *no-vax* movements, biased political information [KAZ19] and radical extremism [Bri18].

In the first section of the chapter, social networks and the biasing phenomena that inhabit them are described, with the purpose of giving an overview of the factors involved in the development of echo chambers. The second part, instead, examines frameworks and approaches used in the literature to identify polarized systems with a focus on echo chamber detection. Ultimately, the last two sections introduce the fields of research of Network Science and Natural Language Processing and give an overview of the techniques leveraged in this work to assess the presence of echo chambers and study the linguistic productions of the users.

2.1 Social Network platforms and *polluted information systems*

Social network sites (henceforth SNSs) radically changed the way human beings perceive social interactions: a message that once could only be vehicolated through paper and several travelling days, now it is sent almost immediately to the addressee, no matter how far they are. SNSs opened the door to a new way to perceive communication and to get in touch with others: they deconstructed the need to be face-to-face to have a conversation and facilitated the acquaintance of new people sharing the same interests, hobbies or, for example, political ideas. In addition to this, SNSs also drastically changed the way individuals interact with information, allowing them to become actively involved in sharing and discussing news within their network.

Furthermore, the success of SNSs seems to be never-ending: in 2022 over the 63% of the world population has access to the global network and almost 60% are social media users too; to the current date, the most important ones are Facebook and Whatsapp (owned by Meta) and Youtube[1].

SNSs transformed users from mere consumers to producers – ranging from posts to videos – and the novelty lies in being able to reach everyone having an Internet connection, in a *many-to-many* communication [JH16]. A formal definition of SNSs is given in [BE07], where they are described as *web-based services* that allow the users to

> (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.

The increased interconnectivity has also led to ineludible effects on the netsurfers' opinion formation and information diffusion, as SNSs have a key role in spreading information. Despite their benefits, SNSs have unfortunately inherited offline in-

[1] Digital 2022 Global Overview Report: https://datareportal.com/reports/digital-2022-global-overview-report, last visited October 10, 2022

dividual and group biases: several studies have shown that people tend to select news and join discussions that are more aligned to their personal beliefs, in order to avoid a mental discomfort, namely *cognitive dissonance*. The need to avoid this discomfort brings the users to adopt behaviours of *confirmation bias* and *selective esposure*: these two biases, allow them to filter contents in a way that they can fit with their ideas. These biases are exarcerbated by SNSs' biases, such as *algorithmic bias* and *filter bubbles*.

In the following, we will briefly focus on the advantages and disadvantages of SNSs, specifically outlining their role w.r.t. news and political debates.

Social networks are well known for increasing the heterogeneity of content and their importance nowadays is also related to access to news, given that they encourage variety also in political discussions and contents [Bru10a] [Bru10b]. This has the effect to soften the boundaries between news and contents and motivates users to move both in time and space through different pieces of news and discussions accordingly to what Brundidge calls *traversability*. Moreover, Tucker in [TTRB17], points out that SNSs facilitate inclusion since they give the opportunity to everyone to be part of political discussions along with easing the likelihood of finding like-minded people, supporting political candidates or parties and organising protests. One interesting point from Tucker is that social media are like a neutral ground, a tool that may be used by *marginalized groups* to increase democracy in democratic countries but, in contrast, they can be used to manipulate anti-democracy and create autocracies in democracies.

One evidence of the intrinsic value of SNSs for political participation is [Bou15]. Here, Boulianne correlates the usage of SNSs to participation in political life; Vaccari et al. [VVB^{+}15], instead, support several theories. The first one is that the online exposure of users can lead them to open up and express their political leaning, while the second one is that they can increase the awareness of users, who may see themselves as more capable of being included in the political process. Thus, can even lead them to expose themselves via other forms of political participation.

So, on one hand, social media heterogeneity is said to aid political participation, but on the other hand, there are major concerns about *social fragmentation* [Sun18][Pap02].

In [Sun07], Sunstein suggests that fragmentation is potentially involved in *polarization*: people tend to form niches of ideologically similar people [Bri18] in a way that they talk only to each other and mostly listen to their own voices. This hypothesis has been widely dissected in the literature: evidence [QSS16, Bar15, CRA14] shows that, to an extent, fragmentation exists in blog and social media, and the phenomenon continues to be concerning since some of these groups may fall under the dome of polarization. This may result in extremism and, even worse, physical and/or psychological violence. Echo chambers have also been cited as the cause of polarization towards extremism since they are repetitions of the same ideas that are bounced back to the people inside.

Moreover, the large-scale diffusion of useful information often floods in *information overload*. To cope with this, users enact actions that lead to *confirmation bias* and *selective exposure* phenomena. At their roots, there is a mental discomfort called *cognitive dissonance*, which is triggered by opinions or information contrary to one's personal ideologies or beliefs. This theory, elaborated by Festinger [Fes62], consistutes an explanation for that exact distress people feel when exposed to different beliefs or ideas. Likewise, it can be also exploited to understand why the exposition to heterogeneous content may trigger that discomfort. According to this theory, if a person holds items of information that are in conflict with each other, would feel a sensation of discomfort since these items – i.e., behaviours, opinions, fragments of knowledge, feelings – are in a psychologically dissonant relation.

The more the dissonant items, the stronger the effects of cognitive dissonance, and although Festinger mentions some ways to escape cognitive dissonance, he also states that it is not always possible to succeed in doing so. A way to lower the pressure of cognitive dissonance is to exploit *dissonance-reduction* strategies, such as reducing the importance of the item that triggered cognitive dissonance and changing an inner belief to match the dissonant one. A typical example of cognitive dissonance is the smoker who discovered that smoking is unhealthy. He will continue to smoke, but this would be in contrast with his discovery. So, to reduce the distress caused by the dissonance, he could stop smoking, so that his behaviour is aligned with his knowledge; otherwise, he could change what he knows about smoking, persuading himself that smoking is healthy.

A large body of works has explored the effects of cognitive dissonance, both in the offline [Fes62, EL86] and online [BKL$^+$19] [JZLC19] world. Two of them are particularly alarming given that they are involved in social network pollution, namely *selective exposure*[Kla60] and *confirmation bias.* Actually, to avoid the discomfort offered by opposite ideas or behaviours, a person tends to look for information or contents that are coherent with their view – thus leading to selective exposure. Confirmation bias, instead, is the tendency to search for information that supports existent beliefs and ignore what, on the contrary, contradicts them [Pet20], making pre-existing ideas the most credible ones.

In conclusion, the risk of cognitive dissonance and the related phenomena are linked to one of the advantages of SNSs, heterogeneity. As effect of this feature users are exposed to a wide range of information and beliefs, even opposite to the ones they have. Furthermore, this heterogeneity is exacerbated by the typical information overload users experience on SNSs, thus usually leading to confirmation bias.

2.1.1 Polarization

One cannot discuss SNSs without considering the concept of *group polarization.* This concept refers to a situation where groups tend toward more extreme positions than the initial inclination of their individual members [Sun99]. This situation, according to Turner [Tur87], may be compared to what happens to polarized molecules: in fact, people, like molecules, as an effect of polarization, become more aligned in the direction in which they already point. Online group polarization, also called by Sunstein *cyberpolarization* [Sun07], sees a group of like-minded people involved in a discussion. At the end of the quarrel, they will end up having the same idea they had before, but sharpened and more extreme.

According to Sunstein, group polarization is vehicolated by two mechanisms. The first one is *social comparison* [Fes54], that is, the need of individuals to be recognized by the other members of a group as favourably at every cost, also the one of shifting their beliefs or positions towards the dominant one. Another means employed by group polarization is *persuasive arguments*, in a way that an individual is prone to accept the most convincing argument among the ones defended by the group members.

Sunstein also claims that polarization may also have a degree of influence on good deliberation and may also foster extremism and fanaticism.

Group polarization is an issue firmly related to online *echo chambers*, such that in [Sun99] can be found the analogy of *people hearing echoes of their own voices* intrinsic in the concept of echo chambers. The concept of group polarisation is mainly discussed in the socio-political sphere and this work, as will be explained in Chapter 3, will cover an example of political polarization.

2.1.2 Filter bubble

The metaphor of *filter bubble* was introduced by the activist and author Eli Parisier [Par11] to describe the personalized microcosm of information created by personalization algorithms to fit as best as possible to user interests, age, gender and other demographic characteristics.

The most significant effect triggered by filter bubbles is confirmation bias 2.1, due to filtering and personalization algorithms in social media and search engines that are designed to show information and news adherent to the ideas and leaning of the user. As a result, there is a continuous reaffirmation of one's own ideas that weakens the democratic circulation of news and discussions for the sake of user engagement.

Filter bubbles, according to Parisier, are characterized by three dynamics:

- Each user is alone in their own filter bubble;
- Filter bubbles are invisible, the personalization algorithm elaborates an intricate puzzle of assumptions about the user, but the user is not aware of the content of this *personalization*;
- The user does not choose to enter the bubble.

While filter bubbles may be useful to receive relevant news that fit also personal interests, permitting to survive the daily *information overload*, they represent at the same time a danger, given that they show to the user only contents that it may

Source	Definition
Garimella *et al.*, 2018 [GMGM18b]	Situations where one is exposed only to opinions that agree with their own
Cinelli *et al.*, 2021 [CMG$^+$21]	Environments in which the opinion, political leaning, or belief of users about a topic gets reinforced due to repeated interactions with peers or sources having similar tendencies and attitudes
Jamieson and Cappella, 2008 [KHJ08]	We mean to suggest a bounded, enclosed media space that has the potential to both magnify the messages delivered within it and insulate them from rebuttal.
Sunstein, 2007 [Sun07]	Sunstein does not give an explicit definition of echo chamber but he states the following: "In a democracy, people do not live in echo chambers or information cocoons"
Nguyen, 2020 [Ngu20]	An echo chamber is a social epistemic structure from which other relevant voices have been actively excluded and discredited.

Table 2.1: Echo chamber definition according to various authors

like or engage with. This leads to emphasising users' ideas and limits the stream and variety of information they have access to, thus also leading to more concerning consequences, including *polarization* and *echo chambers* [GdMTC21].

2.1.3 Echo chamber

In recent years, a large body of studies have focused on the study of echo chambers because of their closer relation to SNSs' pollution and their possible outbreaks in the offline realms, such as interference on political elections or ethnic stigmatization. Although they affect most SNSs, in the literature, there are attestations of this kind of polluted system also in blogs [GBK09] and forums [WMKL15].

Despite being among the most concerning polluting phenomena of SNSs, the concept of "echo chamber" itself is debated. There is still no formal definition or agreement on the various definitions that have been given over the years. Nonetheless, the effects of echo chambers are widely studied in the literature, both from a theoretical [Sun07] [KHJ08] and an applicative perspective [VPV21] [CMG$^+$21] [BJN$^+$15] with the aim to identify them and consequently, study strategies to mitigate their effects.

Generally speaking, what emerges from the various definitions in Table 2.1 is that an *echo chamber* is a polarized system in which some ideas or beliefs reverberate and are magnified for the effect of the repetition inside this closed system, insulated from rebuttal. Their peculiarity is that they do not allow constructive discussions or questioning, fostering problematic effects like *group polarization*, as stated in [KHJ08] [Sun07]. Furthermore, when talking about echo chambers, Garimella [GDGM18] breaks the term into its constituent: the "chamber", that is the "place" where an opinion is shared by someone, that reverberates through "echo", since it is shared also by other users.

The roots of echo chambers are both psychological and intrinsic features of SNSs. The *selective exposure* and *confirmation bias* may play a role (see Paragraph 2.1), enhanced by SNSs' personalization algorithms that show to a user the most engaging contents, thus avoiding situations of *cognitive dissonance* through *filter bubbles.* A concept often confused with echo chambers is the one of *epistemic bubbles.* According to Nguyen [Ngu20], an epistemic bubble is an epistemic structure where other voices are ignored and omitted, in a passive process. Differently, echo chambers involve an active process of exclusion and invalidation of opposing information and ideas. While this kind of *filtering* can be easily broken via exposition to excluded information, echo chambers are much harder to shatter, because the exposition to heterogeneous information would reinforce the grip of the echo chamber. Nguyen, as a way to break up those polluted systems, suggests working on mechanisms that bring to avoiding opposite views, in a way to rekindle the trust between echo chambers' inhabitants and the world outside the chamber.

2.2 *Polarized systems* detection

Online polarized systems are usually identified, according to Garimella [Gar18] [GDGM18] leveraging two approaches, namely the study of users' interactions and the content read or shared by the members of polarized factions.
An example is the work proposed by Garimella *et al.* [GMGM17], where the polarization identification method falls into the category of approaches that leverage *networks.* Here, the authors study five controversial topics (among them, *gun control*, *Obamacare*, and *abortion*) using Twitter as a source of data, and then track the variations in the interaction networks in order to observe if there are notable changes in network topology and polarization levels.
Garimella et *al.* first collected the data from Twitter, then they constructed two

different types of networks: the *endorsement network*, using retweets, and the *communication network* leveraging the replies. The third step is to cluster the network to identify the two sides of the controversy using the *METIS* [KK98] algorithm. Ultimately, network science measures (i.e., clustering coefficient, tie strength), content measures (i.e., sentiment measures, Jensen-Shannon divergence) and *Random Walk Controversy*. This measure allows to quantify the likelihood that a given user is exposed to authoritative content of opposite leaning and was used to test whether the system was polarized. The most interesting result was that an increase in the controversy measure corresponded to an increase in interest in the discussion.

Conversely, in [BMA15] the main focus is textual content. This time, the data are taken from Facebook and describe over 10 million U.S. users who have made public their political leaning and the content published over six months (July 2014 - January 2015). Then, the contents are classified leveraging Support Vector Machines using two labels: "hard", if related to politics or news, or "soft", if they are related to more relaxed content, like entertainment. Then, they created a measure to classify the contents as liberal, neutral, or conservative, and then constructed and analyzed the content network for each user. The result is that the contents shared in the inner circle of Facebook friends are more involved than the Facebook algorithm in terms of the content that a user decides to consume on SNSs.

The issue with content-based approaches is that the data about the content need to be annotated and there is no standard procedure to intervene in an unsupervised context; therefore, different authors leverage different methods. Among the most common, Natural Language Processing algorithms are used to extract the sentiments or ideologies of users or to detect their stances.

Similarly, as concern specifically echo chamber detection, the approaches can be divided into three categories according to the granularity level to which the analysis is linked.

- Macro-scale: the interactions are observed on an aggregate level, with the aim of identifying well-distinguished clusters of users;
- Meso-scale: allows to identify echo chambers through the observation of com-

munities of nodes sharing the same leaning.

- Micro-scale: here, it is considered the single-user behaviour without considering aggregation of users.

Studies on the macro-scale level, include [MMS21], where it is leveraged network science 2.3 to study the political interaction network of Donald Trump and Hilary Clinton supporters on Reddit. Leveragin Reddit structure, they tag each user with their leaning (democrat or republican) and then they analyze their interactions via the joint probability of observe an interaction between nodes of opposite stance, given the leaning of the node u.

A micro-scale approach can be found instead in [CMG+21] where the authors leverage homophily to assess whether or not there are echo chambers in the network. More in detail, it is proposed to infer the individual leaning of each user toward a specific topic through the analysis of the content produced and then to calulate the average leaning of the content the user produced. After obtaining this information, they go on to study the network topology of user interactions, moved by the idea that users surrounded by people with a similar are consequently exposed to similar content(s). To do so, they define for each user i, the average leaning of its neighbourhood, then these value are plotted into contour maps, respectively on the x and the y axes. The brighter areas in the plot, indicates a large density of users with the same leaning, that in their case study is visible in particular in Facebook and Twitter discussions about vaccines and abortion.

Regarding the meso-scale approach, in [MPR21] the authors set up a standard methodology independent of specific features of an SNS and applied it to a case study. This approach was chosen for the detection of polarized systems in this thesis.

The pipeline presented (see also Fig. 2.1) is divided into four steps and allows the identification of echo chambers at a meso-scale level.

1. Identification of a controversial issue, since polarization is more prone to appear when users talk about trigger topics;

2. Inference of users' ideology on the post. The problem is modeled as a *text*

classification task;

3. Creation of the users' debate network, since they will interact with like-minded users, insulating themselves from rebuttal;

4. Identification of users' clusters that are *homogeneous*, both from a topological and ideological perspective. This problem is addressed using Community Detection algorithms.

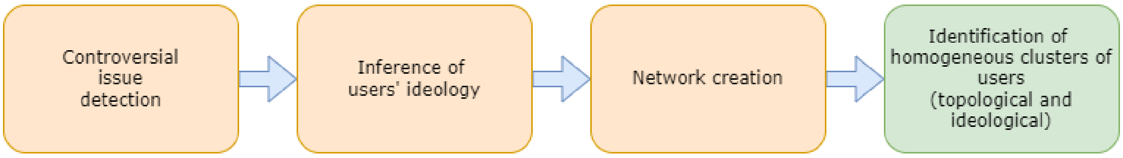

Controversial issue detection
Inference of users' ideology
Network creation
Identification of homogeneous clusters of users (topological and ideological)

Figure 2.1: Pipeline of the method proposed in [MPR21]

Another example of meso-scale approach is described in [VPV21]. Here, the authors study the echo chamber phenomenon using tweets about COVID-19. They proceed through the construction of interaction network, enriching the representation by assigning weights that also consider the *sentiment* or a tweet's topic. Then a community discovery algorithm (METIS) is applied and the last phase is to calculate the polarization and use controversy measures to confirm the presence of echo chambers.

Unfortunately, there is a gap concerning the diachronic analysis of echo chambers in the literature. Indeed, when considered, the social graphs they analyze are usually defined as static entities, representing timeless pictures of the observed phenomenon. Regrettably, such flattened representations, keeping together interactions potentially distant in time and disregarding their temporal ordering, risk an overestimating of users' sociality and failing to understand the real dynamics that led to their formation.

In addition, there is a lack of studies that focus on the linguistic production of users inside echo chambers. In such a way, it would be possible to identify common linguistic patterns across echo chambers members that would help in their detection and in a more precise definition of this phenomenon.

One example of these works is proposed by Brugnoli *et al.* [BCZ$^+$19] which focuses on describing the linguistic behaviour of users inside an echo chamber, showing that for users inside exists a variation in their vocabulary to the point that converge between those ones interacting inside and outside an echo chamber.

2.3 Network Science and Social Network Analysis

Network science is a discipline that intersects graph theory, physics and computer science and it is focused on studying *complex systems*[2]. Although network science has its first theorizations in 1930, its true development is relatively new, as evidenced by the exponential increase in publications at the dawn of XXI century. According to Albert-László Barabási [BP16], two major findings occured: the first is the emergence of tools to map networks with the aid of the Internet, which introduces newer and faster ways to share and store data; the second is that it does not matter which is the origin of the network (e.g., social, biological, etc.): networks share many common properties.

Indeed, society itself can be considered a complex system made of intertwined people, tied by many different kind of relations. This topic is the prerogative of *Social Network Analysis* (SNA), a subfield of network science crossed with sociology, psychology and other behavioural sciences, that investigate social relations and interactions.

In the following we will give an overview of network science, introducing definitions and concepts that are necessary to understand the case study exposed in this work.

2.3.1 Network Science fundamentals

Network Science models reality by the means of graphs, mathematical structure made of single components called nodes or vertices, with related pairs tied through a link called *edge*. A graph $\mathcal{G}$ may be formalized as follows: $\mathcal{G} = (V, E)$, where E is the entire set of edges and V indicates the comprehensive set of vertices.

[2] Complex systems are settings where the interactions of individual entities gives rise to new collective behaviours that are not derivable from the study/analysis of a single individual's behaviour.

Often, the concept of *graph* is wrongly used to talk about *networks*. Whereas networks refer to the 'real object', graphs instead refer to the mathematical modelling of networks.

When discussing graphs, relations are expressed by links, which may be of two types: *undirected* or *directed* (*digraph*). In the first case, the relationship modelled is a mutual one, such as a friend on Facebook, a romantic relationship and so on. Instead, an example of a directed graph is the web, where website A may point to website B but website B does not necessarily point to A.

Another important property of a graph is the so-called degree k, which represents the number of links a node has to other nodes, for instance, in Facebook friendship network k_v denotes the number of friends node v has. The definition above can be applied to undirected networks, whereas if the relations are not mutual, it is necessary to distinguish between *in-degree*, k_i^{in}, that is the number of incoming links, and *out-degree*, k_i^{out}, the number of connections that start from a node v.

Each link (i, j) may also carry a weight $w_{i,j}$. In a network of phone calls, the weight may be the total number of minutes two people mutually talk; in a social network, the number of interactions they had through time (i.e., number of comments or replies to a post, number of shares, likes, etc.).

In addition to the degree, it may be also calculated another related measure, which is the *average degree* $\langle k \rangle = L/N$ in an undirected network. If the network is directed, instead, the average degree is equal to $\langle k \rangle = 2E/V$.

2.3.2 Node-attributed networks

In real-world networks, the entities can be described by attributes; for example, a network of actors can be composed of nodes accompanied by information about the age, the main genre in which they perform, or a network of webpages can be enriched by the topic or another classification label. This type of information may be encoded through attributes leveraging a particular model of so-called *feature-rich networks*[3] [IAG$^+$19]. Information can be stored on the nodes (i.e., user's political

[3] According to [IAG$^+$19], *feature-rich networks* are *all the complex network models that expose one or more features in addition to the network topology.*

leaning, age, and so on), or on the edges, such as the *weight* of a relation or *temporal information.*

Focusing on the node-attributed graphs, according to [ZCY09], a node-attributed graph is defined as $\mathcal{G} = (V, E, \Lambda)$, where V is the set of vertices, E is the set of edges, and Λ is the set of m attributes associated with vertex $v \in V$ to represent the attributes.

The added value to traditional graphs is the presence of semantic information represented by attributes that, when related to network topology, may shed light on the individuals' behavioral patterns. Such a modeling choice allows to estimate additional semantic-related properties of the networks such as homophily [MSLC01] i.e., the tendency of individuals to relate to others with whom they have more common characteristics on different levels (e.g., political leaning, age, gender). This tendency has been widely studied by sociologists and network scientists, for example, to gain insights about gender and interaction patterns in primary schools [SCP$^+$13] and to study segregation or integration [Moo01].

From a network science point of view, a traditional way to estimate homophily is through *Newman's assortativity* coefficient [New03], which allows classifying the behaviour of the nodes in a network as *disassortative* – if there is an inverse correlation w.r.t a property – or *assortative* – if the nodes tend to share the very same property. Such a measure has the main disadvantage to return just an averaged number that flattens a more complex behaviour, e.g. the presence of different mixing patterns across the network or the identification of anomalous patterns.An alternative is to quantify mixing pattern at different levels, as proposed in [PDL18].

2.3.3 Temporal Networks

Real-world phenomena are dynamic, and as a consequence, most networks develop over time: from relationships between users on SNSs to protein interaction, every kind of tie may form or disappear – or simply change – over time. This opens another problem for Network Science, which is the need to find a way to represent topological changes in a network as well as to build data structures capable of manipulating temporal networks [HS13]. One problem may be to represent stable and unstable ties, the first being relations – long and short-term ones – and the second one interactions – instant ones like an email, or over a timespan like phone calls. Among the most important representations, we can distinguish between *graph series*, *stream graphs*, and *interval graphs* with the possibility of performing data transformations from one type to another.

The application of temporal networks has been attested in Community Detection, Diffusion processes, and Link prediction.

A simple representation may be performed using snapshots, where the temporal evolution of the network is observed through the lens of the n static snapshot of the graph within a specific time interval. The timespan between a snapshot and another plays a critical role; if the timespan is too large, there might be an information loss in terms of node and link variation that has not been captured. Likewise, if the timespan is too short, there may be only a few changes in the time window that do not show a temporal correlation with the process occurring in the network [CBWG11]. The criterion suggested by [SEEDT20] is to maintain a balance between the target to be studied and the resolution. As for *pros* and *cons*, the strength of snapshots is parallelized to reduce computation times but at the cost of a *reconciliation* phase of the fragmented results that must be linked between subsequent snapshots [RC18a], which often results in information loss and additional computational complexity. Furthermore, snapshot graphs allow the application of algorithms and measures in single partitions available for static graphs, in addition to avoiding the instability of temporal networks [MFF$^+$17].

Stream graphs [LVM18] are commonly used to model interactions going through time, and are formally described as follows: $\mathcal{S} = (T, V, W, E)$ where T is a measurable set of time instants, V is a finite set of nodes, $W \subseteq T \times V$ is a set of temporal nodes and a set of links $E \in T \times V \otimes V$, such that $(t, uv) \in E$ implies $(t, u) \in W$ and $(t, u) \in W$. $\mathcal{S}$ is defined as a *link stream* if all the nodes are always present in all the timestamps described as $L = (T, V, E)$ where $W = T \times V$ implicitly. The formalism of Stream Graphs has been also adapted in [LVM18] to directed, weighted and bipartite (stream)graphs. According to the authors, Stream Graphs may be useful for modeling traffic networks, mobility traces, and financial transactions.

Interval graphs [HS13] are temporal networks characterized by edges active over a set of temporal intervals. They prove to be useful in proximity networks, i.e., networks where the ties represent a contact between a node i and a node j, or infrastructural systems.

2.3.4 Community Discovery

Static Community Discovery

Community Discovery (or community detection) is the subfield of Social Network Analysis that deals with the identification of topological structures, i.e., communities, inside a network. The concept of a *community* is intuitively simple to define, but currently, there are many definitions, each one of them focusing on different aspects but never converging into a single one. This made the formalization of community an *ill-posed problem.*

In an informal way, it can be stated that communities are clusters of nodes that are strongly similar to each other such that, an entity in the community is closer to the other members of the community than to the other components that belong to other communities.

The chance to have a glimpse into the community structure of networks would aid in different fields, such as biology, that is, to study the interaction network of proteins, sociology, and computer science, where the detection of communities may be useful for categorizing webpages sharing the same topic. Community discovery could give benefits also to research on network visualization, especially when the graph to visualize is large and with a high number of edges, thus leading to an overflow that makes the whole graph unreadable.

Given the instability of the definitions on which community detection rests, the community discovery itself is an arduous task: as stated in [FH16], it is difficult to assess the goodness of the partition obtained using a CD algorithm, due to the lack of a clear and universal definition on what a community is. In [CGP11] the authors conducted a review of both the definitions and algorithms approach. As it regards the definition, the authors propose four different classes of definitions, based on four different ideas, as follows

- **Density-based definitions**: where a community is perceived as a group in which there are many edges between vertices, but between groups there are fewer edges;
- **Vertex similarity-based definitions**: communities are groups of edges similar – using a property to measure – to each other;
- **Action-based definitions**: a community is made of users that perform the same action;
- **Influence Propagation-based definitions**: definitions that have their roots in the concept of tribe as users following the influence of the leader.

In the same work, the authors defined a taxonomy of CD algorithms as follows:

- **Feature distance**: this approach encompasses all CD algorithms that are based on the assumption that a community is composed of entities that share a set of features that may assume similar values[4]
- **Internal density**: communities, in this case, are a set of entities in denser[5] areas of the network. This approaches are based on the *modularity*, described in 2.3.4;
- **Bridge Detection**: a community is seen as a collection of entities close to each other (i.e., a few links separate nodes in the same community). It follows that entities in different communities are distant from each other: communities are dense subgraphs interconnected by a small number of links that, if removed, divide the graph into partitions[6], the desired communities;
- **Diffusion**: in this case, a community is described as a collection of entities close to each other (i.e., few links separate nodes in the same community). It follows that entities in different communities are distant from each other; the communities are clusters of nodes that are influenced by the diffusion of a property inside the network;

[4] In this case, often an edge or node-attribute is used as a feature;

[5] w.r.t. a random graph with the same degree distribution

[6] This approach leverage the concept of *bridges* [Gra83]

- **Closeness**: a community is defined as a collection of components mutually closed, in a way that few links separate them, while the entities outside are far. It follows that the entities in different communities are distant from each other;
- **Structure**: leveraging the idea of a precise structure of edge, a community is defined as a precise and immutable structure;
- **Link Clustering**: this kind of approach is based on clustering the edges of the network and then identifying as communities the resulting groups.

Evaluation

The evaluation of the goodness of a partition identified by a Community Detection algorithm is affected by the ill-posedness of the community definition. As it is not well defined *what* a community really is, it is difficult also to find a unique quality function to evaluate the partitions.

The goodness of a partition is conducted through two types of evaluation: the first one is the *internal evaluation*, a category that includes metrics useful for evaluating the goodness of the partitions *per se*, while the second one, namely *external evaluation*, allows the evaluation of the partitions obtained with a ground-truth or other CD algorithms.

For the scope of this study, it is mandatory to introduce the three metrics used in Section 3.1.2.

- **Modularity**: this quality score allows to measure the observed number of edges inside the given partition minus the expected number of edges if they where distributed following a null model of a random graph. Modularity takes values in the range [-1, 1] and is formalized as follows:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{uw} - \frac{k_v k_w}{(2m)} \right] \delta(c_v, c_w)$$

- **Purity**: the measure, as defined in [CR20], is calculated as the product of the frequencies of the most frequent labels carried by its node. The function lies within the range [0,1].

$$P_c = \prod_{a \in A} \frac{\max\left(\sum_{v \in c} a(v)\right)}{|c|}$$

- **Conductance**: the conductance for a community C is the volume of edges pointing out of it. The aim is to minimize the value of this function such that the average value across all the communities is as low as possible.

$$Conductance_c = \frac{2\,|E_{OC}|}{2\,|E_C| + |E_{OC}|}$$

where E_{OC} is the number of edges exiting the community, E_C is the number of edges remaining inside.

Labeled Community Discovery

Labeled Community Discovery (or Attributed Community Discovery) consists of the community discovery task enhanced by taking into account the attributes of the nodes, therefore, the aim is not just to maximize the topological distance among communities but must also take into account node labels, so that nodes inside a community have attributes that are as homogeneous as possible.

Dynamic Community Discovery

If the Community Discovery task imposes arduous challenges, its temporal counterpart adds a layer of complications. For example, a node in a network or a link may disappear, thus leading to topological variation in the communities. In addition, communities are subject to *events* [PBV07], such as *birth*, if nodes form a brand new community, or *merge* if two communities merge into a single one. All events incurred by a community allow the description of its lifecycle [CR19]. Another peculiar issue in this task is the identification of *smooth* partitions over time and the *identity* of the communities through their temporal evolution, described through the paradox of the *ship of Theseus*.

The detection of communities in dynamic networks is conducted using different approaches, according to [RC18a], and is categorized as follows:

- Instant optimal: the communities are detected at each timestamp t using a CD algorithm; then, the partitions at t are matched looking back to the partitions identified at *t-1*. This approach ensures the quality of the partitions at each evolution step.
- Cross time: this kind of approaches consider all the changes through time at once, resulting in partitions smoothed and coherent over time.
- Temporal trade-off: the communities are detected at every timestamp, keeping an eye also on the past topology/partitions and finding a *trade-off* between time t and the past. This approach is the best one for data that evolve rapidly, but they are prone to avalanche effects, so communities may be subject to drifts w.r.t. static partitions.

2.4 Natural Language Processing

Natural Language Processing (NLP) refers to that interdisciplinary field of research, which intersects linguistics, computer science and artificial intelligence with the aim of developing algorithms to decode and understand natural language so as to obtain results that can help both in linguistic analysis at different levels and in human-machine interaction.

Another definition, often misdefined as NLP, is *text-mining.* While the focus in NLP is on the *representation* of meaning, through the means of concepts borrowed from linguistics, such as *part-of-speech* or *dependency relations*, the focus of text mining is on the extraction of (useful) knowledge from text [KP07].

In most cases, to work on textual data, as will be done in Section 4.5, it is necessary to apply a pipeline to clean data to enhance their quality and obtain better results after the application of a machine learning algorithm. Usually, the first step of text preprocessing is 1) *normalization*, the lowercasing of all the text; the result acts as input for 2) *punctuation removal* and 3) *stopword removal*, to strip words without semantic meaning, such as determiners, conjunctions, or prepositions. The next step is 4) *tokenization*, the task of splitting the entire text into small fragments, the so-called tokens, which roughly correspond to words. The last step is 5) *lemmatization* in such a way that the words are stripped of inflectional endings and obtain the base form of a word, the *lemma.*

For the sake of this work, it is useful to introduce two other tasks typical of *text mining*, namely *sentiment*, *emotion analysis*, and *topic modeling.*

2.4.1 Sentiment and Emotion analysis

Sentiment analysis, also often referred to as opinion mining, is the discipline that, in the field of Natural Language Processing, deals with studying and analysing the opinions, feelings and attitudes expressed by people in written form, in relation to other individuals, products, events or, in general, topics and issues. Studies in this field began in the early 2000s. Although the first article mentioning sentiment analysis dates back to 2003 [NY03], there are, however, the first studies in this field dating back to the 1990s [HM97, WBO99] that focus on the extraction of adjectives that express certain emotions. The years in which this discipline developed, however, were crucial: they were in fact the same years in which the World Wide Web became widespread and the first social networks were born, which can be considered a perfect form of digitally born data rich in opinions, the number of which has grown drastically in recent years. Through social networks, individuals can easily connect with others, get to know other people, who perhaps share the same ideas, and encourage them to share their own ideas, thus adding data that can potentially be used to study the opinions of a large group of people.

A more precise task is the *Emotion analysis*, where algorithms are designed to extract the emotional nuances implied in the text. There are several approaches to emotion analysis, such as *lexicon-based*, if lexical features are used to detect an emotion, or *machine learning* approaches [HMKM17]. In the second case, the emotion(s) to identify is/are the labels of a *classification problem*, where the aim is to find a label (the emotion) that better describes the data in the input (text in our case). These labels are often taken from Ekman's basic emotion model [Ekm05] (*anger, surprise, disgust, enjoyment, fear*, and *sadness*) or from Plutchik's model, which extends Ekman's model with *trust* and *anticipation.* A way to extract this kind of information from text, is via Zero-Shot Learning [LEB08] [CRRS08] (henceforth, ZSL) is a problem setup in machine learning, applied especially in the field of NLP for text classification and in Computer Vision. ZSL is a kind of *jack of all trades* since it allows to perform classification tasks providing a description of the classes to identify. It is usually necessary for a classification model to be trained on a dataset with class labels, which often poses various challenges, since this requires an automatic, semi-automatic or manual annotation phase.

The idea behind ZSL, is that the labels to assign during the test phase contain an inner meaning that can be exploited to better assign the most appropriate label to the text in input. In [CRRS08], is shown that this *data-less classification* achieves very good results, being quite similar in the performances to supervised Machine Learning models.

The meanings behind the labels can be extracted using different approaches, which in [ZLG19] are split into three categories. The first approach is to exploit semantic features (i.e., attributes or properties) of the label, as in [ZSH$^+$19], and another approach is to employ ontologies and knowledge graphs [SGMN13][7], while the last one is to make use of word embeddings to capture latent relationships among words. A hybrid approach was proposed by [ZLG19].

ZSL was initially used to perform topic characterization [YdMdC$^+$19] [ZLG19]; however, it has been applied in many NLP fields, including *sentiment analysis.*

2.4.2 Topic Modeling

The topic modeling task consists of mining reliable information to aid in the identification of the topic discussed in the text given in the input. In addition to text mining, it is also used in the field of bioinformatics [LTD$^+$16] to find similar genomic patterns, as well as to study environmental data [GGD13] and text data extracted from social platforms.

One of the most used algorithms is Latent Dirichlet Allocation (LDA) [BNJ03], which leverages the *bag of words* model: nonetheless, this model has severe limita-

[7] Note that the work is referred to the *Computer Vision* field

tions because the text is seen as a series of words without considering grammar and, in general, the context in which the word is placed.

Other approaches include the use of *neural networks* (*Neural Topic Models*, NTM) [PL20][CLL+15], NMF (Non-negative Matrix Factorization) [FI10] that still relies on bag of words model, hybrid variants of LDA and *neural networks*[WZF12].

An alternative approach is represented by *Transformers*, a particular neural network architecture that implements the *self-attention* mechanism. Transformers [VSP+17] were introduced in 2017, and revolutionised many fields, from computer vision to NLP, having as its best-known models BERT (Bidirectional Encoder Representations from Transformers) [DCLT18] and GPT-3 [BMR+20] just to mention the most famous. Starting from BERT, many other models have been developed, such as RoBERTa [LOG+19] or AlBERT [LCG+19] achieving very good results and being state-of-the-art in the NLP field. These kinds of models aim to create *word embeddings*, representations of every token as a vector keeping track of the context in which every word is in, overcoming the problems of older, simpler approaches.

BERT can also be leveraged to perform topic modeling using BERTopic [Gro22] a topic model that has shown interesting results in recent years (up to the current moment) [EY22] when applied to short review texts [SFRM21].

It operates according to the following three steps:

1. The input text is converted into the respective embedding representation using the Sentence-BERT framework [RG19], but it is possible also to implement other frameworks since they are necessary just to extract the embeddings.

2. The documents are then clustered, starting from the assumption that documents with similar topics should have a similar embedding representation. Before performing clustering, BERTopic applies UMAP [MHSG18] to avoid the curse of dimensionality in a high-dimensional space. The applied clustering is of a density-based type, HDBSCAN [MHA17] because it models noise as outliers and obtains purer topic representations.

3. The final step is the *topic representation* via a modified version of TF-IDF – namely c-TF-IDF – adapted to estimate, for each cluster, the importance of the

words in it. This step is repeated such that the last common topic is merged with the one that is most similar until reaching the number of topics specified by the user.

Chapter 3

Methodology

In this third chapter, we provide an overview of the methodology employed in this thesis to carry out the analysis shown in Chapter 4. All the steps described in this chapter are aimed at understanding, both in terms of network and text features, how communities and users inside them behave as time passes. In detail, we attempt to answer the following questions:

- Are echo chambers diachronically stable w.r.t. the users inside?
- Do echo chambers keep or lose their strong polarization over time?
- Is there a particular feature that distinguishes the linguistic productions of users in the echo chambers?

Firstly, to answer all the questions, and in particular the first two, it was necessary to build the interaction network keeping also into account its temporal evolution. This phase was followed by the extraction of the communities and their evaluation, so as to distinguish the communities more at risk of being echo chambers from the communities with a lower risk. The echo chambers are then monitored over time to assess whether users inside them are "trapped" or if there is a continuous renewal because users constantly come in and out of them.

Then, the answer to the last question was sought using different NLP approaches, including emotion analysis and topic modeling. In this case, we observed the macro differences between echo chambers and lower-risk communities; thereafter, we focused on the communities with a higher risk of polarization. In the following are described the methologies used to deal with each of these tasks.

3.1 Echo chamber and polarized system detection

To identify the *chamber* component of the echo chamber, that is, the virtual place where a polarized discussion reverberates, it is necessary to address network science to gain insights into the topology of relationships and the cohesion of ideologies within them.

More in detail, we followed the last two steps of the pipeline described in [MPR21] (see Section 2.1.1), which involve the creation of the debate network. In this case, we modeled the interaction network as a node-attributed undirected graph, 2.3.2. Each node of this graph is a Reddit user described by the political leaning (see Section 4.1).

After defining the graph, we can proceed with the macro-, meso-, and micro-scale detection of polarized systems. The process described in [MPR21] belongs to the second class of approaches; however, in this study, we investigated polarized systems by exploiting all these three levels of analysis.

3.1.1 Macro-scale

With regard to the macro-scale approach, it allows a qualitative estimation of polarized systems. This was conducted by visualizing the time-flattened network using Gephi[1]. The nodes of the graph were coloured according to their leaning (see 4.1). The Force Atlas 2 [JVHB14] layout was used to arrange the network because this layout is particularly suitable for graphs with a high number of nodes and edges.

The static network used above to conduct the macro-scale analysis was then divided into five smaller graphs. These were used to visualize the network again, snapshot per snapshot. In this way we obtained a better overview of the qualitative changes in the polarization of the network.

[1] Gephi website, https://gephi.org/, last visited: 27/10/2022

3.1.2 Meso-scale

The macro-scale analysis was deepened by observing the meso-scale topology of the network. In this way, the focus shifted from the whole network to communities that may gather like-minded users who may be at risk of being in an echo chamber.

Foremost, it was necessary to construct a temporal network. To do so, the five smaller graphs obtained in Section 3.1 were used to obtain a dynamic representation of the network via snapshot graphs.

In order to perform the analysis, we leveraged two software packages in Python, namely `NetworkX`[2] and `CDlib`[3]. Through NetworkX it was possible to construct, for each semester, the node-attributed graph. In addition, we removed the nodes without a label and the ones representing moderators or bots, as described in Section 4.1.

Then, we performed the Labeled Community Detection on the graph, as described in [MPR21]. Given the attributed nature of this network, it is necessary to consider the cohesion of partitions in terms of both topology and node attributes. The algorithm chosen to address this specific task was EVA [CR20], described in detail in Section 3.1.2. The partitions obtained for each semester were then aggregated in a `TemporalClustering` object, necessary to compute the similarity of the partitions between different timestamps, and then evaluated, as described in Section 3.1.2.

EVA: Louvain Extended to Vertex Attributes

The EVA algorithm [CR20], is a bottom-up low-complexity algorithm, that allows the identification of meso-scale structure through simultaneously optimizing both structural cohesion and label homogeneity. As described in [CR20], the algorithm extends the traditional Louvain [BGLL08] algorithm to node-attributed graphs. It was specifically designed to optimize two quality functions, namely Newman's modularity and *Purity*: this allows to keep as high as possible the modularity of the partition as in the original version of the algorithm, but, at the same time, to maximize the homophily.

[2] NetworkX documentation, https://networkx.org/, last visited: 27/10/2022

[3] CDlib documentation, https://cdlib.readthedocs.io/en/latest/, last visited: 27/10/2022

The algorithm was implemented through `CDlib`, setting the *alpha* parameter[4], which indicates the importance of modularity and purity criteria to 0.5.

Partition Evaluation for Echo chamber detection

After performing the Labeled Community Detection task, the next step was to apply two different constraints in two measures which are typically used for evaluating the goodness of the partitions in Community Detection (Section 2.3.4).

Polarized systems, and, in particular, echo chambers need a sort of closed space in which a discussion can reverberate moving from one member to another. Therefore, the idea suggested in [MPR21], is to evaluate the partitions in terms of *Conductance* (Section 2.3.4), led by the hypothesis that most of the edges should remain inside the partition; this result is then intertwined with another measure, *Purity* (Section 2.3.4), which has the function of estimating the goodness of the partitions in terms of attribute homogeneity. This last measure is necessary because the users inside an echo chamber or polarized system tend to share the same ideology.

According to these two measures, the risk for a community to be an echo chamber is maximized when *Conductance* is equal to 0 and *Purity* is equal to 1.

Following the case study in [MPR21], we considered echo chambers (or communities more at risk of being so), the communities having a *Conductance* equal or lesser to 0.5 and the *Purity* equal or greater to 0.7.

Before proceeding with the next phases of the analysis we also set another threshold in terms of communities' size: in this way, we kept only the communities having at least 15 users inside. This was necessary to remove noisy communities that could lead to scattered and/or not fully representative results.

Polarized system stability

After assessing the presence of echo chambers, the next step involved the observation of the way echo chambers develop over time under two different aspects: the first facet is focused more on users, so as to obtain information on their behaviour

[4] The *alpha* parameter lies within the range [0,1]; a value nearer to 0 will optimise modularity, otherwise the clustering criterion will be maximised

in order to understand whether or not they remain sealed in ECs over time; the second facet, complementary to the first, consists in assessing if a community, that in a timestamp *t* has *Purity* and *Conductance* values that can be associated to an *echo chamber*, remains frozen in that *status* also at *t+1* and so on.

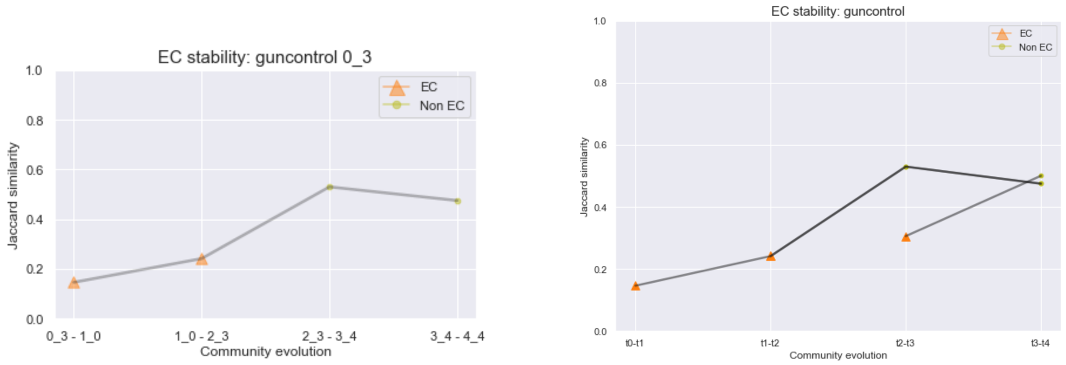
The first aspect has been analysed leveraging the Jaccard index, often used to estimate the *stability* of the partitions in temporal networks. The Jaccard index is defined as follows:

$$J(A_t, B_{t+1}) = \frac{|A_t \cap B_{t+1}|}{|A_t \cup B_{t+1}|}$$

In order to prevent biases due to users joining discussion threads by posting once that may alter the results, it was necessary to perform a quick preprocessing step on the identified communities. For each temporal snapshot, we computed which nodes were in common between adjacent timestamps. Then, we proceeded by removing from each community extracted in the Labeled Community Detection step, the nodes not included in this intersection.

After this preprocessing step, for each pair of timestamps, we estimated their similarity by calculating the ratio between the number of common users in two adjacent timestamps and the number obtained by summing the users forming their union. This allowed assessing whether echo chambers are stable over time or if they make room for variations in their users' composition.

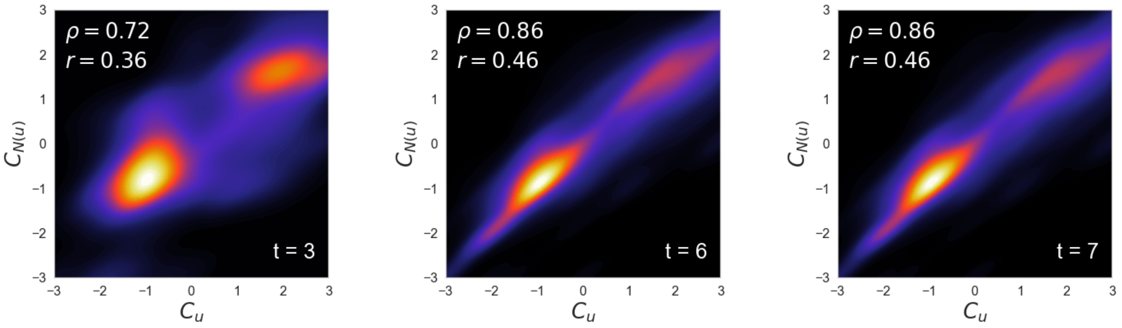
Regarding the second aim of this task, in order to assess whether an echo chamber remains so over time, we observed, for each community at risk during timestamp *t*, if the partition at the next timestamp was still an echo chamber or not. This was analyzed through line plots, where the line corresponds to a community and the marker represents the *status* of the community, i.e., a triangle represents its being an echo chamber while a dot its being not a risk of being an echo chamber. The plot in Figure 3.1a was used to assess the temporal evolution and stability of a single community, while in Figure 3.1b it is shown an example of plot illustrating the same information for all extracted communities.



(a) Example of single community temporal evolution and stability

(b) Example of community temporal evolution and stability

Figure 3.1: Stability plot



(a) t0

(b) t1

(c) t2

Figure 3.2: Example of *contour plot*. Source: [BCC+22]

3.1.3 Micro-scale

The micro-scale approach shifts the focus of the analysis from communities to individual nodes, in order to assess whether the neighbourhood of the node taken into account is composed of other people/nodes carrying the same idea/attribute, as it should be in an echo chamber.

To do so, it was defined for every node u the average neighbours' opinion and then it was calculated the correlation between a user u and its nearest neighbours. The correlation was then visualized through a *contour plot* (Figure 3.2), where the lighter areas correspond to areas of high density of like-minded users. In the case of strongly polarized systems, the contour plot shows denser areas on the bottom left and on the top right.

3.2 Natural language analysis

In this section, we describe in detail the implementation of the linguistic analysis pipeline, divided into several steps (see Figure 3.3), including the extraction of *text-specific features*, the analysis of sentiment and the emotion emerging from the chosen text data, as described in Section 4.1. Finally, we extracted the topic that characterized each discussion thread, in order to assess whether polarized systems have one or more distinguishing features that make them different from non-polarized ones.

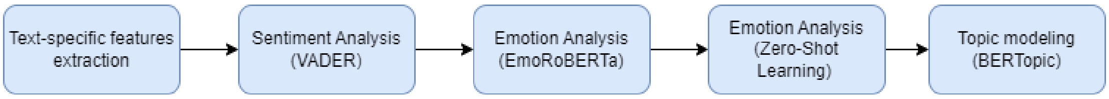

Text-specific features extraction
Sentiment Analysis (VADER)
Emotion Analysis (EmoRoBERTa)
Emotion Analysis (Zero-Shot Learning)
Topic modeling (BERTopic)

Figure 3.3: Linguistic analysis summary

3.2.1 Text-specific features extraction

For this task, the text was analyzed by leveraging two well-known Python libraries in the NLP field. The first is NLTK[5] (*Natural Language Toolkit*), which was used to extract a small range of linguistic features. The second one, instead, is SpaCy[6], leveraged to perform the *Named-Entity Recognition*, that is, the extraction of the so-called *named-entities*, a name, location, organization, etc., which is represented by a *proper name.*

Table 3.1 shows a summary of the extracted features. In this case, we also decided to include two measures of text informativity, that is, *Type Token Ratio* (TTR) and

Feature	**Description**
n_sent	Number of sentences of the post/comment
n_tok	Number of token of the post/comment
avg_sent_length	Average number of sentences per topic
avg_word_length	Average length of a word per topic
ttr	Total number of types divided by the total number of tokens
lexical_density	Number of lexical words (or content words) divided by total number of words. Lexical density is also a measure of informativity of a text
n_ner	Numer of Named Entities extracted from each text.

Table 3.1: Summary of the linguistic features extracted for the analysis

[5] NLTK website: https://www.nltk.org/
[6] SpaCy website: https://spacy.io/

Lexical Density.

The TTR is described as the ratio between the total number of types and the total number of tokens and it is a basic measure that allows estimating the lexical variety in a text. The lexical density, instead, is calculated as the ratio of lexical items, in this case, the sum of nouns, verbs, adverbs and adjectives, to the total number of words. Lexical density is useful to get a more accurate view of the complexity of a text.

Finally, we extracted the named entities (NE) to gain insights into the different distributions of places, people, and so on. This information may be useful to characterize the different topics analysed and, hopefully, communities with a higher risk of being echo chambers from the ones with a lower risk. NE distribution was visualized through a normalized stacked bar chart, so as to observe possible variations in their distribution.

3.2.2 Sentiment analysis

The extraction of the sentiment that characterizes posts and comments was exploited using VADER[7] (*Valence Aware Dictionary and sEntiment Reasoner*) [HG14], a lexicon and rule-based model for sentiment analysis, fine-tuned to extract the sentiment hidden in texts produced on social media.

Through VADER it is possible to obtain four different scores: *positive*, *negative*, *neutral*, and *compound*. While the first three measures indicate the ratios of sections of the text associated with each category, the compound score, instead, is a global measure of the sentiment of a text. This last score was used to extract the dominant sentiment from the input. It can assume values that lie within the range [-1, +1] and can be exploited – setting various thresholds – to extract a unique label that summarizes the general inclination of the input text.

The thresholds set are the following, as described in [HG14]:

- Positive, if $score >= 0.05$;
- Neutral, if $score > -0.05$ and $score < 0.05$;

[7] Vader on Github, https://github.com/cjhutto/vaderSentiment, last visited: 26/10/2022

- Negative, if $score <= -0.05$.

The results were then analyzed using bar charts to observe the variation in the distribution of the three labels among different topics and different kinds of polarized communities.

3.2.3 Emotion analysis

The *emotion analysis* task, introduced in Section 2.4.1, was conducted to gather additional insights into the way people interact inside polarized and non-polarized systems.
The task was exploited by leveraging two different approaches: a fine-tuned model from HuggingFace, namely, *EmoRoBERTa* and *Zero-Shot Learning*.

EmoRoBERTa

EmoRoBERTA[8] is a version of *RoBERTa* [LOG+19] fine-tuned on the *GoEmotions* dataset [DMAK+20]. The choice of this model was justified by the fact that the *GoEmotions* dataset contained 58.009 records taken from the same data source chosen for the analysis, Reddit (Section 4.1.1), in the time window between 2005 and January 2019. In addition, the data were manually annotated with 28 different emotions, with a distribution of 12 positive, 11 negatives, and 4 ambiguous emotions plus the "jolly" of the neutral label.
EmoRoBERTa was implemented using the `transformers` library from HuggingFace [9] using the `sentiment-analysis` pipeline [10] and passing as input model for the pipeline the EmoRoBERTa model.

Zero-Shot Learning

The Zero-Shot Learning approach for emotion detection was used as an alternative approach to EmoRoBERTa. Because EmoRoBERTa extracts a high number

[8] EmoRoBERTa model from HuggingFace, https://huggingface.co/arpanghoshal/EmoRoBERTa, last visited 26/10/2022

[9] `transformers` Github repository https://github.com/huggingface/transformers, last visited 27/10/2022

[10] Sentiment Analysis pipeline https://huggingface.co/docs/transformers/main_classes/pipelines, last visited: 28/10/2022

of labels that may result in a loss in terms of its accuracy, due to the high level of specificity of the labels, it was decided to perform emotion analysis using a smaller subset of emotions. The choice for these labels fell on Ekman's wheel of emotions (Section 2.4.1), but, extended with the labels *approval, disapproval* and *neutral*, in conformity with EmoRoBERTa's *neutral* label. The first two labels were also available in the EmoRoBERTa set of labels, and they were added because it can be possible that a high number of users in an echo chamber approve what has been said by another user inside.

This classification approach was implemented leveraging the specific pipeline[11] offered in the `transformers` library from HuggingFace. This time, we chose to use as model `bart-large-mnli` [LLG$^+$20] to perform the inference task from the input text.

3.2.4 Topic modeling

The topic modeling task was included in this analysis to find out whether there is actually a particularly controversial topic or topics that reverberate within an echo chamber.

The implementation was conducted via the `BERTopic`[12] Python library, which allowed both topic extraction and the visualization.

In fact, the strength of BERTopic – in addition to being extremely accurate – is that it offers different kinds of visualizations of the topics extracted. In this way, it was possible to extract the topics generally discussed in all the polarized systems we identified and then go deeper into the analysis and provide an overview of the most controversial topics that characterize every single community.

For example, in Figure 3.4, it is possible to observe eight different topics, each one described by five different words, namely the most representative words that define that general topic.

[11] HuggingFace available pipeline, https://huggingface.co/docs/transformers/main_classes/pipelines, last visited: 27/10/2022

[12] BERTopic GitHub repository, https://github.com/MaartenGr/BERTopic, last visited: 27/10/2022

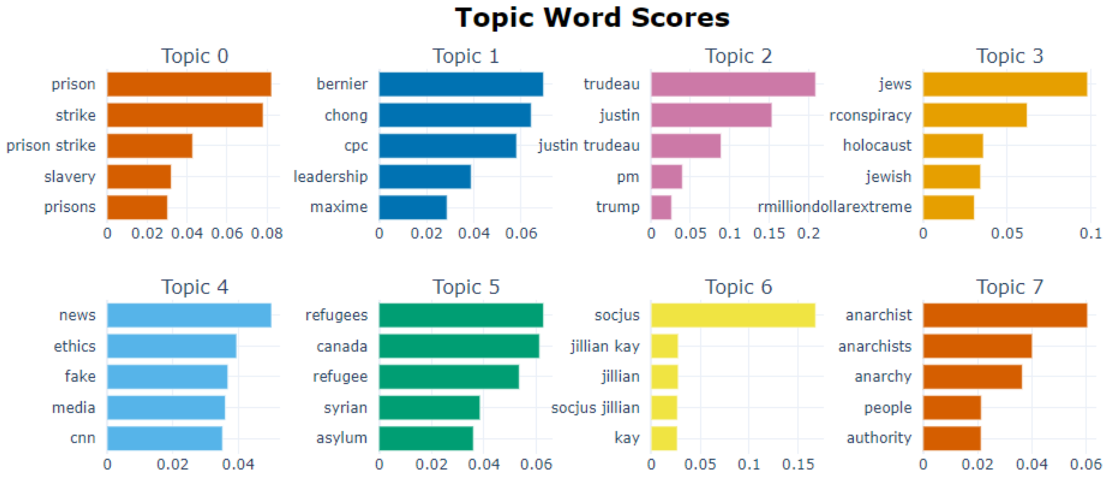


Figure 3.4: Example of barplot of the topics extracted using BERTopic

Chapter 4

Case study: results

After defining a methodology to study the evolution of echo chambers through time and different approaches to analysing users' linguistic production (see Chapter 3), this framework was then applied to a specific case study focused on the study of echo chambers arising on Reddit from the debate about three socio-political topics. The chapter is structured as follows: initially, we discuss the choice of the topics of investigation and the platform from which the data were extracted (Section 4.1); then, the analysis moves to the dataset exploration, giving insights both on the amount of data and on the temporal network necessary to apply the framework. Then, in Sections 4.2, 4.3 and 4.4 are presented the results regarding the echo chamber detection on various levels. Finally, in Section 4.5 the results obtained via linguistic analysis are discussed.

4.1 Data

For the sake of this thesis, we rely on the datasets provided in [MPR21] which cover the first two and a half years of Donald Trump's presidency, from January 2017 to July 2019. In Section 4.1.1 we discuss the choice of the topic as well as the analyzed online social platform (i.e., Reddit); then, in Section 4.1.2 are presented more in detail the datasets used to perform the case study.

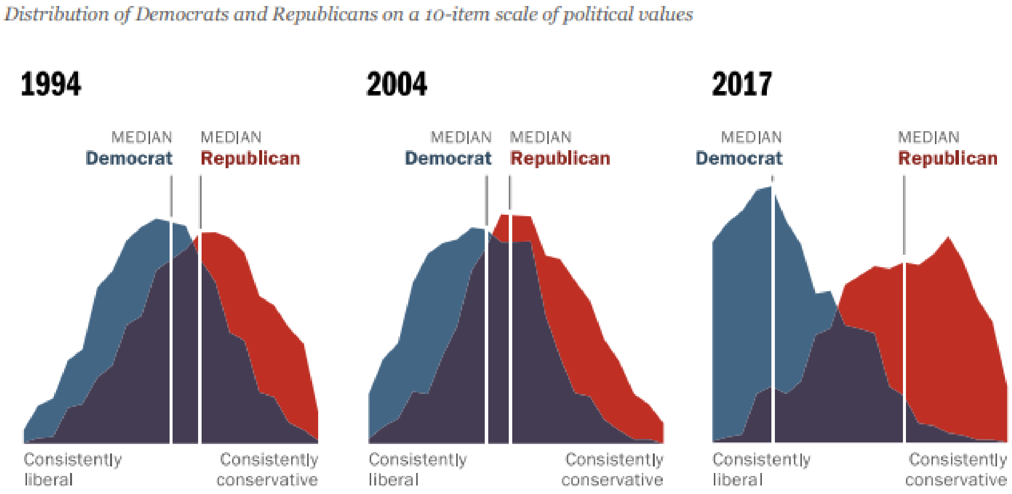


Figure 4.1: Variation of polarization until 2017. Source: Pew Research Center

4.1.1 Choice of the topic

The main topic of analysis is tied to American politics, that, given its bipartite structure, fits well with the concept of *polarized system*. Indeed, a different range of literature has addressed the hypothesis of polarization in the United States, even before the diffusion of SNSs. According to the Pew Research Center[1], the two parties, Democrats and Republicans, are more ideologically distant today than they have ever been before. This tendency can also be observed in the ideological distribution of Americans (see Figure 4.1), where there is a clear increase in the tail of the distribution, both on the liberal and conservative sides.
The focus on Donald Trump's presidency can be justified by looking at Figure 4.1, where it is visible that the heavy polarization towards the extremes was reached in 2017, with his election as President of the United States.

Reddit as data source

Reddit is currently the sixteenth most used social media website in the world[2], following various instant messaging platforms, such as *Whatsapp, Telegram* or *WeChat.*

[1] Partisan divides over political values widen, Pew Research Center https://www.pewresearch.org/politics/2017/10/05/1-partisan-divides-over-political-values-widen/, last visited:24/10/2022

[2] Global social networks ranked by number of users on Statista, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, last visited: 27/10/2022

It was launched in 2005 by Steve Huffman and Alexis Kerry Ohanian, and can be considered an atypical platform, since users do not have a real profile, nor there is a system of followers as found on other social sites, such as Twitter or Facebook. Reddit is described[3] as a source for what is new and popular on the Web. It can be roughly described as a social news and entertainment platform where posts, pictures, and videos circulate.

Reddit's environment is composed of so-called *subreddits*, sort of communities devoted to just one topic specified in the name of the subreddit and preceded by the *r/* prefix. The specificity of the topics covered within a subreddit may vary, so there are both general subreddits such as *r/worldnews, r/Economics, r/gaming, r/Art*, and more specific ones, related, for example, to a specific game or a specific niche related to a hobby (such as *r/StardewValley*, a subreddit for the homonymous game, or *r/FantasyArt*, a subreddit for fantasy-related illustrations).

In subreddits, the subscribed users can post as long as they are compliant with the rules of that space. Their observation is guaranteed by *moderators*[4]. Regarding the contents the users can publish, these are heterogeneous and range from texts to pictures and videos; they can be commented on by other users, with the possibility to reply to a specific comment. The simplest way to interact with a post, however, is via *upvotes* and *downvotes*, which express the opinions of users towards a certain content, allowing it to obtain better exposure in the case of high number of upvotes. Furthermore, each community may define its specific *Flairs*, labels that define the general topic of a post, that appears to the left of the title. For example, in the subreddit r/Steam[5], some of the Flairs are *Discussion, Question, News.*

Reddit users are not even obliged to reveal their real name, it is also discouraged by Reddit itself during registration. This is linked to the fact that there is no obligation to enter any other personal information (i.e., gender, birthday, and provenance), making Reddit an SNSs where the anonymity of users is guaranteed. Even the per-

[3] Reddit FAQ, https://www.reddit.com/wiki/faq/#wiki_what_is_reddit.3F, last visited: 25/10/2022

[4] Redditors that contribute to removing posts, ban spammers, or people that break subreddit rules for free. Source: https://reddit.zendesk.com/hc/en-us/articles/204533859-What-s-a-moderator-

[5] Steam is a well-known video game distribution platform, https://www.reddit.com/r/Steam/, last visited: 25/10/2022

sonal profile is not "personal", since it just shows the comments, the thread(s) started by the user and their *karma*, the reputation of the user inside the platform, gained via upvotes. There is also the absence of a pervasive system of followers/followees as it happens, for example, on an SNS like Facebook; the interactions may happen via comments or chat.

Two are the key features that make Reddit a good data source: first, users' anonymity, which can push a user to speak without filters on every topic, even those ones that are more controversial. The second one is Reddit's structure: the data are already divided by topic, leveraging the internal division into subreddits. In addition, it is also possible to find data about niches or particular populations [ABKK19], for which researchers have always had problems finding participants.

Moreover, the topic of choice we discussed in the previous section (see Section 4.1.1), fits well with Reddit, since almost half of the active accounts write from the United States (47.13%) [6] making Reddit the sixth most used social in U.S.[7].

4.1.2 Dataset

The Reddit dataset, as described in [MPR21], is made of three different macro-categories of socio-political issues, as follows:

- **Gun Control**: contains six subreddits linked to *gun control*, the laws and politics supporting the making, selling and ownership of firearms. The discussions in this dataset pertain to the legalization of firearms and their consequences. The subreddits included are the following: *r/guncontrol, r/antiwar, r/liberalgunowners, r/Firearms, r/guns*

- **Minority Discrimination**: this topic is related to groups that are considered minorities and may therefore be subject to discrimination. The subreddits under this umbrella are variegated and defend or are discriminatory towards

[6] Reddit global active user distribution: https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/, last visited: 24/10/2022

[7] Market share of the most popular social media websites in the US. https://www.statista.com/statistics/265773/market-share-of-the-most-popular-social-media-websites-in-the-us/, last visited: 25/10/2022

gender, racial, sexual equality.

The subreddits included are the following: *r/MensRights, r/KotakuInAction, r/metacanada, r/racism, r/AgainstHateSubReddits, r/Anarchism*

- **Political Sphere**: the data contained are tied to American politics and its ideologies.

  The subreddits included are the following: *r/esist, r/democrats, r/MarchAgainstTrump, r/Conservative, r/Libertarian, r/Republican*

All these categories are composed by subreddits representing both sides of a controversy (i.e., r/democrats and r/Republican), in order to have a better view of the debate. The data, posts and comments, from discussion threads in these subreddits, were scraped using the *Pushshift API* and refer to the period between January 2017 and July 2019.

In order to proceed with the meso-scale ECs' detection pipeline (see Section 2.2), the datasets were enriched with the leaning of each user inferred from their posts, namely *protrump*, *antitrump* or *neutral*. To perform this annotation, the authors leveraged $BERT_{BASE}$ trained and tested on a ground truth of polarized posts, i.e., datasets extracted from three subreddits[8] known to be openly Anti-Trump or Pro-Trump.

The model was then applied to the Reddit sociopolitical dataset and allowed to extract a score – corresponding to the model confidence – ranging from 0 to 1. If a post is tagged with 1, then it is perfectly aligned with Pro-Trump ideologies, while, on the contrary, a 0 as label means that is aligned with Anti-Trump ones. From this score, it was extracted a *leaning score* L_u, computed as the average value of their posts' leaning

$$L_u = \frac{\sum_{i=1}^{n} PredictionScore(p_i)}{n}$$

where $p_i \in P_i$ corresponds to a post shared by a user u and $n = |P_i|$ is the cardinality of the set of posts from this user. The obtained value was discretised into intervals, as follows: *antitrump* if $L_u \leq 0.3$, *protrump* if $L_u \geq 0.7$, *neutral* otherwise. These arbitrary thresholds have also been maintained in this work.

[8] The subreddits are: *r/The_Donald*, *r/Fuckthealtright*, *r/EnoughTrumpSpam*

Afterwards, the authors created the interaction network for each topic: starting from the dataset, were obtained five smaller subsets, each representing a semester, and then the network was reconstructed in a way that each labeled user u has a link to user v, if and only if u directly replied to a post or a comment of user v or vice versa. Each edge u, v is enriched with the weight of that relation, equal to the number of times that interaction between u, v happened.

The datasets containing the *edgelists* of interactions for each topic, were used in Section 4.3 to perform the creation of the network and then the extraction of ECs. In this phase, nodes representing moderators or bots were removed, in order to avoid biases due to non-representative interactions. The linguistic analysis conducted in Section 4.5, instead, was exploited by leveraging the textual data from the posts and comments datasets extracted from the API.

4.2 Macro-scale analysis

The macro-scale level of analysis allows, as described in Section 3 to perform a qualitative estimation of the presence of polarized systems.

For each topic, we plotted the time-flattened network. The nodes were colored according to the leaning of each user, i.e., magenta for Trump supporters and blue for the anti-Trump partition. The neutral users, were instead colored with grey.

The most visible division is, without a doubt in the *Minority* graph in Figure 4.2, where it can be seen a clearer division between two distinct centres, one characterized by a higher density of pro-Trump nodes and the other one slightly more polarized on the opposite side of the discussion – even if merged with other Trump supporters.

In the *Politics* graph, differently from what happens in *Minority*, there is a dense connection in terms of edges between the two sides of the controversy. *Gun Control* seems to have been monopolised by the *pro-Trump* faction, with an almost imperceptible cluster of well-separated democrats. We also decided to visualize, for each topic, the network timestamp per timestamp. This made it possible to better observe how the network evolved over time and to note a progressive strengthening

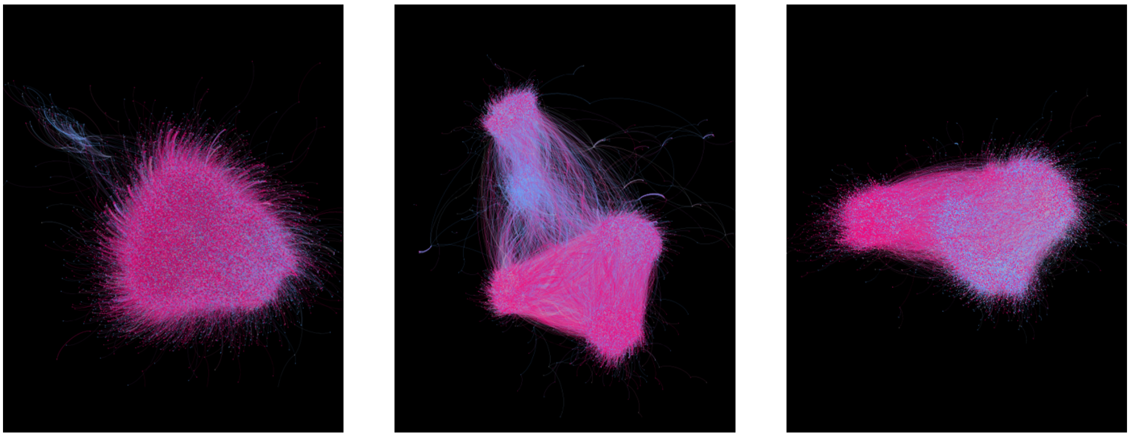


Figure 4.2: Time-flattened network (*Gun Control*, *Minority*, *Politics*)

of the pro-Trump polarization – characterised by three different pro-Trump poles – and a smaller anti-Trump pole.

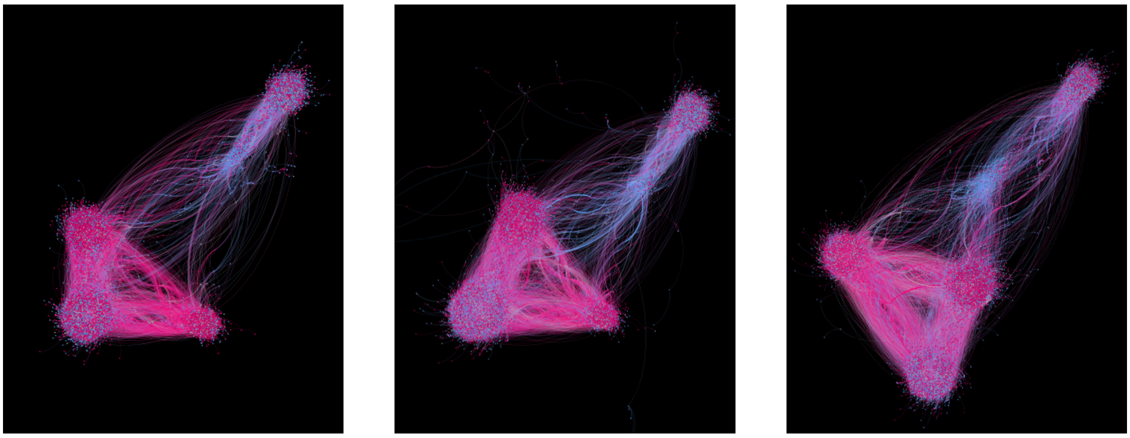


Figure 4.3: *Minority* network evolution *t0, t1, t3*

4.3 Meso-scale analysis

After identifying the presence of echo chambers through the visualization, for each topic, of the flattened networks, the analysis moved to the meso-scale level, where the presence of polarized systems was studied more in detail, by investigating the topological structure of the network and the ideological homogeneity. The section is divided into two parts: in Section 4.3.1, are presented the results obtained through the Labeled Community Discovery task, while in Section 4.3.2, the focus will shift onto the temporal analysis of echo chambers.

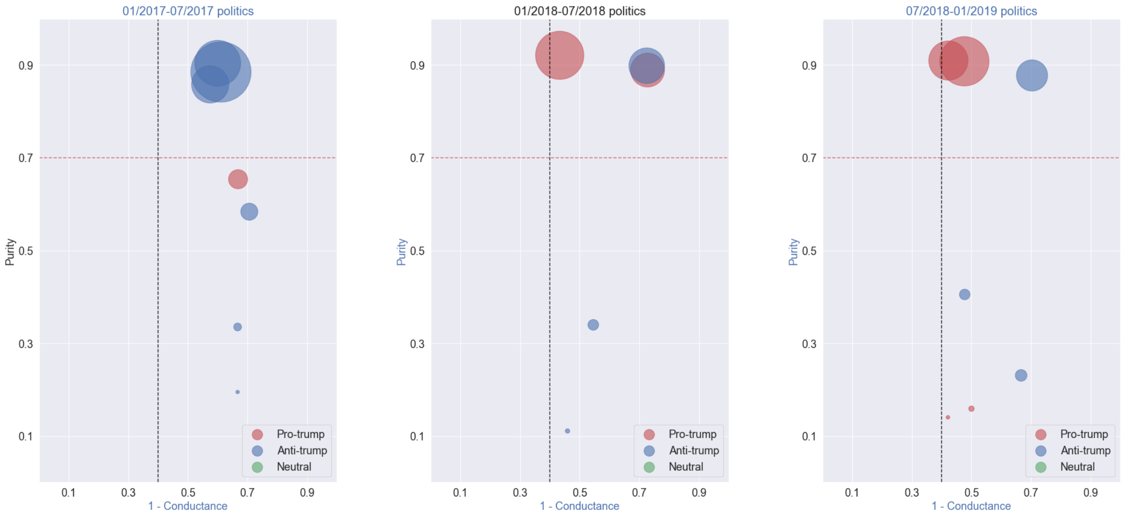


Figure 4.4: *Politics* communities and risk of being ECs

4.3.1 Echo chamber detection

The detection of echo chamber was conducted via the approach described in Section 3.1.2. After the creation of the dynamic network using the five datasets (Section 4.1.2), it was carried out the Labeled Community Discovery step, allowing to obtain for each temporal snapshot belonging to each topic, the most likely partitions. From the extracted communities, the smaller ones were then removed to avoid not representative results. Subsequently, the remaining partitions were evaluated via *Purity* and *Conductance* (see Section 2.3.4), thus allowing to discriminate between potential ECs and Not-ECs.

Politics. Through the scatter plots in Figure 4.4, it is possible to identify heterogeneous communities, in the first timestamp in Figure 4.4a, all the bigger communities are ECs that belong to the Democrat side of the controversy. By lowering the *Purity* threshold, we would also include two smaller communities, including one with a Republican majority. The tendency varies in the subsequent timestamps (Figures 4.4b, 4.4c), where can be found, in addition to other anti-Trumps ECs, even a few *pro-Trump*.

Minority. Minority resulted to be the dataset with the higher number of ECs, both *anti-* and *pro-Trump*, all above the *Purity* threshold that marks strong ECs.

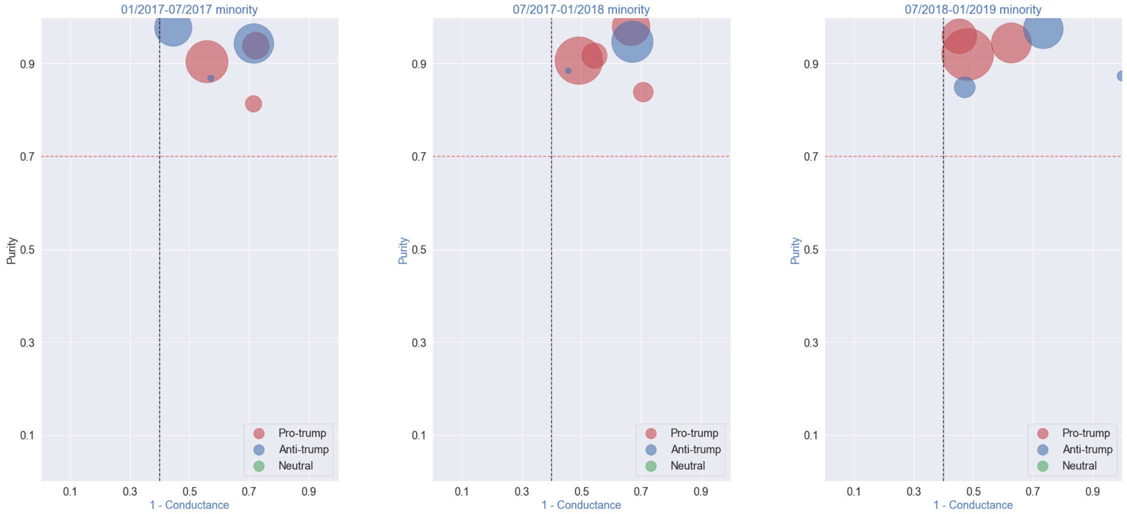


Figure 4.5: *Minority* communities and risk of being ECs

Through time, it is possible to observe that there is an increase in *pro-Trump* ECs over democrat ones.

Gun control. In contrast to what happens in the two other datasets, in *Gun Control* there is not a strong polarization of communities, even though there is a significant number of communities just below the *Purity* threshold. The majority, in this case, consists of *pro-Trump* communities.

4.3.2 Assessing echo chamber stability

The extracted partitions were further investigated, as described in Section 3.1.2, to study whether ECs maintained their strong polarization over the five semesters under analysis or not. For each of the topics, we cleaned the network from the nodes that interacted just a few times in one semester as described in 3.1.2. Then, for each topic and semester, it was computed and visualized the evolution of each community and the similarity in users' composition through the Jaccard index.

Politics. From the very first results obtained on the *Politics* network, it can be seen that echo chambers are persistent over time and often keep their users inside. From Figure 4.7a, it can be seen that from the very first pair of adjacent timestamps, ECs

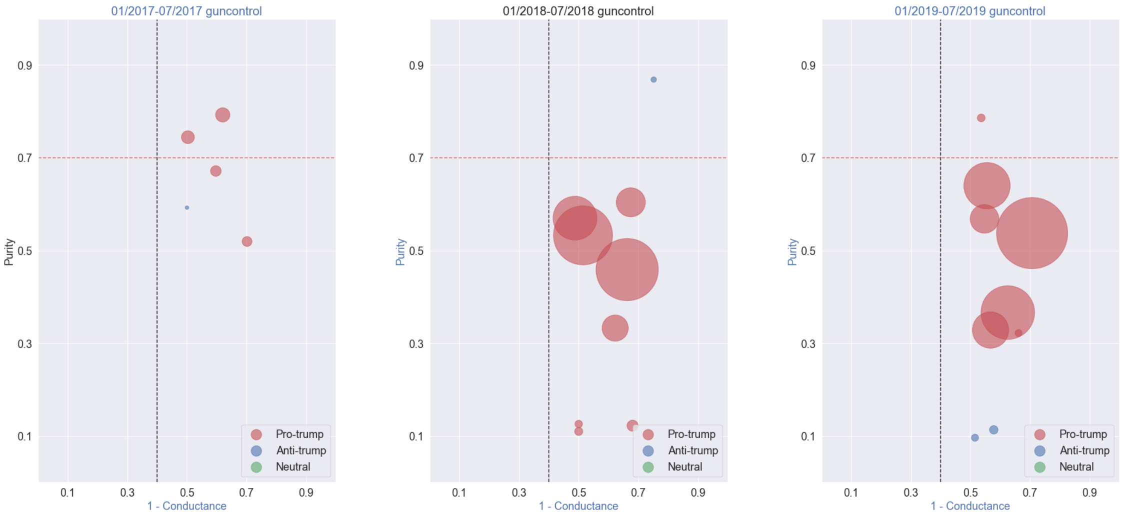
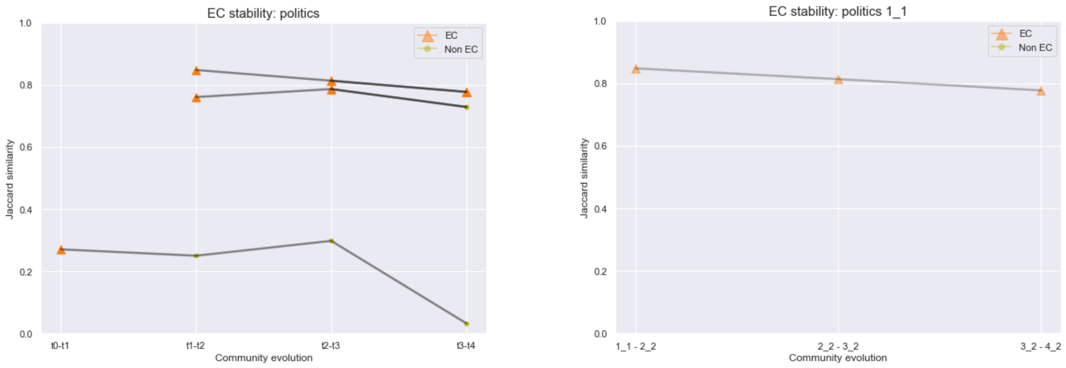


Figure 4.6: *Gun Control* communities and risk of being ECs



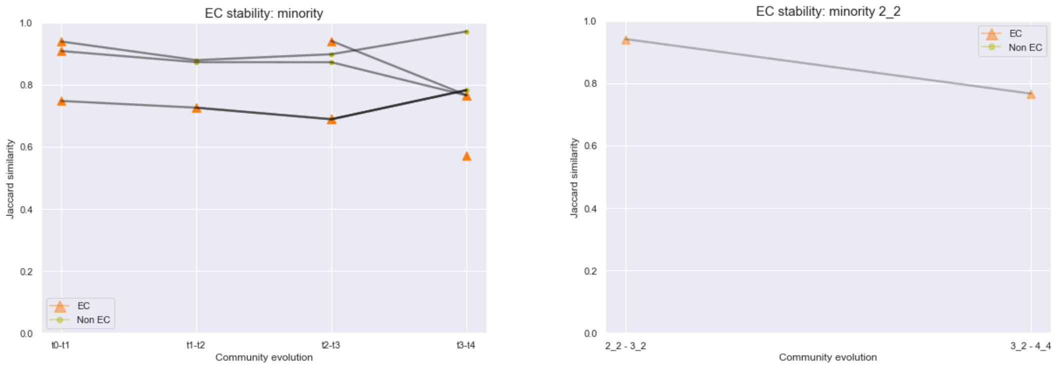
(a) ECs evolution stability (all)

(b) Evolution of highest risk EC

Figure 4.7: Stability of *Politics* ECs

have a high value of Jaccard index (≈ 0.8) in four different communities. Furthermore, two of them maintain high stability even between another pair of semesters. The community more at risk of being an EC, is also the most persistent one: *community 1_1* in fact maintains its being an EC for three adjacent timestamps, starting from the second timestamp until the very end, and it keeps most of its initial members.

Minority. In *Minority*, as in *Politics* it is visible again the stability of ECs: also in this case there is an echo chamber that lasts for a year and a half, *community 0_4*. This community, though, it is not the one more at risk, since that place is taken by



(a) ECs evolution stability (all)

(b) Evolution of highest risk EC

Figure 4.8: Stability of *Minority* ECs

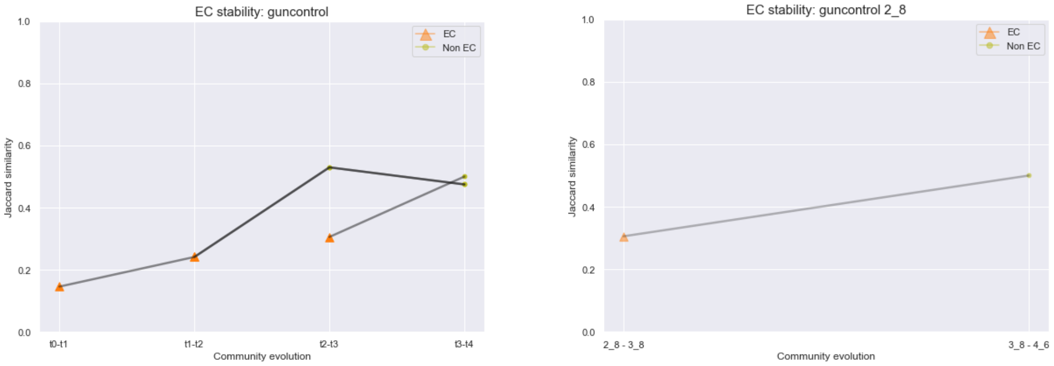
community 2_ 2, which lasts for one year. Even in this stability plot, it is clearly visible that the similarity values between timestamps are very high, although there are – as in *Politics* – stability values that are slightly lower than the ones registered in the previous pair of timestamps.

Gun control *Gun Control*, unlike *Minority* and *Politics*, has only one EC that lasts for two adjacent temporal snapshots. Even by looking at the community with higher likelihood of being an EC (see Figure 4.9b), it is just possible to find out that its lifecycle as EC lasts just between the second and the third timestamp, then it becomes a community with lower risk, increasing its stability. This pattern is repeated also in the other communities, and this result can be considered anomalous *w.r.t.* the ones obtained in other graphs. However, it must be recalled that *Gun Control* ECs' were very small and often on the verge of the EC's risk threshold (Figure 4.6), so the results may be reasonable if we think to their instability.

4.4 Micro-scale analysis

For each of the three topics, the analysis was ultimately deepened by comparing the users' opinions with the ones of their neighbours, semester by semester. To visualize the polarization, for each node u we plotted the correlation between the user's opinion and the average opinion of their nearest neighbours $C_N(u)$.

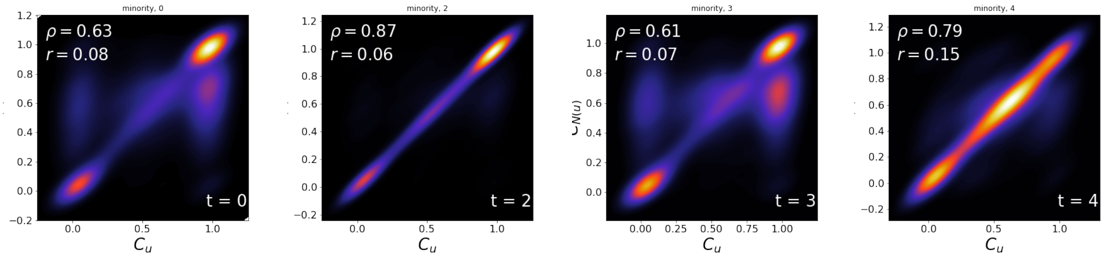
Not all the plots were included, but only the most relevant ones for each dataset.



(a) ECs evolution stability (all)

(b) Evolution of highest risk EC

Figure 4.9: Stability of *Gun Control* ECs



(a) First semester

(b) Third semester

(c) Fourth semester

(d) Fifth semester

Figure 4.10: *Minority* contour plot

Minority. Minority, according to the meso-scale analysis, was the dataset with the highest concentration of potential echo chambers. This can be confirmed also by the micro-scale analysis shown in Figure 4.10. In the first semester, it can be already identified a clear lighter area towards the extremes on the bottom left and, even clearer on the top right of the contour map. In between semesters, there is a reinforcement of the polarization of opinion, especially on the pro-Trump side, which culminates in the third semester 4.10b, where the lightest area is located right at the pro-Trump pole, having a Pearson coefficient $\rho \approx 0.87$. Instead, when looking at network's assortativity, r, we can see that the network is slightly assortative, but the assortativity value is quite low in every temporal snapshot. It results higher in the last six-month observation period, but despite the fact that both ρ and r are higher in comparison to the other observations, there is still a clear loss of the strong polarization seen in the previous semesters, due to a high number of *neutral* users.

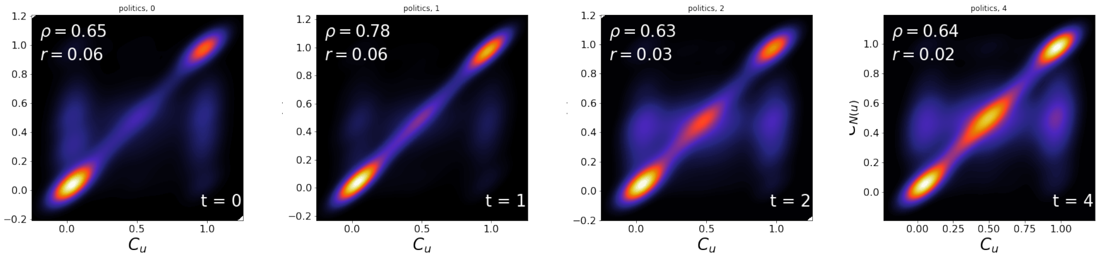


Figure 4.11: *Politics* contour plot

Politics. In the *Politics* network, opposite to what happens in the *Minority* one, we witness a more evident polarization towards the Democrats' side, which remains stable across the first two snapshots (Figures 4.11a, 4.11b). In Figure 4.11b, the Republican polarization becomes also more evident. In the fifth snapshot (Figure 4.11d), becomes clearer the presence of a neutral side of the controversy, which already appeared in the third semester but here becomes more evident. Even here, Pearson's coefficient ρ is around 0.65, reaching its highest value ($\rho \approx 0.78$) in the second semester (Figure 4.11b). Despite the high ρ value, there is again no match in terms of assortativity, which does not exceed 0.06.

Gun Control. In *Gun control* there is a very strong polarization towards the pro-Trump side. This confirms what has been discovered with the macro- and meso-scale analysis, that is the almost exclusive presence of republican echo chambers.

In the first network's snapshot, the users are already arranged along the top-right of the diagonal, even if there is a slightly lighter area around the neutral section of the diagonal. The Democrat side of the controversy described in the dataset, instead, does not seem to be polarized, given the feeble lighter area on the bottom left. In *Gun Control* we register also the lowest value of assortativity, equal to 0 in the second semester (Figure 4.12b) so, apparently, in this case, it does not exist a correlation between the opinion of the user's neighbourhood and their leaning.

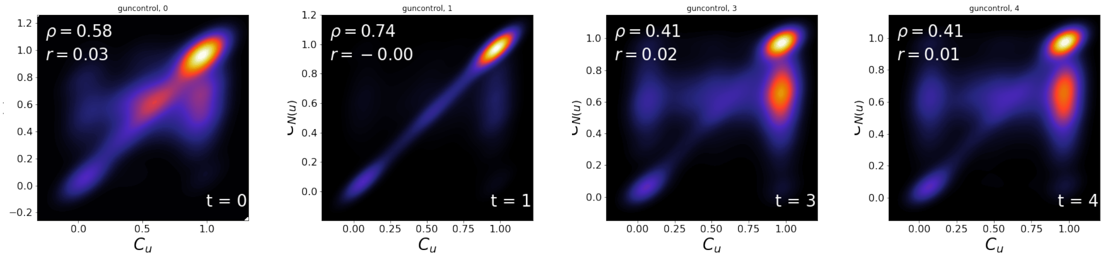


Figure 4.12: *Gun control* contour plot

	Dataset					
	Politics EC	**Politics Non EC**	**Minority EC**	**Minority Non EC**	**Gun Control EC**	**Gun Control Non EC**
avg_n_sent	1.43	1.74	2.67	4.14	3.22	2.06
avg_n_tok	19.06	26.75	52.87	88.89	47.76	31.20
avg_sent_length	12.20	12.40	15.06	16.67	10.79	12.39
avg_word_length	4.94	5.00	4.85	4.77	4.13	4.88
ttr	0.99	0.99	0.99	0.99	0.98	0.993554
lexical_density	0.709737	0.70	0.66	0.67	0.66	0.73
n_NER	1.05	1.19	0.88	1.90	0.96	0.95

Table 4.1: Summary of the averaged values of linguistic features per each post dataset

4.5 Text analysis

This section will give an overview of the results of the linguistic analysis of the linguistic productions of Reddit users. The aim of these results is to assess if there is/are discriminant feature(s) in the way people interact via posts or comments on Reddit. At first, we will proceed with the extraction of text-specific features (Section 4.5.1), then the analysis is deepened via sentiment and emotion analysis (Sections 4.5.2, 4.5.3) and will end up with the extraction of the most relevant topics (Section 4.5.4) from the posts and the comments, so as to understand whether there is a definite polarized topic around which the echo chambers focus. In the following sections, the results are discussed according to the type of analysis performed, topic by topic.

4.5.1 Text-specific features

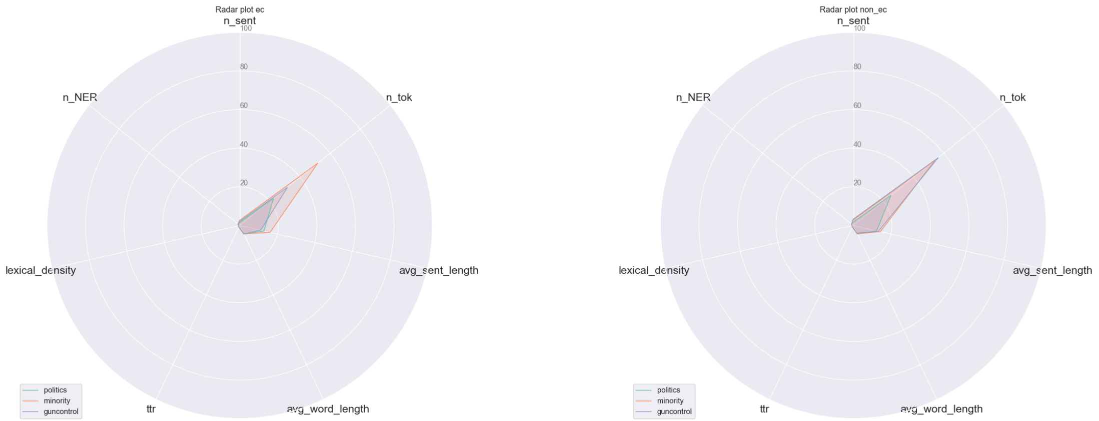
This analysis, as described in Section 3.2.1, allowed to extract, for each topic, a small range of text-specific features, namely the number of sentences and tokens, the average length of sentences and words, the Type-Token Ratio, the lexical density and the number of Named-Entities. Generally speaking, the results mapped within the radar plots in Figure 4.13 show that there is not too much difference in terms of language – at least with the features extracted in this study. Most of the simpler features describing ECs and Not-ECs do not seem to follow a pattern that allows to fully distinguish between ECs and Not ECs. As for Named Entities, instead, in Table 4.2, we reported the five most common NEs for echo and non-echo chambers. Given the diversity in terms of the amount of data in one dataset and the other, we decided also to plot a normalized stacked bar chart (Figure 4.14) to visualize whether there is a variation or not in the proportions of the NE extracted using SpaCy.

Politics. In *Politics*, there are not too many major differences between communities considered to be echo chambers and those less at risk. The only exception is a slightly higher number of sentences and tokens per sentence. As for NEs, in the Not-ECs it may be noted there is a higher number of `DATE` w.r.t. the ECs.

Minority. A macro-difference emerged in the *Minority* dataset, is that users within ECs would be more prone to write short texts than users outside them ($avg\_n\_tok \approx 88.8$ vs. ≈ 52.86). In addition to this, their texts are richer in term of NEs.

From Figure 4.14 it can be also observed that in the Not-ECs there is a lower proportion of words labeled as `CARDINAL` in the most polarized contents, while the others are almost equal in percentage.

Gun control. The results obtained using the text data in this dataset showcase a higher number of tokens per sentence in ECs. In addition to this, in *Gun Control* there is the highest difference between ECs and Not-ECs in terms of lexical density (0.66 vs. 0.72 respectively), so the posts outside ECs, at least in this specific case, tend to be more informative. Looking at the NEs distribution, instead, it seems not to be any relevant difference among ECs and Not-ECs.



(a) EC

(b) Not EC

Figure 4.13: Text-specific features radar plot

Dataset					
Politics EC	**Politics Not EC**	**Minority EC**	**Minority Not EC**	**Gun Control EC**	**Gun Control Not EC**
PERSON (26606)	PERSON (39101)	PERSON (8801)	PERSON (31333)	DATE (3909)	DATE (12094)
ORG (21786)	GPE (33024)	GPE (7855)	GPE (26054)	GPE (3295)	GPE (11384)
GPE (20231)	ORG (30493)	NORP (6823)	DATE (21643)	PERSON (2752)	PERSON (9553)
NORP (20060)	NORP (30045)	ORG (5608)	ORG (18699)	ORDINAL (2162)	CARDINAL (6630)
DATE (9322)	DATE (14408)	DATE (5001)	NORP (17512)	CARDINAL (2096)	ORG (6343)

Table 4.2: First five most common NE in the post dataset

4.5.2 Sentiment analysis

The sentiment analysis was conducted via VADER (see Section 3.2.2) and applied both on the post and comment dataset. The results, were similar between posts and comments, as we will discuss in the sections below. Both ECs and Not-ECs posts, in fact, are characterized by a predominance of posts tagged as *negative*, followed by an almost equivalent number of *positive* and *negative* posts. With regard to the comments, on the contrary, the number of positive comments tends to be closer to the number of negative comments. This may be due to the sparseness of the comments data, which are shorter than comments and this can make challenging to identify the correct sentiment.

Politics. In *Politics* (see Figure 4.15), we observe the general pattern described in the introduction to this section – the predominance of negative texts over positive and neutral – for ECs posts, while in Not-ECs there is a similar number of

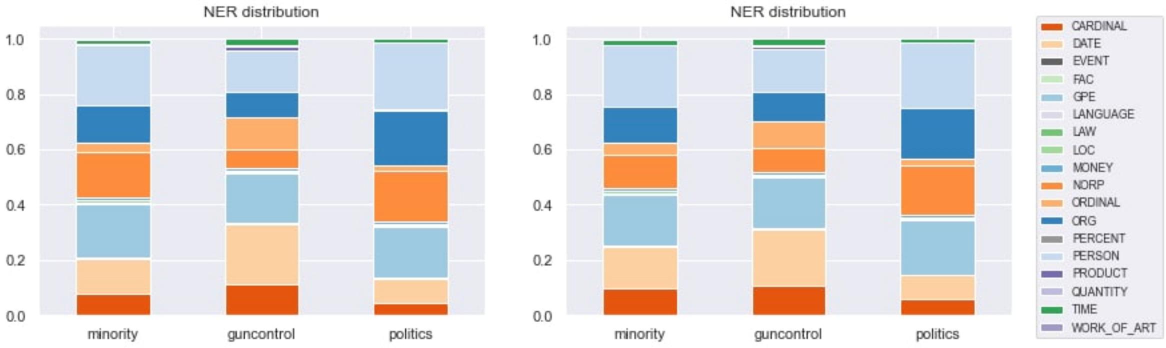
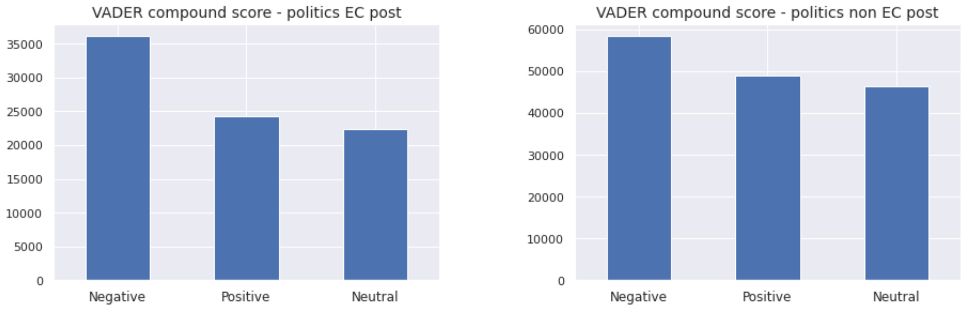


Figure 4.14: Named Entity distribution across echo chambers and non-echo chambers posts



(a) ECs

(b) Not-ECs

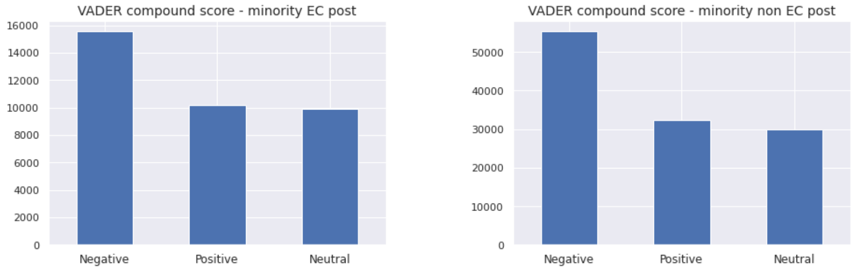
Figure 4.15: *Politics* sentiment results

neutral and positive posts, but with higher values than in ECs.

From the comments, instead, almost the same information was extracted, which is the equivalent high value of negative and positive comments.

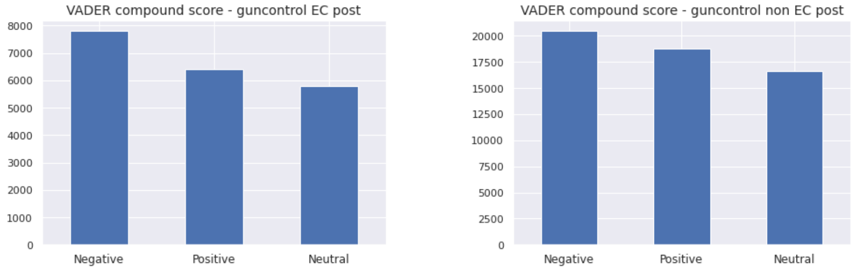
Minority. *Minority* posts and comments show all the tendencies described at the beginning of the section (Figure 4.16) both in ECs and in Not-ECs. The general pattern of having a very high number of positive texts, is also noticeable in this case, although not nearly as high as the number of negative ones as registered in *Politics*.

Gun control. Even in *Gun Control* (Figure 4.17) there are no relevant differences among ECs and Not-ECs and between posts and comments. A thing that may be underlined is that text in *Gun Control* ECs are more similar, in terms of results, to the ones of *Politics*' Not-ECs (Figure 4.15b).



(a) ECs (b) Not-ECs

Figure 4.16: *Minority* sentiment results



(a) ECs (b) Not-ECs

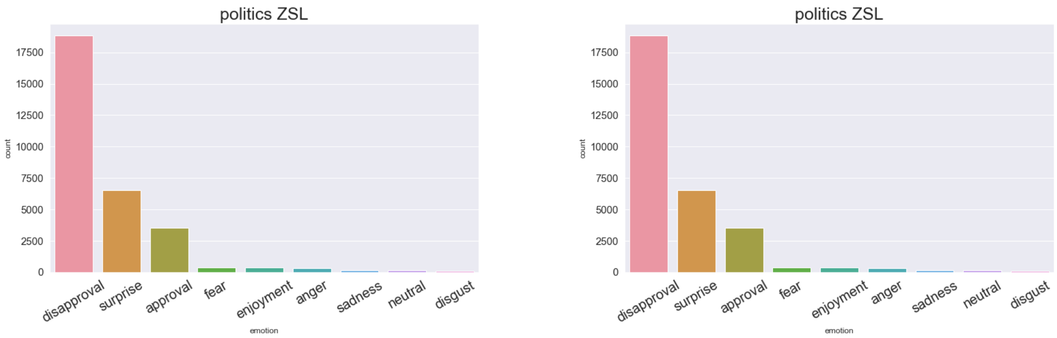
Figure 4.17: *Gun Control* sentiment results

4.5.3 Emotion analysis

The first approach employed to extract the general emotional leaning from posts and comments – with sparser and similar results – was *EmoRoberta*. The algorithm was applied to the posts and the comments and returned as output the most probable label among the likely classes on which the model was trained. A common trait among all the EmoRoBERTa results is that the majority of the input data has been tagged as *neutral*. The majority of labels obtained with both approaches

Post	**Community type**	**Dataset**	**Label EmoRoBERTa**	**Label Zero-shot learning**
so glad my babies are back!	EC	Gun control	joy	approval
a tommy gun? yes please!	Not EC	Gun control	curiosity	approval
ask the npcs who get upset over ""it's ok to be white"" this question. so you're saying it's not ok to be white????	EC	Minority	curiosity	surprise
new drunk driving laws don't increase safety. they just reduce rights.	Not EC	Minority	neutral	disapproval
trump hits new polling low as base shrinks missing	EC	Politics	surprise	neutral
a great comic in support of gun restrictions, to further my previous argument that you can be libertarian and support gun control	Not EC	Politics	approval	enjoyment

Table 4.3: Example from the tagged dataset



(a) EmoRoBERTa ECs

(b) ZSL ECs

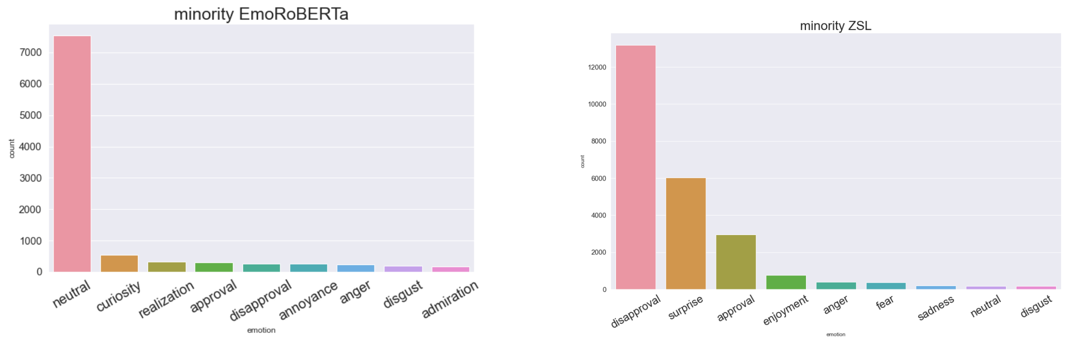
Figure 4.18: *Politics* sentiment results

to describe the data of each dataset is very similar, although for EmoRoBERTa's own nature the labels assigned describe the data from a finer-grain perspective. The explanation for this similarity might lie in the fact that Reddit is a virtual place to exchange opinions and discuss events or news both from the virtual and real world: this may foster common emotions, such as *curiosity*, *surprise* and even *disapproval* from users. **Politics.** In the *Politics* dataset, according to EmoRoBERTa, the majority of textual data is neutral, with a smaller heterogeneous composition of posts tagged as *curiosity, approval, disapproval* and *realization* (see Figure 4.18a), both in ECs and Not-ECs. The results may be interpreted by thinking at the discussions taking place in these subreddits, which might arise the curiosity of the readers or (dis-)approval *w.r.t.* an event of American politics.

Through ZSL, instead, the majority of emotion labels is *disapproval, surprise, approval* and *fear* inside ECs, while outside them there is a more consistent corpus of *enjoyment* labels.

Minority. In *Minority* apparently, the results are very similar among ECs and Not-ECs. Most of them are *neutral* for EmoRoBERTa (Figure 4.19a) or denoted by *curiosity, realization* and *approval*, both in ECs and Not-ECs.

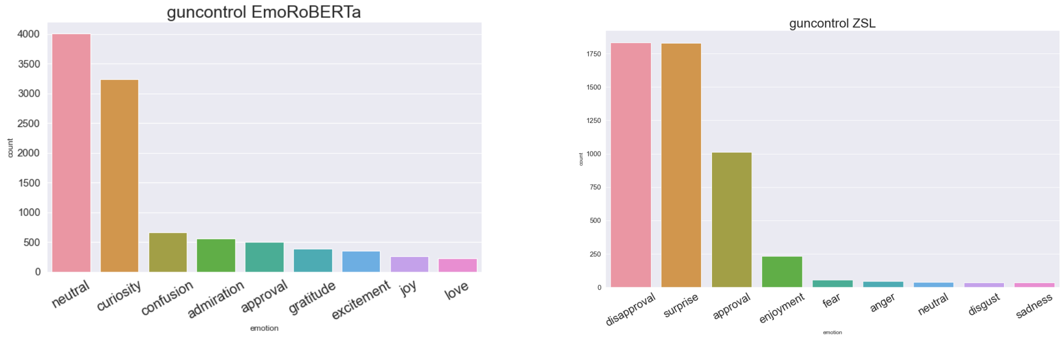
ZSL's results were very similar – as the ones obtained by applying EmoRoBERTa – between ECs and Not-ECs: the majority of posts in ECs/Not-ECs is tagged as *disapproval, surprise, approval* and *enjoyment.*



(a) EmoRoBERTa ECs

(b) ZSL Not-ECs

Figure 4.19: *Minority* sentiment results



(a) EmoRoBERTa ECs

(b) ZSL ECs

Figure 4.20: *Gun Control* sentiment results

Gun Control. *Gun Control* is maybe the most peculiar case, with a lot of posts tagged by EmoRoBERTa with the label *curiosity* and a smaller component of *confusion* in ECs (Figure 4.20a). This may be explained by thinking at the topics themselves and the aim of the subreddits since often *r/guncontrol* is used to ask for advice about weapons. In Not-ECs instead, there is a great component, almost equivalent to the neutral one, of *surprise* posts, followed by *approval.*

ZSL instead tagged most of the ECs' posts as *surprise*, *approval* and *disapproval*, which intuitively may be consistent with the concept of echo chamber itself: it is possible that many users inside them approve what is being said in a discussion or are surprised by an argument that confirms their beliefs.

4.5.4 Topic modeling

BERTopic was the algorithm chosen for the extraction of the most representative topics that characterize the three macro-categories on which we decided to focus this case study.

At first, after the application of the algorithm, the results were analysed as a whole and only later the focus shifted to a finer-grain analysis of the most characteristic topics of each community. In particular, referring to what was done in Section 4.3 to assess the ECs stability, we focused on analysing the results related to the community more at risk of being a strongly polarized EC.

Politics. In Politics' echo chambers (Figure 4.21) what emerge is that the resulting topics are often source of heated debates: among these we find several topics related to Trump's executive orders, such as the *2017 Trump travel ban*, by which Trump applied travel restrictions to people coming to U.S. from Middle-Eastern Countries, such as Iran, Iraq, Libya – all Muslim-majority countries. Other debates are about *net neutrality*, enacted by Barak Obama in 2015 and repealed in 2018 [9], sexual harassment and the Tax Cuts and Jobs Act, law passed in 2018.

The first topic, however, concerns the Alabama elections for the US Senate, a 2017 event where the Democratic candidate Doug Jones defeated Republican candidate Roy Moore becoming the first Democratic candidate after over 20 years. Focusing on the community more at risk of being an echo chamber, the users' interest is mainly focused on this last described event (Figure 4.22), on the *net neutrality* and *tax reform* of 2017. The more specific analysis of ECs, instead is interesting since it give a better explanation for the presence of the topics of net neutrality and abortion: these ones are often accompanied by the name of the journalist Ben Shapiro, who actually discussed these two issues during the period under analysis. It is possible to think that there is an echo-chamber-like system where users discuss the declarations of Shapiro, in an environment composed of people belonging to the same ideological leaning.

[9] New York Times, Net Neutrality Has Officially Been Repealed. Here's How That Could Affect You.: https://www.nytimes.com/2018/06/11/technology/net-neutrality-repeal.html, last visited: 02/11/2022

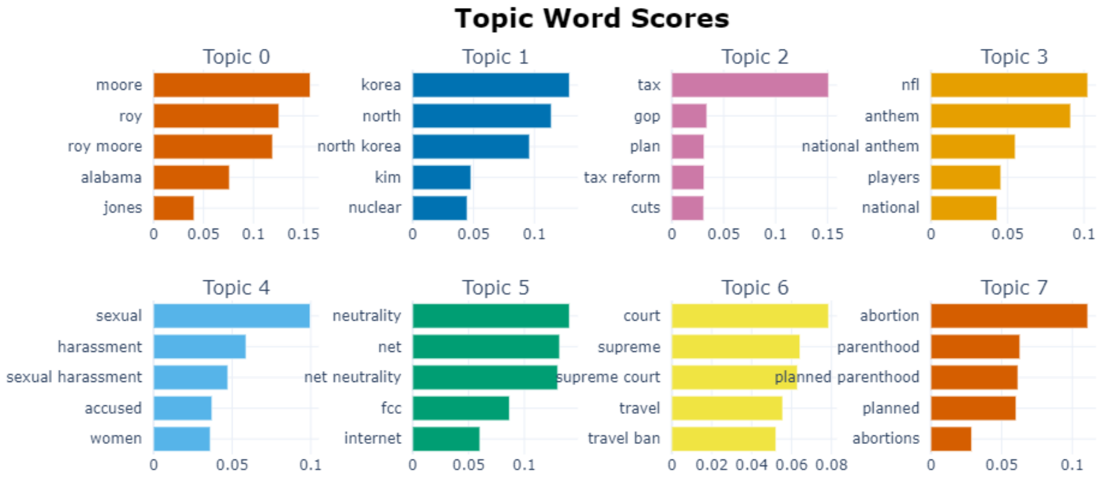


Figure 4.21: Main topics in *Politics* ECs

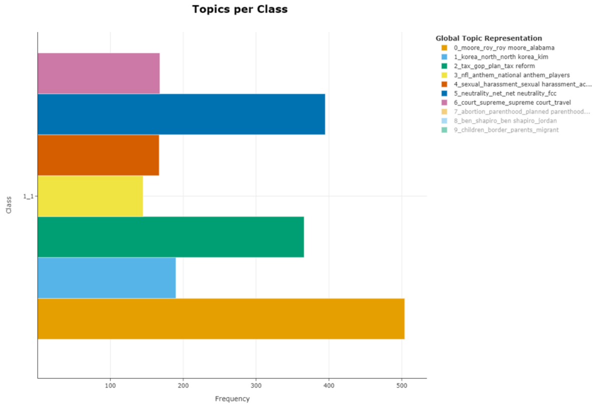


Figure 4.22: Topics of the community with a higher risk of being EC (*Politics*)

In Not-ECs, apparently there are significant topics, as is the case in ECs (Figure 4.23). Focusing instead on the analysis of the community least likely to be an EC, it is possible to unveil that most of the topics are actually Trump-related and considered as noise by BERTopic, while just a few others are actually "polarising" topics, such as climate change and US healthcare.

Minority. *Minority* has undoubtedly the most interesting results since most of the topics extracted from ECs are more controversial from a political point of view than the ones found in the Not-ECs. In Not-ECs, in fact, in addition to controver-

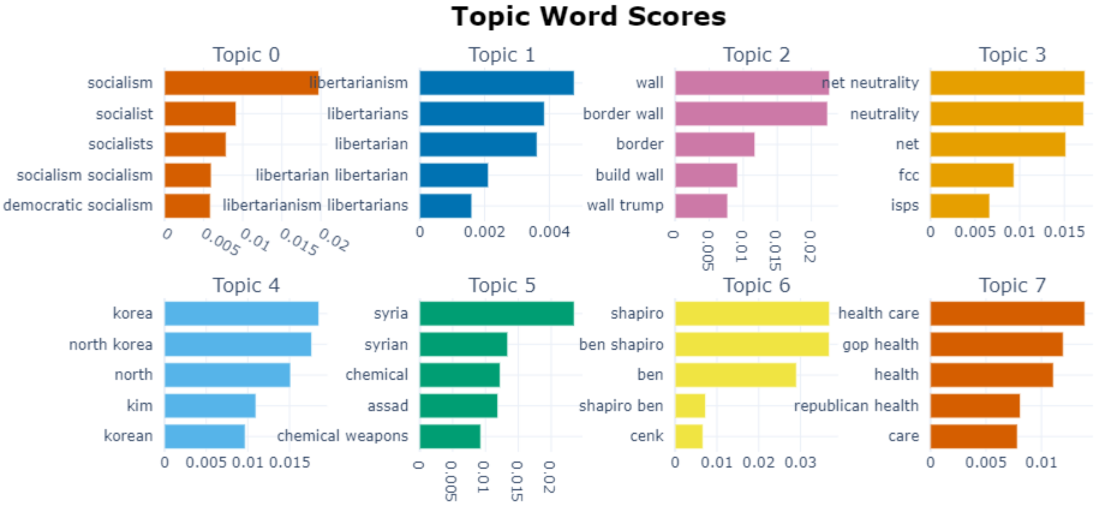


Figure 4.23: Main topics in *Politics* Not-ECs

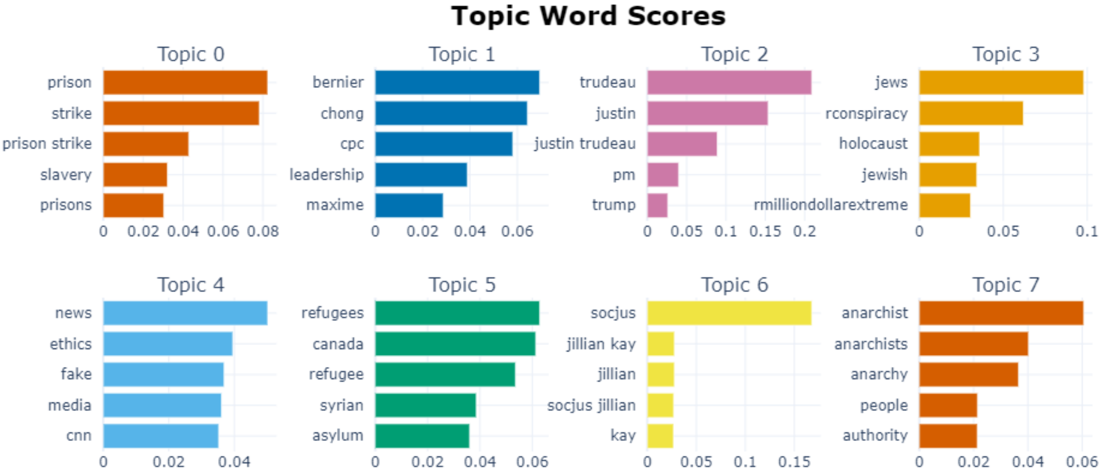


Figure 4.24: Main topics in *Minority* ECs

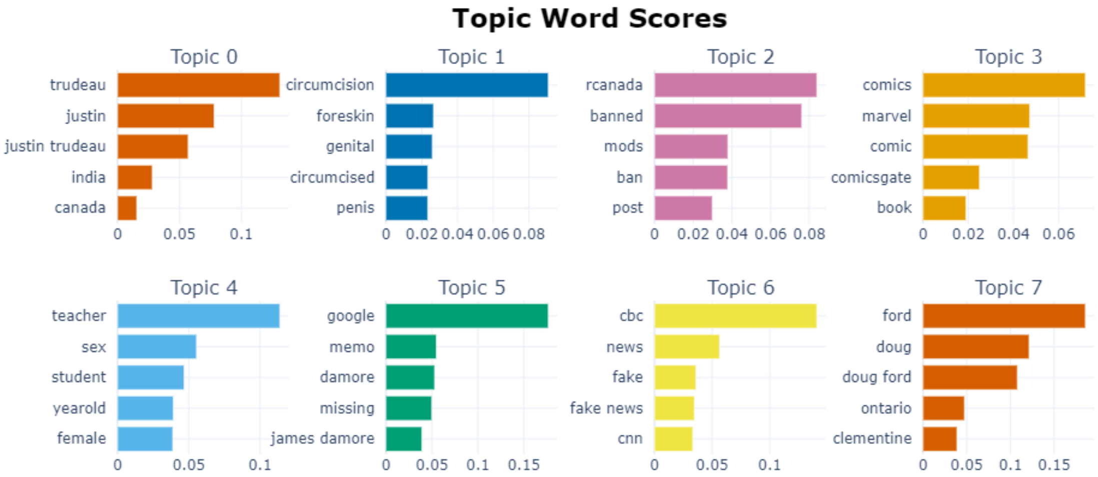


Figure 4.25: Main topics in *Minority* Not ECs

sial topics, there are others that are less involved in political polarization, memes about the ban-wave in the subreddit *r/canada*. An interesting topic that may be ideologically polarized but was not discussed inside a topological EC is the so-called *Comicsgate*, a boycott campaign of the far right against the "forced diversity" in U.S. superhero comic books; other topics that have raised a lot of discussions are the travel to India of the Canadian Prime Minister Justin Trudeau and the *memorandum Google's Ideological Echo Chamber* described in the Topic 2 in Figure 4.26). In this case was truly necessary to analyse the community more at risk of being an echo chamber to fully understand the behaviour of *Minority* ECs, as shown in Figure 4.26. In this case, two main topics emerge – the U.S. prison strike in 2016 and anarchism – are intertwined. Both of them appear in the same subreddit (*r/anarchism*) and the first event, born in response to the underpaid jobs that prisoners are forced to do, was well received by anarchist movements that see prisons as the embodiment of power and domination against people. The other topics in the EC, instead are unrelated and of marginal importance compared to the preponderant theme.

Gun Control. As concerns *Gun Control*, in ECs there are more controversial topics, even if there are potential triggering topics also in the Not-ECs.

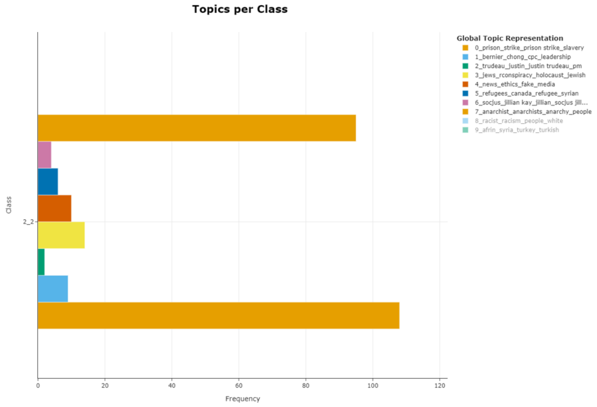


Figure 4.26: Topics of the community with higher risk of being EC (*Minority*)

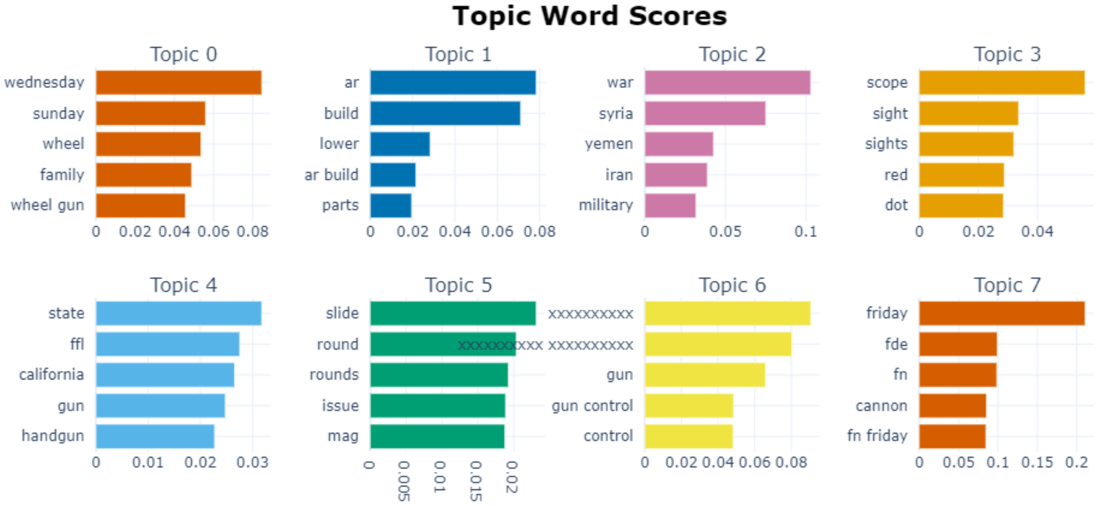


Figure 4.27: Main topics in *Gun Control* ECs

The most representative topic inside ECs (Figure 4.27) is the war in Syria and the California FFL, the licence that allows buying and bringing guns in California. Other topics are sparser, given also the nature of the debate, and they are tied to general talks about guns or advice.

In Not-ECs, instead, there are heterogeneous topics, including talks about gun collections and specific models (i.e., Sig Sauer). There are also more specific and controversial topics, such as school mass shootings and the scandal about NRA and its alliance with Russia. Even in *Gun Control*, it was necessary to look at the community with a higher risk of being EC, namely *community 2_8*. As it can be

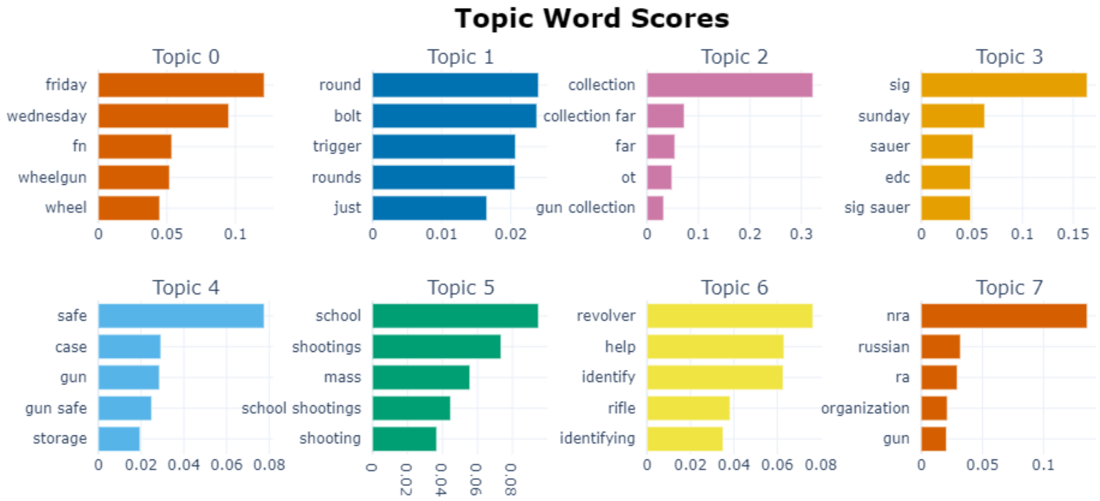


Figure 4.28: Main topics in *Gun Control* Not ECs

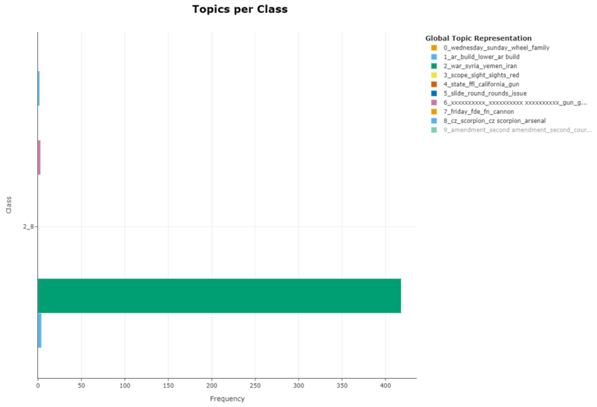


Figure 4.29: Topics of the community with a higher risk of being EC (*Gun Control*)

seen from Figure 4.29, there is only one topic that outweighs all the others, and that is the war in Syria. This topic can be found also in other ECs, but never with the same importance.

Chapter 5

Conclusions

In this thesis, we introduced a methodology to both analyze echo chambers' diachronic evolution and to characterize possible users' behaviours therein through their linguistic productions.

This multidisciplinary framework was applied to a specific case study. It was chosen an extremely controversial topic, namely American politics during the first two years and a half of Donald Trump's presidency, with the aim to track and characterize democratic and republican echo chambers. These polarized systems were studied leveraging Reddit socio-political discussions.

After introducing the problem and conducting the literature review, we described two different approaches.

At first, we leveraged network science to assess whether echo chambers are actually stable over time. To do so, for each semester under analysis we extracted ideologically homogeneous communities using the *EVA* algorithm. Then, we evaluated the communities extracted in terms of *Purity* and *Conductance* [MPR21], which allowed us to set a boundary (i.e., $Purity \geq 0.7$ and $Conductance \leq 0.4$) between communities more at risk of being echo chambers and the ones with lower risk.

Then, for each community of each semester, we leveraged the Jaccard index to assess its evolution in the next semester. Here, we observed how high-risk communities tend to be such for a long period of time, often persisting for up to a year and a half. In addition to this, we have proven that echo chambers do clearly tend to keep their internal composition in terms of users, thus reinforcing the theory that echo chambers are isolated systems, with a long lifecycle.

We proceeded then to study the linguistic productions of the users, by exploiting a wide range of approaches to address the problem. First of all, we extracted a small set of linguistic features, but in this case, they did not give relevant insights on the users involved.

Subsequently, the focus shifted to sentiment and emotion analysis. The former, implemented using VADER, revealed that Reddit users, both inside and outside polarized systems, write texts tagged as negative. On the other hand, instead, emotion analysis was implemented through two different approaches: a transformer model (i.e., EmoRoBERTa) fine-tuned on Reddit comments tagged with a range of 28 different emotions, and Zero-Shot Learning, which allowed tagging posts with 8 arbitrarily chosen labels. The results unveiled that there are no relevant distinctions between textual data produced inside and outside echo chambers. This may be due to Reddit's own nature since it is a platform born to share interesting contents and news that may arise emotions such as *curiosity* or *surprise*, two of the most common labels extracted.

Undoubtedly, the most interesting results were obtained through topic modeling: through BERTopic it was possible to demonstrate that the majority of users in echo chambers tend to discuss only one specific controversial topic, rather than multiple ones, as happens outside these polluted systems.

Contributions. The diachronic study of echo chambers and the users' profiling conducted in this work tried to advance the knowledge on the evolution and characterization of the behaviour of polluted information systems, in which there is still a gap in the literature. First of all, to the best of my knowledge, previous works have approached the first problem only from a static point of view, using a flattened snapshot of interactions over a specific time window. This may easily introduce biases since such representations do not keep into account the temporal ordering of relations, thus leading to an overestimation of the interactions.

The other focus of the work has been the study of the behaviour of users inside echo chambers in a finer grain, an aspect that has often been ignored by previous works.

Through the analysis of linguistic productions, it was possible to assess that the most persistent echo chambers over time focus on the discussion of a single topic – or two closely related topics – that is strongly controversial. This does not happen in less polarized systems, where the users tend to split their attention over a wider range of different topics. This pattern, repeated in the different topics of analysis, is, without doubt, an insight that might be useful also to better define mitigation strategies for this kind of polluted phenomenon.
Additionally, all the steps proposed in this framework are generalisable since they do not leverage platform-specific features: in this way, they can be applied also to other platforms (i.e., Twitter, Gab, Facebook) and on different topics than the socio-political ones.

Limitations. Nonetheless, this work is not free from limitations. Given its data-driven nature, it carries all the limits of this kind of approach, e.g., the data sparseness, which may prevent the identification of patterns that recur just a few times within the collected data sample that might bring a loss of useful insights.
In addition to this, it is necessary to perform further validations, both of the framework itself, i.e., testing it on data extracted from other platforms and/or about other topics, and of the results themselves, since in this thesis, we worked using an unsupervised approach in particular *w.r.t.* sentiment and emotion analysis.

Future developments As for future developments, an interesting aspect would be to study echo chambers moving from pairwise to high-order interactions. In this case, we could rely on structures such as hypergraphs and simplicial complexes, that have never been used to study echo chambers. The advantage would be to be able to capture further homophilic behaviours – e.g., peer pressure – that might go unnoticed using traditional graphs and may be useful to gain additional insights on echo chambers.
Other improvements might regard also testing different algorithms and approaches for the text analysis. For example, it would be possible to fine-tune other models for emotion analysis or to perform linguistic profiling of the text in analysis, so as to

have more accurate results and gain other insights that may have gone unnoticed. Lastly, another key aspect is to define the role of attractors, i.e. *hubs*, in the formation of echo chambers and in their lifecycle. This would make it possible to gain new insights into additional patterns that characterize this phenomenon and define new strategies to predict and consequently mitigate their polluting behaviour.

Bibliography

[ABKK19] Ashley Amaya, Ruben Bach, Florian Keusch, and Frauke Kreuter. New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review*, 39(5):943–960, December 2019.

[ACT$^+$21] Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. A survey on echo chambers on social media: Description, detection and mitigation. *CoRR*, abs/2112.05084, 2021.

[Bar15] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91, 2015.

[BCC$^+$22] Chiara Buongiovanni, Roswita Candusso, Giacomo Cerretini, Domenico Febbe, Virginia Morini, and Giulio Rossetti. Will you take the knee? Italian Twitter Echo Chambers' Genesis during Euro 2020. 2022.

[BCZ$^+$19] Emanuele Brugnoli, Matteo Cinelli, Fabiana Zollo, Walter Quattrociocchi, and Antonio Scala. Lexical convergence inside and across echo chambers. *CoRR*, abs/1903.11452, 2019.

[BE07] Danah M. Boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, October 2007.

[Bes16] Alessandro Bessi. Personality traits and echo chambers on facebook. *Comput. Hum. Behav.*, 65(C):319–324, dec 2016.

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.

[BJN$^+$15] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. Tweeting from left to right. *Psychological Science*, 26(10):1531–1542, August 2015.

[BKL$^+$19] Jie Bai, Qingchao Kong, Linjing Li, Lei Wang, and Daniel Zeng. Exploring cognitive dissonance on social media. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, July 2019.

[BMA15] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132, June 2015.

[BMR$^+$20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003.

[Bou15] Shelley Boulianne. Social media use and participation: a meta-analysis of current research. *Information, Communication & Society*, 18(5):524–538, March 2015.

[BP16] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, Cambridge, 2016.

[BPV$^+$15] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Viral misinformation. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, May 2015.

[Bri18] Jonathan Bright. Explaining the Emergence of Political Fragmentation on Social Media: The Role of Ideology and Extremism. *Journal of Computer-Mediated Communication*, 23(1):17–33, January 2018.

[Bru10a] Jennifer Brundidge. Encountering "difference" in the contemporary public sphere: The contribution of the internet to the heterogeneity of political discussion networks. *Journal of Communication*, 60(4):680–700, November 2010.

[Bru10b] Jennifer Brundidge. Political discussion and news use in the contemporary public sphere: The "accessibility" and "traversability" of the internet. *Javnost / The Public*, 17:63–82, 06 2010.

[BZK$^+$20] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839, May 2020.

[CBWG11] Rajmonda Sulo Caceres, Tanya Berger-Wolf, and Robert Grossman. Temporal scale of processes in dynamic networks. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 925–932, 2011.

[CGP11] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, 4(5):512–546, September 2011.

[Chu20] Petr Chunaev. Community detection in node-attributed social networks: A survey. *Computer Science Review*, 37:100286, August 2020.

[CLL$^+$15] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A novel neural topic model and its supervised extension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), February 2015.

[CM19] Uthsav Chitra and Christopher Musco. Understanding filter bubbles and polarization in social networks. *CoRR*, abs/1906.08772, 2019.

[CMG$^+$21] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), February 2021.

[CR19] Remy Cazabet and Giulio Rossetti. Challenges in community discovery on temporal networks. In *Computational Social Sciences*, pages 181–197. Springer International Publishing, 2019.

[CR20] Salvatore Citraro and Giulio Rossetti. Identifying and exploiting homogeneous communities in labeled networks. *Applied Network Science*, 5(1), August 2020.

[CRA14] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, 64(2):317–332, March 2014.

[CRRS08] Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press, 2008.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018.

[DMAK$^+$20] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July 2020. Association for Computational Linguistics.

[Ekm05] Paul Ekman. Basic emotions. In *Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons, Ltd, January 2005.

[EL86] Roger A. Elkin and Michael R. Leippe. Physiological arousal, dissonance, and attitude change: Evidence for a dissonance-arousal link and a "don't remind me" effect. *Journal of Personality and Social Psychology*, 51(1):55–65, 1986.

[EY22] Roman Egger and Joanne Yu. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7, May 2022.

[Fes54] Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, May 1954.

[Fes62] Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4):93–106, 1962.

[FGR16] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(Specialissue1):298–320, 2016.

[FH16] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, jul 2016.

[FI10] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the beta-divergence, 2010.

[For09] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, jun 2009.

[Gar18] Kiran Garimella. *Polarization on Social Media*. Doctoral thesis, School of Science, 2018.

[GBK09] E. Gilbert, T. Bergstrom, and K. Karahalios. Blogs are Echo Chambers: Blogs are Echo Chambers. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10, 2009.

[GDGM18] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, pages 913–922. ACM, 2018.

[GdMTC21] Quentin Grossetti, Cedric du Mouza, Nicolas Travers, and Camelia Constantin. Reducing the filter bubble effect on Twitter by considering communities for recommendations. *International Journal of Web Information Systems*, 17(6):728–752, 2021.

[GGD13] Yogesh Girdhar, Philippe Giguère, and Gregory Dudek. Autonomous Adaptive Underwater Exploration using Online Topic Modeling. In *Experimental Robotics*, pages 789–802. Springer International Publishing, 2013.

[GMGM17] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. The effect of collective attention on controversial debates on social media. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, June 2017.

[GMGM18a] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press, 2018.

[GMGM18b] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying Controversy on Social Media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.

[Gra83] Mark Granovetter. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, 1:201, 1983.

[Gro22] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[HG14] C. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[HM97] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting on Association for Computational Linguistics -*. Association for Computational Linguistics, 1997.

[HMKM17] Nida Manzoor Hakak, Mohsin Mohd, Mahira Kirmani, and Mudasir Mohd. Emotion analysis: A survey. In *2017 International Conference on Computer, Communications and Electronics (Comptelix)*. IEEE, July 2017.

[HS13] Petter Holme and Jari Saramäki, editors. *Temporal Networks*. Springer Berlin Heidelberg, 2013.

[IAG+19] Roberto Interdonato, Martin Atzmueller, Sabrina Gaito, Rushed Kanawati, Christine Largeron, and Alessandra Sala. Feature-rich networks: going beyond complex network topologies. *Applied Network Science*, 4(1), January 2019.

[JH16] Klaus Bruhn Jensen and Rasmus Helles. Speaking into the system: Social media and many-to-one communication. *European Journal of Communication*, 32(1):16–25, December 2016.

[JVHB14] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, 9(6):e98679, June 2014.

[JZLC19] Myeongki Jeong, Hangjung Zo, Chul Ho Lee, and Yasin Ceran. Feeling displeasure from online social media postings: A study using cognitive dissonance theory. *Computers in Human Behavior*, 97:231–240, August 2019.

[KAZ19] Sayeed Ahsan Khan, Mohammed Hazim Alkawaz, and Hewa Majeed Zangana. The use and abuse of social media for spreading fake news. *2019 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2019 - Proceedings*, pages 145–148, 6 2019.

[KHJ08] Joseph N. Cappella Kathleen Hall Jamieson. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment.* Oxford University Press, USA, 2008.

[KK98] George Karypis and Vipin Kumar. Multilevelk-way partitioning scheme for irregular graphs. *Journal of Parallel and Distributed Computing*, 48(1):96–129, January 1998.

[Kla60] J.T. Klapper. *The Effects of Mass Communication.* Foundations of communications research. Free Press, 1960.

[KP07] Anne Kao and Stephen R. Poteet, editors. *Natural Language Processing and Text Mining.* Springer London, 2007.

[LCG$^+$19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2019.

[LEB08] H. Larochelle, D. Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[LLG+20] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[LMV19] Matthieu Latapy, Clémence Magnien, and Tiphaine Viard. Weighted, bipartite, or directed stream graphs for the modeling of temporal networks. *CoRR*, abs/1906.04840, 2019.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[LTD+16] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1), September 2016.

[LVM18] Matthieu Latapy, Tiphaine Viard, and Clémence Magnien. Stream graphs and link streams for the modeling of interactions over time. *Social Network Analysis and Mining*, 8, 10 2018.

[MFF+17] Matteo Morini, Patrick Flandrin, Eric Fleury, Tommaso Venturini, and Pablo Jensen. Revealing evolutions in dynamical networks. *CoRR*, abs/1707.02114, 2017.

[MHA17] Leland McInnes, John Healy, and Steve Astels. HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

[MHSG18] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, September 2018.

[MMS21] Gianmarco De Francisci Morales, Corrado Monti, and Michele Starnini. No echo in the chambers of political interactions on reddit. *Scientific Reports*, 11(1), February 2021.

[Moo01] James Moody. Race, school integration, and friendship segregation in america. *American Journal of Sociology*, 107(3):679–716, 2001.

[MPR21] Virginia Morini, Laura Pollacci, and Giulio Rossetti. Toward a Standard Approach for Echo Chamber Detection: Reddit Case Study. *Applied Sciences*, 11(12):5390, June 2021.

[MS20a] Moritz Markgraf and Manfred Schoch. Quantification of echo chambers: A methodological framework considering multi-party systems. In *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019*, 2020.

[MS20b] Moritz Markgraf and Manfred Schoch. Quantification of echo chambers: A methodological framework considering multi-party systems. In *27th European Conference on Information Systems - Information Systems for a Sharing Society, ECIS 2019*, 2020.

[MSLC01] Miller McPherson, Lynn Smith-Lovin, and James Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–, 01 2001.

[MUMM07] François Mairesse, Marilyn A Uk, Matthias R Mehl, and Roger K Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[New03] M. E. J. Newman. Mixing Patterns in Networks. *Physical Review E*, 67(2), February 2003.

[Ngu20] C. Thi Nguyen. Echo Chambers and Epistemic Bubbles. *Episteme*, 17(2):141–161, 2020.

[NY03] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis. In *Proceedings of the international conference on Knowledge capture - K-CAP '03*. ACM Press, 2003.

[Pap02] Zizi Papacharissi. The virtual sphere. *New Media &amp Society*, 4(1):9–27, February 2002.

[Par11] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Group , The, 2011.

[PBV07] Gergely Palla, Albert-László Barabási, and Tamás Vicsek. Quantifying social group evolution. *Nature*, 446(7136):664–667, apr 2007.

[PDL18] Leto Peel, Jean-Charles Delvenne, and Renaud Lambiotte. Multiscale mixing patterns in networks. *Proceedings of the National Academy of Sciences*, 115(16):4057–4062, April 2018.

[Pet20] Uwe Peters. What is the function of confirmation bias? *Erkenntnis*, 87(3):1351–1376, April 2020.

[PL20] Dang Pham and Tuan Le. Auto-encoding variational Bayes for inferring topics and visualization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5223–5234, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[PSÅ19] Mina Young Pedersen, Sonja Smets, and Thomas Ågotnes. Analyzing Echo Chambers: A Logic of Strong and Weak Ties. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11813 LNCS, pages 183–198, 2019.

[QSS16] Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein. Echo chambers on facebook. *SSRN Electronic Journal*, 2016.

[RC18a] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks. *ACM Computing Surveys*, 51(2):1–37, June 2018.

[RC18b] Giulio Rossetti and Rémy Cazabet. Community discovery in dynamic networks: A survey. *ACM Comput. Surv.*, 51(2), feb 2018.

[RG19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[SCP$^+$13] Juliette Stehlé, François Charbonnier, Tristan Picard, Ciro Cattuto, and Alain Barrat. Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks*, 35(4):604–613, October 2013.

[SEEDT20] Mohamed Salama, Mohamed Ezzeldin, Wael El-Dakhakhni, and Michael Tait. Temporal networks: a review and opportunities for infrastructure simulation. *Sustainable and Resilient Infrastructure*, 7(1):40–55, February 2020.

[SFRM21] Manuel J. Sánchez-Franco and Manuel Rey-Moreno. Do travelers' reviews depend on the destination? an analysis in coastal and urban peer-to-peer lodgings. *Psychology & Marketing*, 39(2):441–459, October 2021.

[SGMN13] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 935–943, Red Hook, NY, USA, 2013. Curran Associates Inc.

[Sun99] Cass R. Sunstein. The law of group polarization. *SSRN Electronic Journal*, 1999.

[Sun07] Cass R. Sunstein. *Republic.Com 2.0.* Princeton University Press, USA, 2007.

[Sun18] Cass R. Sunstein. *#Republic.* Princeton University Press, April 2018.

[TTRB17] Joshua Tucker, Yannis Theocharis, Margaret Roberts, and Pablo Barberá. From liberation to turmoil: Social media and democracy. *Journal of Democracy*, 28:46–59, 10 2017.

[Tur87] J.C. Turner. *Rediscovering the Social Group: A Self-categorization Theory.* Basil Blackwell, 1987.

[VPV21] Giacomo Villa, Gabriella Pasi, and Marco Viviani. Echo chamber detection and analysis. *Social Network Analysis and Mining*, 11(1), August 2021.

[VSP$^+$17] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[VVB$^+$15] Cristian Vaccari, Augusto Valeriani, Pablo Barberá, Rich Bonneau, John T. Jost, Jonathan Nagler, and Joshua A. Tucker. Political expression and action on social media: Exploring the relationship between lower- and higher-threshold political activities among twitter users in italy. *Journal of Computer-Mediated Communication*, 20(2):221–239, January 2015.

[WBO99] Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics -*. Association for Computational Linguistics, 1999.

[WMKL15] Hywel T.P. Williams, James R. McMurray, Tim Kurz, and F. Hugo Lambert. Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, 32:126–138, May 2015.

[WZF12] Li Wan, Leo Zhu, and Rob Fergus. A hybrid neural network-latent topic model. In Neil D. Lawrence and Mark Girolami, editors, *Proceed-*

ings of the Fifteenth International Conference on Artificial Intelligence and Statistics, volume 22 of *Proceedings of Machine Learning Research*, pages 1287–1294, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[YdMdC^{+}19] Dani Yogatama, Cyprien de Masson d'Autume, Jerome T. Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. Learning and evaluating general linguistic intelligence. *CoRR*, abs/1901.11373, 2019.

[YHR19] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019.

[ZCY09] Yang Zhou, Hong Cheng, and Jeffrey Xu Yu. Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow.*, 2(1):718–729, aug 2009.

[ZLG19] Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1031–1040, 2019.

[ZSH^{+}19] Bo Zhao, Xinwei Sun, Xiaopeng Hong, Yuan Yao, and Yizhou Wang. Zero-shot learning via recurrent knowledge transfer. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1308–1317, 2019.