HAL open science

# A Multi-Factorial Analysis of Polarization on Social Media

Celina Treuillier, Sylvain Castagnos, Armelle Brun

## To cite this version:

Celina Treuillier, Sylvain Castagnos, Armelle Brun. A Multi-Factorial Analysis of Polarization on Social Media. UMAP '23 Adjunct: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, Jun 2023, Limassol, Cyprus. 10.1145/3563359.3597393 . hal-04108988

**HAL Id: hal-04108988**

**https://hal.science/hal-04108988v1**

Submitted on 30 May 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A Multi-factorial Analysis of Polarization on Social Media

**Célina TREUILLIER**
Université de Lorraine, CNRS, LORIA
Nancy, FRANCE
celina.treuillier@loria.fr

**Sylvain CASTAGNOS**
Université de Lorraine, CNRS, LORIA
Nancy, FRANCE
sylvain.castagnos@loria.fr

**Armelle BRUN**
Université de Lorraine, CNRS, LORIA
Nancy, FRANCE
armelle.brun@loria.fr

**Abstract**

Polarization is an increasingly worrying phenomenon within social media. Recent work has made it possible to detect and even quantify polarization. Nevertheless, the few existing metrics, although defined in a continuous space, often lead to a unimodal distribution of data once applied to users' interactions, making the distinction between polarized and non-polarized users difficult to draw. Furthermore, each metric relies on a single factor and does not reflect the overall user behavior. Modeling polarization in a single form runs the risk of obscuring inter-individual differences. In this paper, we propose to have a deeper look at polarized online behaviors and to compare individual metrics. We collected about 300K retweets from 1K French users between January and July 2022 on Twitter. Each retweet is related to the highly controversial vaccine debate. Results show that a multi-factorial analysis leads to the identification of distinct and potentially explainable behavioral classes. This finer understanding of behaviors is an essential step to adapt news recommendation strategies so that no user gets locked into an echo chamber or filter bubble.

***Keywords*** Polarization Metrics, Social Media, User Behavioral Classes, Opinions, Sources

## 1 Introduction

The influence of social media (SM) is tremendously growing worldwide. A recent study from the Pew Research Center [1] has shown that one in five adults gets her news primarily through SM, and tends paradoxically to be less well-informed. The impact of SM on polarization can then be explained in two ways. On the one hand, a lack of information due to filter bubbles and echo chambers can affect the degree of polarization of users [1]. On the other hand and paradoxically, the more a polarized user is exposed to opposing views on SM, the more her degree of polarization increases [2]. The balance to be found in terms of diversity of opinions, sources, and content is therefore extremely delicate.

Recent studies suggest that it would be more appropriate to tailor the level of diversity and recommendation strategies to behavioral classes [3, 4, 5] rather than to maximize diversity in the same way for all users [6, 7, 8]. In order to do so, user polarization behaviors need to be understood and modeled in detail. Our work is a first step towards this goal of adapting the recommendations.

Polarization has been investigated in the literature from two different perspectives: a network perspective which mainly relies on the SM structure in order to quantify the polarization of a community and to highlight the content

[1] https://www.pewresearch.org/journalism/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/

of discussions [9], and an individual perspective to measure the polarization and the impact of recommender systems through the diversity level of accepted recommendations [10]. In this paper, we focus on this second perspective, since we aim to quantify the level of polarization at the individual level, and we define 2 research questions. **RQ1**: Do the current individual polarization metrics contribute to distinguishing polarized users from non-polarized users? **RQ2**: Can a multi-factorial analysis identify different classes of polarization behavior?

In the rest of this paper, section 2 consists of a literature review about polarization metrics on SM. Section 3 details the experimental setup. The multi-factorial analysis is presented in Section 4, while conclusions and perspectives are drawn in Section 5.

## 2 Literature Review

SM play an important role in polarization [11, 12]. Conover *et al.* were the first to study polarization on Twitter [13]. Since then, many polarization metrics have been put forward and used in the literature, such as modularity [14], controversy [15], or metrics based on a probability density function [16]. Though these polarization metrics rely on a varied set of information, they share a characteristic: they quantify overall polarization using a graph representing users and their interactions. However, as well as being explained by social or technological filters, polarization is also influenced by individual factors [17] and only a few metrics are designed to quantify individual polarization.

Among these individual metrics, we can find the *polarization score* proposed by Becatti *et al.* [18]. The latter is based on the identification of a set of communities $C$. The polarization score of a user $u$ depends on the ratio of interactions in each community, with $N_{u,c}$ the number of interactions of $u$ in community $c$, and $N_u$ her total number of interactions. The polarization score $\rho(u)$ of user $u$ corresponds to the maximum of this ratio, as presented in Equation (1).

$$\rho(u) = \max_{c \in C} \{ \frac{N_{u,c}}{N_u} \} \tag{1}$$

Highly polarized users, accessing a unique community have $\rho = 1$, while users equally accessing all the communities have $\rho = 1/C$. In our view, this metric has several limits: its minimum bound depends on the number of communities, it only takes into account the community with which the user has interacted the most, and it does not inform about which communities are accessed. In a two-community context, the polarization score presented by Schmidt *et al.* is similar but ranges between $-1$ and $1$ and is oriented: the value informs about which community is accessed more often [19].

To go further, Cicchini *et al.* [20] propose the *Lack of Diversity* (LD) metric that is, to some extent, highly similar to the *polarization score* of Becatti *et al.* [18]. The main difference lies in the fact that it considers sources of information a user interacted with, concretely a set of $M$ media outlets, rather than communities. Each user $u$ is represented by her number of interactions $N_{u,m}$ on news from media $m$. LD is computed as follows:

$$LD(u) = \max_{m \in M} \{ N_{u,m} \cdot log(\frac{|U|}{|U_m|}) \} \tag{2}$$

$|U|$ is the total number of users and $|U_m|$ is the number of users interacting with media $m$. The term $log(\frac{U}{U_m})$ corrects a potential bias introduced by $m$ when shared by a large number of users. As in [18], $LD$ represents the maximum value within the vector. Thus calculated, the LD metric is not bounded, and should therefore be normalized.

These graph-based metrics, which quantify polarization at the individual level, face a main limit: they are only computed on a single factor (communities or media outlets). However, polarization is known to occur over the influence of multiple factors [21, 22]. As a consequence of considering only a single factor, several users may be identified as similarly polarized, but may in fact exhibit a wide range of behaviors and distinguish themselves in different ways.

To summarize, while different metrics have been proposed in the literature, individual polarization metrics are still scarce. Besides, to the extent that they are based on each user's preferred behavior (*i.e.*, the maximum value of the observed variable), we question their ability to accurately differentiate polarization behaviors on SM. This is why we propose to gradually conduct a multi-factorial analysis.

## 3 Experimental Setup: Polarization about the Vaccine Debate on Twitter

To answer our research questions (**RQ1**) and (**RQ2**), we propose to study a real SM context. We focus on the highly controversial vaccine debate, which was widely discussed following the COVID-19 crisis.

### 3.1 Data Collection

We used the Twitter API (*v2*), with academic research access. Data collection relies on the concept of elite users [23] that represent users who are relevant to the subject matter. We assume that elite users' tweets, when related to the selected topic, are always in line with their beliefs. Inspired by the methodology of Primario *et al.*, we fix conditions that elite users must satisfy to ensure their legitimacy: (1) have a significant number of followers; (2) personally manage their Twitter account; (3) are known by the general audience, through media or government interventions; (4) are qualified by education and/or profession to address the subject of matter.

Elite users are an effective entry point for collecting data about a specific topic because their opinions are publicly known [23]. Nevertheless, as our objective is to analyze standard users' interaction behaviors, it is necessary to have a faithful overview of standard users' interactions about the selected topic during a specific period. The dataset must be balanced in terms of opinion carriers, and representative of behaviors adopted on SM about a specific topic.

To obtain such a dataset, we carried out several steps, run after having chosen the topic, identified a relevant set of elite users, and defined a collection period. These steps are: (1) Collect all tweets published by the set of elite users during the predefined period; (2) Filter tweets about the topic of interest; (3) Collect information about a random subset of interacting standard users for each collected tweet; (4) Identify the most active standard users among those selected in Step 3; (5) Collect all interactions of selected standard users on collected elite users' tweets during the defined period.

Following the procedure detailed above, we manually identified 20 French-speaking elite users having a legitimate voice in the vaccine debate (10 pro-vaccine and 10 anti-vaccine). Their opinion is known because they have clearly expressed it publicly, and the community to which they relate is therefore unambiguous. To preserve their confidentiality and meet Twitter policy, we do not share the names or usernames of the selected accounts. We collected all elite users' tweets between January 1, 2022 and July 31, 2022. Based on relevant vaccine-related hashtags (that we stripped from the tweets) and a random tweet corpus [24], we trained a two-class classifier based on BertTweetFR [25]. This classifier allowed us to keep only elite users' tweets dealing with the vaccination debate. Here, we focus on retweets, which are signs of approval and thus give information about what users agree with [13]. Thus, we collected information about 100 randomly selected retweeters for each collected tweet, that we hope to be representative of all users. Among the selected retweeters, we focused on the 1,000 most active ones (500 pro-vaccine and 500 anti-vaccine).

### 3.2 Data Analysis

We collected 6,697 tweets in the period, divided into 1,869 tweets from pro-vaccine elite users, and 4,828 tweets from anti-vaccine elite users. From the 1,000 most active retweeters, we collected 11,449,936 retweets. 299,879 of these retweets were on elite users' tweets, with 16,791 retweets on pro-vaccine tweets, and 283,088 retweets on anti-vaccine tweets. This reflects a more intensive activity on the anti-vaccine side, which is consistent with the fact that anti-vaccine supporters are more engaged on Twitter, especially by doing many replies and retweets [26]. Looking at the structure of the graph, we identify 2 highly connected sets of nodes (*modularity*=0.55 [27]), with few edges between them. Besides, the controversy of the vaccine topic, computed using the Random Walk process described by Garimella *et al.* [15], is equal to 0.89. This indicates that it is difficult to move from one community to the other one. Altogether, these results first confirm that the selected elite users are tweeting pro-vaccine and anti-vaccine according to the opinion for which they were chosen. Second, they corroborate polarized attitudes towards the vaccination debate within our dataset.

## 4 Towards a Multi-factorial analysis of Polarization Behaviors

In this section, we first analyze the information returned by each polarization metric separately, relying on a single factor. Second, we conduct a multi-factor analysis (bi-factor and tri-factor) by considering several metrics. We evaluate the ability to accurately differentiate and characterize polarization behaviors in these different experimental conditions.

### 4.1 A Single Factor Analysis: Metrics from the Literature

To study polarization, we separately study two factors: (1) **Opinion factor**, where opinions are assessed from the standard users' retweets on each community (pro- or anti-vaccine); (2) **Source factor**, where sources are assessed from standard users' retweets on each elite user, who act as sources. To quantify polarization on these two factors, we rely on two individual metrics: the *polarization score* $\rho$ [18] and the *Lack of Diversity* $LD$ [20] (see Section 2).

Here, as we deal with single factor data, we use Kernel Density Estimation (KDE) rather than well-known multidimensional clustering algorithms to differentiate potential clusters. KDE estimates the probability density function of the studied single factor clusters based on local minima and maxima.

Applied to $\rho$, the estimated kernel density has no local maxima or minima. This does not allow the differentiation of well-separated clusters. Looking closer at the values of $\rho$ for all standard users, we note that it is not well distributed in $[0, 1]$, and the minimum value is $0.5$. This is one of the limits of the metric (See Equation (1)), which is bounded in $[0.5, 1]$ in a 2-community context. Users having at least one interaction in each of the communities represent only $18.8\%$ of users, and KDE does not estimate a probability density function allowing to clearly differentiate them. Thus, $\rho$ cannot accurately and finely characterize behaviors depending on whether users interact with 1 or 2 communities. In the same way, the kernel density estimated on $LD$ does not have local minima or maxima as values are distributed continuously. Therefore, this metric does not allow for a differentiation of polarization behavior in terms of access to information sources neither.

### 4.2 A Bi-factor Analysis

In this second step, we rely on a clustering algorithm to identify behavior clusters by combining opinion and source factors. We choose $k$-means [28] as it is well suited when dealing with numerical features. The number of clusters $k$ is optimized by maximizing 2 traditional metrics: Davies-Bouldin Index [29] (the lower the better) and Silhouette Index [30] (the higher the better).

#### 4.2.1 Use of Metrics from the Literature

The optimal number of clusters obtained with $\rho$ and $LD$ factors is $k = 2$, where Davies-Bouldin Index $= 0.60$ and Silhouette index $= 0.56$. Unlike the single factor analysis, considering the two metrics together does allow for the identification of two distinct groups of users, who adopt different polarization behaviors. This bi-factor analysis results in a finer-grained modeling. Identified clusters are shown in Figure 1a. Here again, $\rho$ does not allow a clear differentiation as users interacting with a unique community and those interacting in both communities are not associated with different clusters. Users are clustered according to the $LD$ value, with the two clusters being separated by a threshold fixed around $LD = 0.6$. Users in the orange cluster C1 with higher values of $LD$ are those with a high polarization according to accessed sources (*i.e.* retweeted elite users), while the other users (blue cluster C2) retweet more elite users and in a balanced way. However, following the KDE estimation presented in Section 4.1, this threshold is difficult to interpret. Although allowing for the differentiation of two groups of users, which was not possible with a single factor analysis, the limitations identified with respect to the distribution of $\rho$ values and the $LD$ threshold used to delimit the clusters question the quality of the clustering step. We expect the identified clusters to group together users that are likely to adopt well-differentiated polarization behaviors.

#### 4.2.2 Refining Metrics from the Literature: Use of Entropy

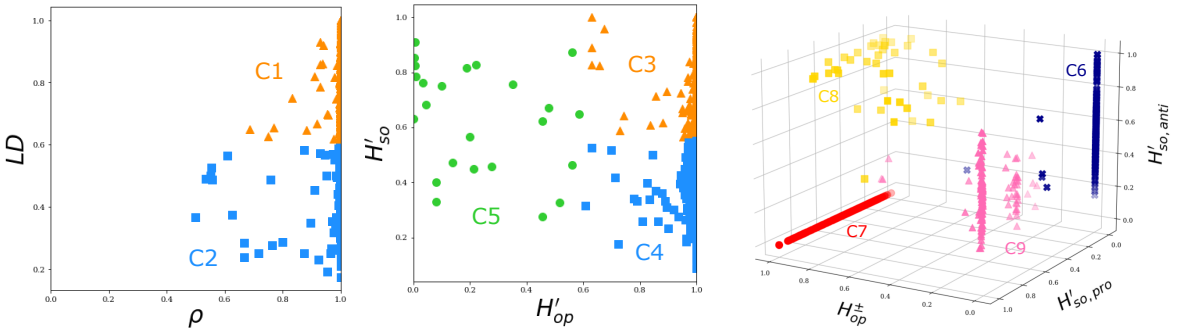
As previously mentioned, from our analysis, assessing individual polarization by only considering the predominant opinion ($\rho$) or source ($LD$) limits the ability to differentiate between users, and to understand polarization behaviors. To address this limitation, we propose to consider all interactions and represent them as a probability distribution. Following Information Theory, this makes it possible to compute entropy [31]. We thus propose to compute entropy-based metrics, measuring the uncertainty of access to opinions or sources.

Precisely, the more homogeneously distributed the probability mass, the higher the entropy and the greater the uncertainty. As the maximal entropy depends on the number of entities, we propose to use the normalized entropy, $H_N(Z) = \frac{-\sum_z^n P(z) log(P(z))}{log(n)}$, where $Z$ is a discrete random variable that takes $n$ possible values and $P(z)$ is the probability of entity $z$. For simplicity's sake, from now on, we will refer to normalized entropy as entropy. As entropy is null when there is no randomness and equal to one when the behavior is very heterogeneous, in this work we use $H'(X) = 1 - H_N(X)$ to get high scores for polarized users. To the best of our knowledge, no individual polarization metrics from the literature rely on it.

We note $H'_{op}$ the entropy-based opinion metric, and $H'_{so}$ the entropy-based source metric. First, comparing $\rho$ and $H'_{op}$, we can notice that unlike $\rho$, $H'_{op}$ ranges in $[0, 1]$. Though, as a very large proportion of users only interact with a single community, their polarization on opinion remains maximal, with $\rho = H'_{op} = 1$. The potential contribution of entropy to the differentiation of polarization behaviors will therefore be for the other users ($H'_{op} \neq 1$), representing $18.8\%$ of standard users. Second, comparing $LD$ and $H'_{so}$, we notice that values are distributed differently. First of all, $LD$ values range in $[0.17, 1]$ while $H'_{so}$ values range in $[0.09, 1]$. Second, the mean of $LD$ values is $0.59$, while the mean of $H'_{so}$ values is $0.50$. To go deeper into this comparison, we analyze the ranked values of $LD$ and $H'_{so}$, and get a Spearman score of 0.92 (*p-value*$= 0$). Although relatively high, this score indicates that a significant proportion of users observe large rank variations. Actually, $49.5\%$ of users have a variation higher than $5\%$ of the positions. Given a user $u_1$, who has more than half of her interactions with one elite user, $LD(u_1) = 0.52$. Looking at $u_1$'s other retweets, they are on only 3 other elite users, with two of them having only one retweet. This is reflected in $H'_{so}(u_1) = 0.91$,

which indicates a high level of polarization. For a significant proportion of users, the information returned by $H'_{so}$ is well different from that returned by $LD$, as for $u_1$, and could potentially allow for the identification of different classes of polarization behaviors.

Applying the $k$-means algorithm on bi-factor data, namely $H'_{op}$ and $H'_{so}$, the optimal number of clusters is $k = 3$, with Davies-Bouldin Index $= 0.58$ and Silhouette Index $= 0.57$. Performance is thus close to the one with $\rho$ and $LD$, but one additional cluster is identified. Clusters are represented in Figure 1b, with orange and blue clusters (C3 and C4) quite similar to the ones identified in Section 4.2.1. The green cluster (C5), corresponds to a subset of 24 users with lower $H'_{op}$ values, thus interacting with both communities. In an unprecedented way, entropy-based metrics thus allow differentiating standard users both on opinion and source factors, which was not the case with $\rho$ and $LD$. More importantly, clustering on $H'_{op}$ and $H'_{so}$ contributes to identifying a new subset of users, interacting with both communities and potentially acting as intermediates between pro-vaccine and anti-vaccine users. The resulting clusters thus provide useful observations about the polarization of users on SM. A bi-factor analysis, coupled with the use of entropy-based metrics, has made it possible to differentiate new classes of behavior. We now wish to assess to what extent additional factors could further improve the quality of the polarization behaviors modeling.




(a) $\rho$ and LD. (b) $H'_{op}$ and $H'_{so}$. (c) $H^{\pm}_{op}$, $H'_{so,pro}$ and $H'_{so,anti}$

Figure 1: Clusters identified with bi-factor data (3a and 3b) and tri-factor data (3c).

### 4.3 A Tri-factor Analysis

One main limit of the literature and of the previous sections is that the metrics evaluate to what extent users are polarized, but do not inform towards which community. We assume that this could improve the clustering.

In this respect, and considering the opinion metric, we propose to apply a transformation factor, as follows:

$$H^{\pm}_{op} = \frac{\pm H'_{op} + 1}{2} \tag{3}$$

The plus-minus sign in front of $H'_{op}$ depends on the predominant community. We set $H'_{op} > 0$ if interactions are in favor of the pro-vaccine community, and $H'_{op} < 0$ otherwise. The final $H^{\pm}_{op}$ values range in $[0, 1]$, with $H^{\pm}_{op} = 0$ indicating a very high polarization in the anti-vaccine community, $H^{\pm}_{op} = 1$, corresponding to an extreme polarisation in the pro-vaccine community and $H^{\pm}_{op} = 0.5$ reflecting balanced interactions between the two communities.

Considering the source metric, and still to inform about each community, we propose to split it into two metrics. The entropy-based metric $H'_{so}$, which initially combined sources from both communities, is split into two metrics $H'_{so,pro}$ and $H'_{so,anti}$, corresponding to the source factor computed on either pro or anti-vaccine community. This is designed to better differentiate users with unbalanced polarization in each of the two communities.

The optimization of $k$-means algorithm on the tri-factor data indicates that the optimal value of $k$ is 4, with Davies-Bouldin Index $= 0.51$ and Silhouette Index $= 0.74$. Performance is thus significantly higher than with two factors.

Looking at Figure 1c, representing identified clusters, we see that the clusters identified are quite different from those described in Section 4.2. First, blue and red clusters (C6 and C7) respectively correspond to highly polarized anti-vaccine users and pro-vaccine users. These users, accessing a unique community, do not all behave the same way according to the source factor as values are distributed in $[0, 1]$. Besides, the yellow and pink clusters (C8 and C9) are of particular interest. Looking more closely at the opinion and source metrics within these clusters, users in the yellow cluster (C8) are those having $H^{\pm}_{op}$ values close to $0.5$, indicating a balanced activity between the two

communities. In each accessed community, these users interact with a variety of sources, as $H'_{so,pro}$ and $H'_{so,anti}$ values are evenly distributed among users. Besides, users associated with the pink cluster (C9) interact predominantly with the anti-vaccine community. Nevertheless, they all have at least 1 interaction in the pro-vaccine community, in which they mostly interact with a few elite users ($H'_{so,pro} \approx 1$). Users taking part in these two unprecedented clusters are intermediate users as they interact in both communities. Moreover, in addition to being quite different from those identified in Section 4.2.2, they are also much more numerous. Yellow and pink clusters (C8 and C9) contain 43 and 140 standard users respectively, while only 24 intermediate users were previously identified. It appears that a significant proportion of users do not engage in extreme polarizing behaviors.

Overall, the identification of four patterns of polarization, only possible with the last analysis, is very interesting and reflects the multi-factorial nature of polarization. Identified behavioral classes are well-separated according to specific characteristics, which was not the case when considering only one or two factors on traditional or entropy-based metrics (Sections 4.1 and 4.2). Nevertheless, a few users are at the intersection between clusters, whose membership to one or other of the groups is questionable. For example, some users associated with the blue cluster C6 (*i.e.* highly polarized users in the anti-vaccine debate community) are also very close to the pink cluster C9. These users may be in a transitional phase which might be of interest.

## 5 Conclusion and Perspectives

The literature still lacks individual polarization metrics allowing fine-grained modeling of users' polarization behaviors. As the latter are the result of multiple influences, we conducted a multi-factorial analysis of polarization. Experiments confirm, first, that metrics from the literature using maximum value are too restrictive, and that the proposed entropy-based metrics allow a finer distinction between polarization behaviors, both for opinion and source factors. It does indeed contribute to the identification of an additional cluster of under-represented users, who have retweet interactions in both communities. These users adopt moderately polarized behavior about a highly polarized topic on SM. Altogether, these results indicate that current polarization metrics do not distinguish polarization behaviors properly (**RQ1**), and that entropy-based metrics seem better adapted. Besides, conducting a tri-factor analysis allows an unprecedented identification of well-separated behavioral clusters, which emphasizes that an adequate combination of factors leads to more reliable modeling of polarization behaviors (**RQ2**).

In a process of opening the filter bubbles, and reducing the polarization phenomenon, such a multi-factorial analysis could be greatly beneficial. In a strongly polarized context, within which users have formed strong opinions, an input of diverse items does not always have the desired effect. In fact, providing diversity can be tricky due to the strong opinions held by users, who are potentially very wary of being exposed to contrary ideas. It may even reinforce the polarization phenomenon [2]. An accurate characterization of adopted polarization behaviors could help to adapt solutions, including through recommendations. Each identified behavioral class could benefit from different recommendation strategies. To go further, the intermediate classes of users identified through the multi-factorial analysis could help to gradually expose highly polarized users to different viewpoints. This could limit the potential adverse effects of diversity, and help build trust-based recommendations. Especially, users whose membership to one unique cluster is uncertain could serve as bridges between the different classes close to them. The future development of depolarizing recommender systems could probably rely on a multi-factorial analysis of polarization to limit this worrying phenomenon.

## References

[1] Marina Azzimonti and Marcos Fernandes. Social media networks, fake news, and polarization. *European Journal of Political Economy*, page 102256, 2022.

[2] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[3] Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A Beam, Marc P Hauer, Lucien Heitz, Pascal Jürgens, et al. Diversity in news recommendations. *arXiv preprint arXiv:2005.09495*, 2020.

[4] Jonathan Stray. Designing recommender systems to depolarize. *First Monday*, 27, 2021.

[5] Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, and Armelle Brun. Being diverse is not enough: Rethinking diversity evaluation to meet challenges of news recommender systems. In *Fairness in User Modeling, Adaptation and Personalization (FairUMAP 2022)*, 2022.

[6] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207, 2018.

[7] Gabriel Machado Lunardi, Guilherme Medeiros Machado, Vinicius Maran, and José Palazzo M. de Oliveira. A metric for filter bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing*, 97:106771, 2020.

[8] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism*, pages 1–21, 2022.

[9] Mahsa Badami. *Peeking into the other half of the glass: handling polarization in recommender systems*. PhD thesis, University of Louisville, 2017.

[10] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, 2018.

[11] Emily Kubin and Christian von Sikorski. The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206, 2021.

[12] Jay J. Van Bavel, Steve Rathje, Elizabeth Harris, Claire Robertson, and Anni Sternisko. How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11):913–916, 2021.

[13] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.

[14] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[15] 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.

[16] Alfredo Jose Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114, 2015.

[17] Daniel Geschke, Jan Lorenz, and Peter Holtz. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1):129–149, 2019.

[18] Carolina Becatti, Guido Caldarelli, Renaud Lambiotte, and Fabio Saracco. Extracting significant signal of news consumption from social networks: the case of twitter in italian political elections. *Palgrave Communications*, 5(1):1–16, 2019.

[19] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. Polarization of the vaccination debate on facebook. *Vaccine*, 36(25):3606–3612, 2018.

[20] Tomas Cicchini, Sofia Morena Del Pozo, Enzo Tagliazucchi, and Pablo Balenzuela. News sharing on twitter reveals emergent fragmentation of media agenda and persistent polarization. *EPJ Data Science*, 11(1):48, 2022.

[21] Tinggui Chen, Qianqian Li, Jianjun Yang, Guodong Cong, and Gongfa Li. Modeling of the public opinion polarization process with the considerations of individual heterogeneity and dynamic conformity. *Mathematics*, 7(10):917, 2019.

[22] Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999.

[23] Simonetta Primario, Dario Borrelli, Luca Iandoli, Giuseppe Zollo, and Carlo Lipizzi. Measuring polarization in twitter enabled in online political conversation: The case of 2016 us presidential election. In *2017 IEEE international conference on information reuse and integration (IRI)*, pages 607–613. IEEE, 2017.

[24] Nicolas Turenne. The rumour spectrum. *PloS one*, 13(1):e0189080, 2018.

[25] Yanzhu Guo, Virgile Rennard, Christos Xypolopoulos, and Michalis Vazirgiannis. BERTweetFR : Domain adaptation of pre-trained language models for French tweets. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 445–450, Online, November 2021. Association for Computational Linguistics.

[26] Federico Germani and Nikola Biller-Andorno. The anti-vaccination infodemic on social media: A behavioral analysis. *PloS one*, 16(3):e0247642, 2021.

[27] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[28] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, pages 451–461, 2003.

[29] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[30] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, pages 53–65, 1987.

[31] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.