Transformer-Based Quantification of the Echo Chamber Effect in Online Communities

VAHID GHAFOURI, IMDEA Networks Institute, Spain and Universidad Carlos III de Madrid, Spain
FAISAL ALATAWI, Arizona State University, USA
MANSOOREH KARAMI, Arizona State University, USA
JOSE SUCH, King's College London, UK and VRAIN, Universitat Politecnica de Valencia, Spain
GUILLERMO SUAREZ-TANGIL, IMDEA Networks Institute, Spain

An Echo Chamber on social media refers to the environment where like-minded people hear the echo of each others' voices, opinions, or beliefs, which reinforce their own. Echo Chambers can turn social media platforms into collaborative venues that polarize and radicalize users rather than broadening their exposure to diverse information. Having a quantified metric for measuring the Echo Chamber effect can aid moderators and policymakers in tracking and mitigating online polarization and radicalization. Existing methods for Echo Chamber detection are either one-dimensional, only considering the network behavior of users while ignoring their semantic behavior, or require demanding supervised labeling, which is both expensive and less generalizable.

This paper proposes a new metric to quantify the Echo Chamber effect using Transformer models for context-sensitive processing of natural language (NLP). Our metric quantifies (1) the effect of an Echo Chamber through the inverse effect of *user diversity*, and (2) polarization by means of *user separability* between two Echo Chambers in a topic. Leveraging this metric, we further propose an NLP-based embedding that represents the users' activity. Our model is simultaneously effective, computationally cheap, and unsupervised. As our method is unsupervised, it makes existing collaborative moderation efforts to thwart Echo Chamber effects more efficient by addressing the problem of identifying narrow information bases for algorithmic biases and misinformation detection. We run our analysis on three recent highly controversial political topics and a non-controversial topic: Russo-Ukrainian War, Abortion, Gun-Control, and SXSW music festival. Our results offer data-driven findings such as a higher Echo Chamber effect among Republicans over Democrats and diverse explicit support for Ukraine, especially among Democrats. We also observe a direct relationship between the Echo Chamber effect and polarization while observing that the low Echo Chamber effect for the Russo-Ukraine war is accompanied by a low polarization; and vice versa for Gun-Control.

CCS Concepts: • **Information systems** → **Social networks**.

Additional Key Words and Phrases: echo chambers, polarization, social networks, NLP, sentence transformers

ACM Reference Format:
Vahid Ghafouri, Faisal Alatawi, Mansooreh Karami, Jose Such, and Guillermo Suarez-Tangil. 2024. Transformer-Based Quantification of the Echo Chamber Effect in Online Communities. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 467 (November 2024), 27 pages. https://doi.org/10.1145/3687006

Authors' Contact Information: Vahid Ghafouri, vahid.ghafouri@imdea.org, IMDEA Networks Institute, Leganés, Madrid, Spain and Universidad Carlos III de Madrid, Leganés, Madrid, Spain; Faisal Alatawi, faalataw@asu.edu, Arizona State University, Tempe, Arizona, USA; Mansooreh Karami, mkarami@asu.com, Arizona State University, Tempe, Arizona, USA; Jose Such, jose.such@kcl.ac.uk, King's College London, London, UK and VRAIN, Universitat Politecnica de Valencia, Valencia, Ohio, Spain; Guillermo Suarez-Tangil, guillermo.suarez-tangil@imdea.org, IMDEA Networks Institute, Leganés, Madrid, Spain.

ACM 2573-0142/2024/11-ART467
https://doi.org/10.1145/3687006

1 Introduction

Online Echo Chambers are both the cause and the effect of the polarized political environment existing across the globe. An Echo Chamber could be thought of as an environment where ideas are reinforced by repeated interactions between users with similar tendencies and attitudes [18, 46].

Social media platforms are fertile grounds for these polarizing repeated interactions that lead to the formation of Echo Chambers [21]. In addition, users are often exposed only to the content they agree with due to social media over-personalization [9, 61], further confirming their existing beliefs — see confirmation bias [54], and shielding them from exposure to the other side of the argument — see selective exposure [43].

One of the key drivers of Echo Chambers on social media platforms is the interplay between algorithmic-driven and human-driven curation of content [35]. While algorithms play a significant role in shaping the content that users see, human curation through sharing and reposting also amplifies certain viewpoints and reinforces existing beliefs. This dynamic can create a self-reinforcing cycle that further entrenches users in their own Echo Chambers. As a result, it is important to understand the mechanisms that contribute to the formation of Echo Chambers and to develop strategies to promote a more diverse and inclusive online discourse.

Echo Chambers stifle the free flow of ideas, hindering the exchange of diverse perspectives and the formation of well-rounded opinions [67]. By limiting exposure to opposing viewpoints, Echo Chambers foster a climate of intolerance and prejudice, where individuals become increasingly entrenched in their own beliefs and less receptive to alternative views [13]. This intellectual insularity can lead to a decline in critical thinking skills and a diminished capacity to engage in constructive dialogue.

Moreover, Echo Chambers amplify the spread of misinformation, posing a significant threat to public discourse and decision-making. In these self-reinforcing environments, false or misleading information can gain traction and go unchecked [24, 40, 63, 64, 68, 69], potentially influencing individuals' actions and behaviors in detrimental ways. The proliferation of misinformation in Echo Chambers can undermine trust in institutions, erode public confidence in democratic processes, and exacerbate social and political tensions. The COVID-19 pandemic has been one of the recent critical cases in which society had been affected by Echo Chambers driven public mistrust in the vaccination and precaution mechanism propagated by governments and the mainstream media [72].

In the quantitative domain, the study of Echo Chambers and political polarization has gained significant attention within the field of computer-supported cooperative work [12, 48, 62] as researchers strive to understand the societal impact of online collaboration and information sharing. Cooperative work provides a unique lens through which to analyze the dynamics of Echo Chambers, as it explores how individuals interact, collaborate, and engage with computer systems and technologies in social and political contexts. By leveraging computational methods and social network analysis, we can uncover patterns of online collaboration, information diffusion, and the formation of ideological clusters.

In this study, we employ an unsupervised approach to estimate the Echo Chamber effect. Echo Chamber effects are overly dynamic. Thus, using manually labeled data to measure polarization and Echo Chambers limits considerably the generalizability of the study. Labeling efforts include identifying seed accounts (e.g., influencing politicians, users, or news channels) [10] or establishing predefined sets of domain-specific polarized hashtags and keywords [2, 25, 57]. On the contrary, unsupervised methods are more scalable, as they do not require manual data labeling, which can be time-consuming and resource-intensive. Our unsupervised approach allows for increased scalability and flexibility in analyzing the Echo Chamber effect, and by not relying on manually labeled data, we assist and reduce the need for collaborative efforts in crowd-sourcing data annotations.

Our first computational step is to detect Chambers — communities — for every topic based on the retweet network clusters. Then, we select a random sample of users from each Chamber and embed the users into a vector space by averaging the sentence transformer embeddings of their tweets. We use the diversity of user embeddings in every Chamber to measure its Echo and the separability of two Chambers' users to estimate polarization across Chambers.

In Section 3, we break down the concept of Echo Chamber and define "Echo", "Chamber", "Echo Chamber", and "Polarization" aligned with our computational model. In Section 4, we show how we embed users using sentence encoders and quantify "Echo" per "Chamber" and "Polarization" across "Chambers". In Section 7, we apply our method to three recent controversial topics and a non-controversial topic: "war on Ukraine", "Abortion Ban", "Ulvade school Gun Shootings", and "SXSW music festival". We compare the level of "Echo" per "Chamber" and "Polarization" across "Chambers" for each topic. In summary, we make the following observations:

- The diversity of users in Republican Chambers is lower than in Democratic Chambers. We interpret this as a higher Echo Chamber effect in Republican stances, which is consistent with previous literature [10].
- The diversity of pro-Ukraine users is higher than in the other controversial case studies. In addition, Ukraine-related Chambers, as a case of foreign national conflict, has caused the least polarization in comparison to the other topics. However, we also observe that the most explicit supporters of Ukraine seem to be Democrats.
- The use of mean-pooling in sentence-transformer encodings to generate user embeddings is fast and effective for distinguishing users based on their political stances. This has useful implications for future work leveraging user classification tasks.

We address the challenge of modeling Echo Chambers through the combination of cutting-edge methods in different disciplines, including the use of sentence transformers, network analysis, and social sciences. By integrating these approaches, we bridge the gap between computational techniques and social science theories to gain a comprehensive understanding of Echo Chambers as collaborative phenomena. We hope to contribute to the aim of designing technologies and interventions that support effective collaboration in various domains (e.g., political discourse analysis, gender studies, etc.)

2 Related Work

In this section, we will initially discuss the social implications of Echo Chambers and how they can cause online harm according to the social science literature. Then, we discuss previous quantitative methods of Echo Chamber detection. We also allocate a separate section to previous methods of embedding users as it is a key element in our method of quantifying online Echo Chambers and polarization.

2.1 Echo Chamber and Social Harms

Research has consistently demonstrated the negative impacts of Echo Chambers on online communities and society. For instance, a study by Colleoni et al. [21] found that users who were exposed to ideologically homogeneous information on Twitter were more likely to exhibit polarized attitudes. Similarly, Bakshy et al. [9] demonstrated that social media algorithms can exacerbate polarization by recommending content that aligns with users' existing beliefs.

The proliferation of misinformation in echo chambers has also been documented by a multitude of studies. Del Vicario et al. [24] found that Echo Chambers on Twitter played a significant role in the spread of misinformation about the 2016 US presidential election. Similarly, Shu et al. [64]

demonstrated that the consumption of misinformation in Echo Chambers can lead to decreased trust in mainstream media and increased belief in conspiracy theories.

The harmful effects of Echo Chambers extend beyond the realms of political polarization and misinformation. A study by Cinelli et al. [18] found that Echo Chambers on YouTube can lead to increased prejudice and discrimination against minority groups. Similarly, Jiang et al. [40] demonstrated that Echo Chambers on social media can contribute to social unrest and violence.

In conclusion, previous research underscores the substantial threat posed by Echo Chambers to the health and well-being of online communities and society at large. Recognizing this, the development of effective tools for detecting online Echo Chambers becomes paramount in fostering healthier and more inclusive digital discourse.

2.2 Echo Chamber Detection

We could split Echo Chamber detection methods into three types: network-based [22],

content-based [16], and hybrid detection methods [70]. The network-based methods utilize well-known community detection algorithms to detect communities in interaction graphs built using social media interactions such as retweets and replies. The content-based methods [44] cluster users based on the content they use by extracting features such as the sentiment about a topic or embedding of content. Finally, the hybrid approach [40, 50] incorporates the knowledge from both content and topology to find Echo Chambers.

In this paper, we utilize the network feature to detect communities (Chambers) as it is the most common method to detect Echo Chambers. Moreover, network-based methods were used in related work on measuring polarization [31]. Then, we use the content generated by users to measure the Echo Chamber effect to verify if the detected communities are indeed Echo Chambers.

2.3 User-level Embeddings

User-level embeddings are used to model the behavior of the users for various tasks. Recent common methods utilize neural encoders to encode the user behavioral data (e.g., recent tweets on social media or recent queries for search engines) into low-dimensional vectors. These approaches reduced the amount of feature engineering and manual feature extraction labor by automating the relations between the user's own data as well as its relation to other users' data. User-specific data on social media can be divided into four different categories: (i) user's profile information, (ii) user's activity, (iii) user's network connectivity, and (iv) user's generated content. In the behavioral analysis of the users on social media, researchers utilized different conjunctions of the aforementioned categories for creating task-specific as well as universal user representations [37].

Most of the user embedding research models the user's behavior through their generated content by utilizing models that optimize the conditional probability of the texts, given their authors. These aggregated texts per user can be modeled using different methods such as Latent Dirichlet Allocation (LDA) [57], Convolutional Neural Network (CNN) [3], Matrix Representations [41], and Word-Embeddings [2, 25, 57]. Moreover, the network connectivity of the users is also common in modeling the users' attributes. These methods try to map the social networks into low-dimensional representations such that the local and global topological structures are preserved [55]. Community detection algorithms and Graph Neural Network models are among the common methods used to model social networks such as "friendship", "retweet", and "endorsement" social graphs [26, 73]. Auxiliary information such as profile information would also help in modeling the user behavior and improving the methods [41, 42, 74].

However, all the user-level embedding methods for Echo Chamber detection rely on a labeled and cherry-picked set of ground-truth political users, keywords, and hashtags. This would make them less robust, more demanding for manual effort, and less generalizable to later social network analysis tasks since supervised methods are vulnerable to concept drift [52]. In other words, as time passes, seed political celebrities, political hashtags, and the use of language will change.

In Section 4.2, we explain how we propose an unsupervised, computationally cheap, and effective way of embedding users based on sentence transformers to tackle the mentioned short-come.

3 Terminology

The terms "Echo Chamber" and "Filter Bubble" are often used interchangeably in the literature [14, 15] while sometimes being integrated with the concept of "Polarization" [59]. Although there is a common core idea underlying these terms, it is hard to find prior work that makes a unique, universally settled definition for each of the terms. Therefore, in this section, we explicitly state the definitions we consider most relevant to our study from previous literature.

Chamber is a discussion forum where interactions occur and users share content or ideas. In our work, a *Chamber* equates to an Internet forum, where users post messages to other members of that forum. On Twitter, we represent a Chamber as a cluster of users linked by interactions (i.e., retweets, quotes, mentions, and replies) on a topic. Our rationale is that these clusters represent a network where users interested in a specific topic get exposed to a particular discussion on Twitter. This definition is derived from Garimella et al. work, where they establish that a Chamber is "the social network around the user, which allows the opinion to echo back to the user, as it is also shared by others" [30].

Echo is the level of homogeneity among the members of a discussion in a Chamber. It is a common notion in the literature that online Echo Chambers happen in environments with homogeneous sets of users [28, 66]. This homogeneity can stem from similarities in users' political leaning (e.g., traditional left or right), socio-economic statuses, or demographic features (like age or gender) [36].

Echo Chamber in our terminology is a "Chamber" with high levels of "Echo". In our domain this is a retweet network with low user diversity (high homogeneity). For instance, if all the members of an anti-abortion Chamber are from the right wing in political opinion, we call that Chamber an "Echo Chamber" where like-minded people hear the echo of their own voice [10].

Polarization is the extent to which the members of a Chamber formed around a topic can be separated/distinguished from the members of its opposing Chamber on the same topic. Similar to Garimella et al. [29], we take into account the Oxford Dictionary definition of Polarization as "the act of separating or making people separate into two groups with completely opposite opinions." Let's take the case of abortion as a running example. If we observe that only hard-core left-leaning users attend Chamber A (which can presumably be the place where pro-abortion opinions are being shared) and only hard-core right-leaning users attend Chamber B (which instead can presumably be the place where anti-abortion content is being shared), we would say that the topic "abortion" is polarized between Chambers A and B based on political leaning. However, if both the pro-abortion Chamber and anti-abortion Chamber embrace diverse users from all parts of the political/demographic/economic/gender spectrum, in a way that a pro-abortion user is hardly distinguishable from an anti-abortion one by an explicit factor, our definition would label the abortion topic as less polarized.

Our definition of polarization is also aligned with Esteban and Ray [27]. Similarly, we also argue that polarization can theoretically happen by gender (i.e., mostly men opposing abortion rights and mostly women supporting it), age, location, political leaning, and any other features from users that can be automatically stored in our black-box user embedding approach which we explain in Section 4.2. This multi-dimensionality of polarization in our method is particularly useful in environments where polarization extends beyond the traditional left-right divide; a division that is primarily defined for the US as an effect of the cold war [6]. For instance, in Taiwan, polarization centers around attitudes towards having closer ties with the US versus having closer ties with

China [71], while in Western Asian countries such as Iran and Turkiye, the degree of desired secularism forms the primary axis of division [5, 7].

4 Methodology

Our method returns two main measures, the Echo of every Chamber and the Polarization across Chambers. Our first step is to detect the top important Chambers, for which we use the retweet network of a set of controversial topics. Our second step makes a per-user analysis by looking at the type of content posted by the users of the detected Chamber to embed their general stance. The final step is to utilize the user-embeddings to estimate the homogeneity of users (Echo) per Chamber and their polarization across Chambers.

Figure 1 shows an overview of our computational architecture.

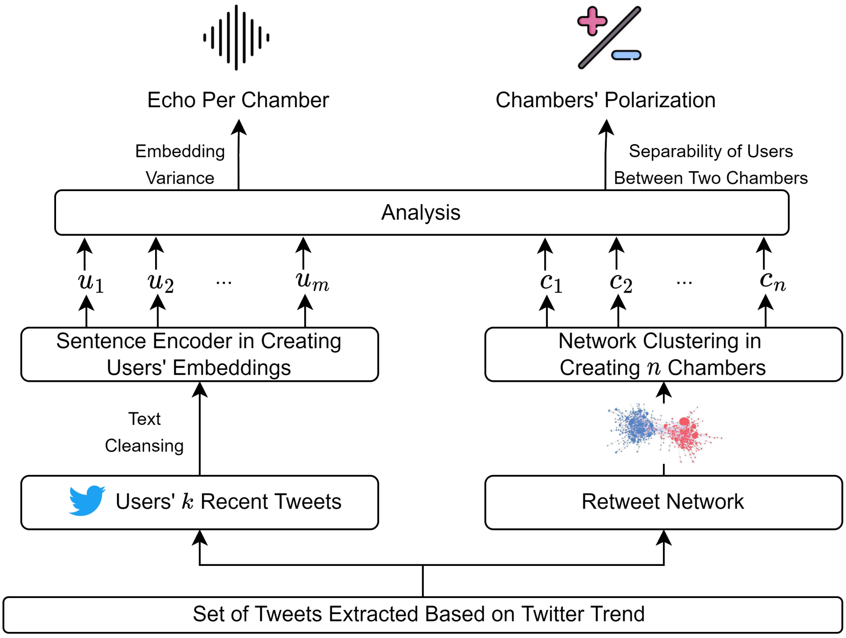
Echo Per Chamber
Chambers' Polarization
Embedding Variance
Separability of Users Between Two Chambers
Analysis
u_1
u_2
...
u_m
c_1
c_2
...
c_n
Sentence Encoder in Creating Users' Embeddings
Network Clustering in Creating n Chambers
Text Cleansing
Users' k Recent Tweets
Retweet Network
Set of Tweets Extracted Based on Twitter Trend

Fig. 1. Scheme of our method's architecture.

4.1 Detecting Chambers (Network Clusters)

Our initial step is to identify Chambers.

Our method departs from a large set of trending tweets around controversial topics. Our analysis focuses on three topics *abortion*, *gun control*, and the *Ukraine war* selected for being either well-established controversial topics (i.e., abortion and gun control) or recently established topics (i.e., the Ukraine war). We also add SXSW 2022 music festival a commonly analyzed case of a non-controversial topic [20]. However, our methodology is generic and can be applied to any other topic.

Overall, we collect the retweet network of $\approx 20k$ users for each of the topics using relevant keywords explained in Section 6.

We then create a retweet graph per topic in which the nodes represent the users, and a link between two nodes A and B represents that user A retweeted user B. Then, we use the Louvain algorithm [11] over the retweeted tweets to unfold communities into clusters. Louvain is known to work well with polarized communities [18, 23].

It is common for the retweet networks of controversial topics that the two largest network clusters represent the main sides of the debate. To verify this, we ran a cursory inspection that proved most of the tweets were aligned with the partisan stances of the entire Chambers. We label the Chambers' stances as "Democratic" or "Republican" based on the stances of tweets we observe in each Chamber.

It is worth noting that this only labels the political stance of the "content" in each Chamber which is presumably either pro or against the debated topic, not the "general ideology" of the "users" inside those Chambers. One of our main objectives is to check user diversity inside each Chamber. Therefore, we expect a significant amount of moderate or non-political users to appear in each of the partisan Chambers.

4.2 Embedding Users

The next step in our analysis is to characterize Twitter users' ideology according to their produced content. We start by extracting the features for the 200 tweets that have recently been generated by a user. After preprocessing the tweets' text (removing mentions, URLs, etc.), we represent them using a vector of embeddings. We use the state-of-the-art[1] pretrained sentence transformer model (all-mpnet-base-v2)[2] from *Hugging-Face*.[3] The model is fine-tuned to map sentences and short paragraphs to a 768-dimensional dense vector space in a way that preserves semantic features of the text so that the vectors can be utilized for tasks such as clustering or semantic search. Then, we represent users through the average pooling of his/her tweets' embedding vector.

In our methodology for user representation, we deliberately opted for state-of-the-art pretrained sentence transformer models like all-mpnet-base-v2 due to their adeptness in capturing semantic essence from individual tweets efficiently. Unlike LSTM models applied to concatenated tweets, which assume continuity in text sequences and might struggle with discrete, independent tweets, sentence transformers excel in encoding short texts without imposing such assumptions. Their transformer architecture enables effective capture of semantic relationships within tweets, aligning with our goal to represent users based on their varied and discrete tweet content. Specifically choosing the all-mpnet-base-v2 model was driven by its balance between performance and computational efficiency, ensuring effective mapping of tweets into a 768-dimensional vector space while preserving semantic features crucial for downstream tasks like clustering and semantic search, thereby enabling a robust user representation based on tweet content. Moreover, all-mpnet-base-v2 is open-source and downloadable for offline use. When it comes to large-scale use, this makes it a more practical option than the recently developed advanced LLMs that require paid plans for using their APIs at limited rates.

4.3 Quantifying Echo

We quantify the Echo of every Chamber by the inverted effect of the variance among user-embeddings of all members in a Chamber:

$$echo = \frac{1}{\widehat{\mathrm{Var}}(U)}. \tag{1}$$

[1] https://www.sbert.net/docs/pretrained_models.html
[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[3] https://huggingface.co/

This quantification captures the level of homogeneity among the members of a Chamber, which is aligned with the definition of "Echo" in Section 3. Thus, a lower variance of users indicates a higher "Echo".

We compute the variances across 768-dimensional vectors representing user embeddings. This involves assessing the variability present in each dimension of the user embeddings, capturing the multidimensional nature of the data. Specifically, we calculate the echo by averaging the variances observed across all elements within these vectors. This comprehensive approach ensures that the echo metric accurately reflects the level of homogeneity or consistency among users across all dimensions represented in the data space.

4.4 Quantifying Polarization

In addition to the variety of users in every Chamber, we are interested in quantifying the polarization of users across pairs of selected Chambers formed on a topic. We begin by measuring the level of linear separability among user embeddings of pairs of Chambers. To this end, we train a linear SVM classifier with the user embeddings (cf. Section 4.2) as features and the Chamber that the users belong to as the labels. We also apply a similar pipeline with hashtags as labels.

Note that our goal differs from the classical usage of a prediction task and we do not aim at classifying users based on the Chamber they belong to. Instead, we intend to deduce which pair of Chambers have the highest level of separation among their users judging by the performance of multiple pairwise classification tasks. Thus, it is critical to have a consistent set of elements for all classification experiments, including the parameters and sample size. Therefore, we take equal random samples of users (1,500) per Chambers/hashtag, and split one half to train and the other half to test the model. We take the accuracy of the test set as the final indicator of linear separability among users.

We chose a Linear SVM due to its inherent use of hyperplanes to split data points. Our rationale is that stances are in a continuous spectrum. For instance, when it comes to political leanings, a user can stand in the alt-left, the alt-right, or somewhere in between. Therefore we expect a line/hyperplane to be able to clearly split users based on this spectrum in cases of strong polarization. The accuracy of the SVM classifier would indicate the separability of the users.

In addition to reporting classification accuracy, we also report the weighted average of the model's confidence for each data point in the classification. This supplementary metric is to take into account the difference between pairs of points that are closer to the separating hyperplane (less polarized) and those that are farther from the hyperplane (more polarized). The confidence score provided for each data point indicates how far the data point is from the SVM decision boundary.

Then, the weighted average of confidence scores is computed as in Equation 2 while setting weights to 1 for correct predictions and −1 for incorrect ones.

$$\text{Average Confidence} = \frac{\sum_{i=1}^{n} \text{confidence}_i \cdot \text{weights}_i}{\sum_{i=1}^{n} |\text{weights}_i|} \qquad \text{weights} = \begin{cases} 1 & \text{if } \hat{y} = y \\ -1 & \text{if } \hat{y} \neq y \end{cases} \tag{2}$$

5 Evaluation

We evaluate our metric on a dataset of tweets from congresspeople[4] and senators labeled as Republican or Democrat. The users in this analysis are the ground-truth for a set of users who are separated by their political views. Our evaluation measures our model's capability to separate them.

[4] Obtained from: github.com/alexlitel/congresstweets

We sample 200 tweets per user and embed them by the average of their tweets' embeddings as introduced in Section 4.2. We use UMAP [49] to visualize the 768-dimensional user embeddings into 2D space. UMAP is one of the state-of-the-art dimensionality reduction algorithms at the time of writing [34]. Figure 2 shows the political affiliations color-coded. We see that most points are well-separable by a linear hyperplane. In higher dimensions (e.g., the original 768D vectors), where we have more features, separation becomes even easier due to the increased dimensionality of the data space. Therefore, an n-dimensional hyperplane can yield similar or more separable results than the 2D data points in Figure 2. This is due to the fact that the additional features provide more discriminative power, enabling better separation of data points in the higher-dimensional space.

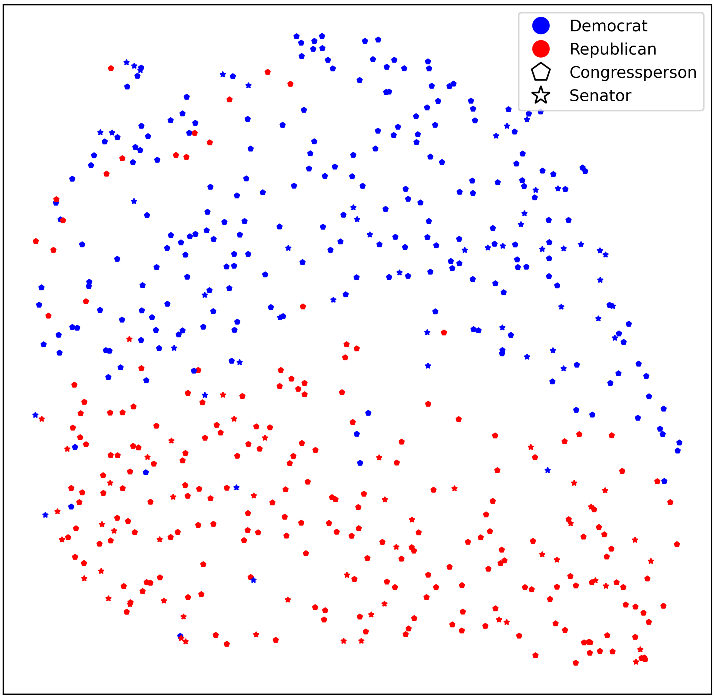

Fig. 2. 2D projection of US congresspeople and senators' user-embeddings.

To quantify and validate this separability, we train a linear SVM classifier on half of the data and validate on the other half as our test-set. The classifier yields a 93% F1-Score (macro), suggesting that a promising set of features are stored in our user embedding vectors and that our method can be used to distinguish the political stances of users. Note that this performance is given after using only 130 users per class (Republican vs. Democrat) and 200 tweets per user, which offers a promising measure for scarce datasets. We examine the performance of our method when we increase the number of tweets per user to 500, obtaining an improvement of the F1-Score up to 95%. We, however, stick to 200 tweets per user due to constraints in our Twitter API rate limit.

As we deal with pairs of Chambers that are formed on the basis of a Republican vs Democratic leaning idea over a topic, the user separability we measure across these Chambers is mapped to the level of political polarization across Chambers of a topic. We further discuss the scope of our evaluation in Section 8.

As per the performance, the whole process of collecting 200 tweets from a user, transforming them into vectors, and averaging all the vectors, took approximately 3 seconds per user on Google-Colab's GPU.

6 Datasets

We consider top trends on Twitter associated with three recent controversial events next to a non-controversial one: (1) the Uvalde school shooting which triggered yet another discussion around gun control; (2) the US Supreme Court's decision on June 2022 to overturn Roe v. Wade sparked a nationwide debate on abortion rights in the US [5]; (3) the Russo-Ukraine War; and (4) the SXSW 2022 music festival.

Our data is collected over one month period since the events related to the topics. We utilize the "Network Tool" [6] developed by Indiana University Observatory On Social Media to query top trending hashtags related to the topics on Twitter. Table 1 shows the list of hashtags and dates that we used for collecting retweets for every topic.

Topic	Queried Keywords/Hashtags	Start Date	End Date	# of Users
Abortion-ban	Abortion, #RoeVsWade, #Pro-life, #Prochoice, #WhatIsAbortion, #MyBodyMyChoice #AbortionIsHealthCare, #AbortionIsMurder	1/6/2022	30/6/2022	≈ 29000
War on Ukraine	Ukraine, #StandWithUkraine-(the latter was used only for Section 7.1)	20/2/2022	20/3/2022	≈ 21000
Texas Gunshooting	Gun, Ulvade, Shooting, #GunControl, #GunOwnersForSafety, #ProGun, #AntiGun, #GunRights, #GunViolence, #MassShooting, #2ndAmendment, #RighttoCarry, #EndGunViolence	24/5/2022	23/6/2022	≈ 25000
SXSW Festival	#SXSW	1/3/2022	30/3/2022	≈ 11000

Table 1. Queried hashtags for data collection.

Next to the basic keywords of the topics we used for querying (e.g. "abortion" for the Abortion topic), we tried to maintain equal numbers of partisan hashtags for both sides of the debates on every topic. We sorted trending hashtags per topic based on their popularity and picked as many neutral hashtags as existed in the trends (e.g. #RoeVsWade has no clear partisan position on its own) and an equal number of partisan hashtags from both sides down-sampled to the less populated side. For example, if a topic has 3 right-wing and 10 left-wing partisan hashtags, we pick all the 3 right-wing hashtags and 3 top most trendy left-wing ones. However, for the case of "War on Ukraine", despite multiple pro-Ukraine hashtags, we were unable to find any pro-Russian invasion hashtag in the English Twitter, thus, we only used tweets that contained the word "Ukraine" for forming the retweet network. In this way, we represent both sides of the debate, if any, fairly on

[5] https://reproductiverights.org/global-trends-abortion-rights-infographic/

[6] https://osome.iu.edu/tools/networks/

the retweet network. Also, for the case of the SXSW, there was no notion of right-wing or left-wing hashtags since it is not a politically polarized topic, so we only queried the keyword "SXSW".

Later on in Section 7, we select subsets of the users of these keywords, based on the partisan hashtags they used (cf. Section 7.1) or the retweet network (Chamber) they appeared in (cf. Section 7.2), and collect the latest 200 tweets of their timeline using Twitter's official API.

7 Experiments and Results

We next run two separate experiments. First, we analyze the level of Echo per hashtag and hashtag-wise Polarization by characterizing the users who have used any of those hashtags. Then, we measure the Echo of every two Chambers for all topics and their Polarization.

7.1 Echo per Hashtag

On most social media platforms, including Twitter, clicking on a hashtag fills the timeline of the user with top-tending tweets around the hashtag. Thus, a hashtag offers a specific environment of content. Therefore disregarding the position of users in the retweet networks, we only look into partisan hashtags (i.e., hashtags with clear political stances) to measure the diversity and polarization of users across the hashtags.

For this, we gather a sample of users who have used pro-gun hashtags (e.g., #GunRights), anti-gun (e.g., #EndGunViolence), pro-abortion (e.g., #AbortionIsHealthCare), anti-abortion (e.g., #AbortionIsMurder), and pro-Ukraine (e.g., #StandWithUkraine) — i.e., there is no explicit *anti-Ukraine* hashtag on Twitter to be added to the analysis. We also add one case of a non-partisan hashtag, namely #SXSW, for comparison.

We obtain a novel embedding of each of the users in an unsupervised fashion following the step in Section 4.2.

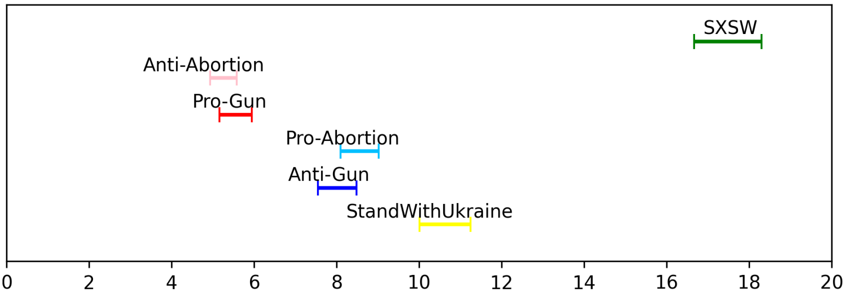

Fig. 3. Variances of user embeddings for partisan hashtags' users + #SXSW as a non-partisan case

Figure 4 shows the 2D projection of user embeddings color-coded by the type of hashtags they have used. We see that the Republican stances discussing Pro-Gun and Anti-Abortion (red and pink) stem from users that are more densely embedded in the spectrum. These users have a high overlap with each other. Instead, the Democratic stances discussing Anti-Gun and Pro-Abortion (blue and light-blue) are represented by a more diverse set of users on Twitter. The users of #StandWithUkraine hashtag are also widely distributed in the plot with higher overlap with Democratic users than the Republicans. These results provide an initial intuition about the variety and overlap of users who had supported specific political stances, yet we are interested in quantifying these concepts statistically.

To quantify variety, we use a multidimensional variance of the user embeddings per hashtag portrayed. These variances are calculated by taking the mean of all element-wise variances for

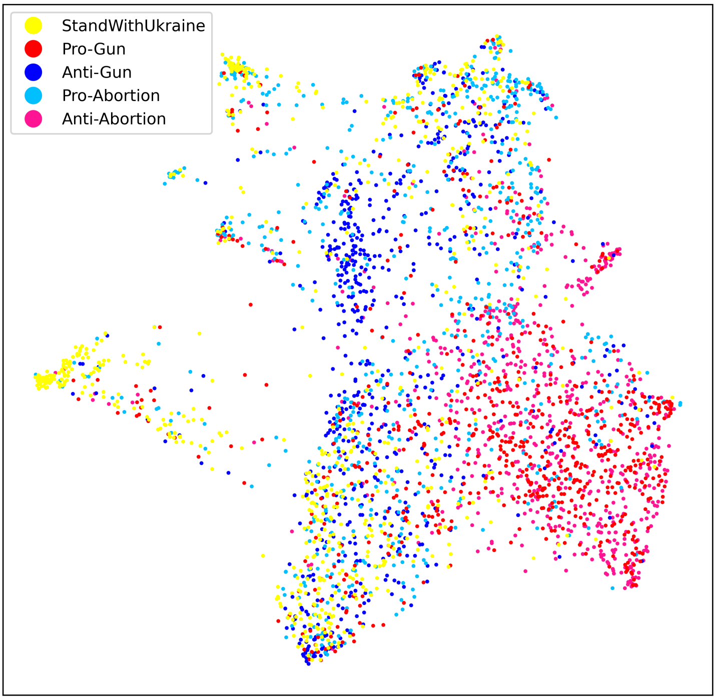

Fig. 4. 2D projection of user-embeddings for polarized hashtags' users.

a multidimensional set of vectors. The 95% confidence intervals are calculated based on 1,000 bootstraps each containing random 1,500 samples. Figure 3 shows that the users of the Republican-leaning hashtags have significantly lower diversity than the Democratic hashtags' users. The users of #StandWithUkraine hashtag preserve the highest diversity, showing a possibly vast demographic support among the users.

Finally, we quantify the polarization according to the ability of a Linear SVM to separate users of two classes (hashtags). Table 2 shows the F1-Score per hashtags class. Recall that a low F1-Score means a high rate of overlap between the users of two hashtag classes as discussed in Section 4.4. We see that the Democratic and Republican hashtags have lower separability among themselves and higher separability across hashtags supported by the other party. For instance, the separability of pro-abortion vs anti-gun is low (70%) in two democratic Chambers. At the same time, there is a high (91%) separability between anti-abortion and anti-gun as the members of a Republican stance are presumed to be separable from a Democratic one.

Table 2. F1-Scores for linear separability between pairs of user embeddings across hashtags.

Hashtag Class	Pro-Ukraine	Pro-Gun	Anti Gun	Pro-Abortion	Anti-Abortion
Pro-Ukraine	50%	83%	77%	74%	90%
Pro-Gun	83%	50%	82%	77%	60%
Anti-Gun	77%	82%	50%	70%	91%
Pro-Abortion	74%	77%	70%	50%	87%
Anti-Abortion	90%	60%	91%	87%	50%

We also observe a higher separability of pro-Ukraine users with Republican supporters than when compared to Democrats, meaning that although the pro-Ukraine stance is more diversely supported, discussions are more popular among Democrats. Note that even the most partisan hashtags can have an underlying political agenda. Although this effect may influence the intuitiveness of the results, our method is good at quantifying these nuances.

7.2 Echo per Chamber

This section measures the Echo for every Chamber. In other words, we quantify the Polarization of the retweet clusters across topics.

Unlike in our experiment in Section 7.1 where we select users that use specific partisan hashtags, we retain here all users that appear in the retweet network cluster. This is done to compare user embeddings with the stances of the users on each of these topics. This comparison let us measure the Echo Chamber effect and Polarization.

First of all, we validate the network clustering step by manually labeling a random sample of 210 retweets for all network clusters. Each retweet network cluster in our dataset is composed of approximately 300 seed tweets, thus, our sample will look at around 12% of the entire seed tweets (6×300).

Although the homogeneity of the stance of each Chamber is visible from a cursory inspection, the purpose of this experiment is to systemically verify this. Table 3 shows the number of each tweet's stance per retweet network and the rate of alignment with the hypothesized stance of the entire Chamber in the first cursory glance. We see that each Chamber is formed around a certain stance toward a topic, as for every Chamber, the identifiable stances of tweets are almost entirely pro or anti. Unidentifiable tweets' stances include tweets with reference to broken links or quotations of news without expressing any explicit opinion about them.

Our annotation guideline is based on the main positions of each political party in the US on each of the controversial topics. Tweets with references such as "women's right to decide about their own body", "health-related risks of banning abortion", etc. are labeled as *Democratic* whereas those with references to "the right of the embryo to live", "religious teachings against abortion", etc. are labeled as tweets with *Republican* stances. Regarding the Ulvade school shootings, tweets emphasizing the significance of the tragedy with direct or indirect blame on the gun law in the US are labeled as *Democratic* and those referring to the "2nd Amendment rights to carry firearms" or arguing that "gun-rights is not the actual reason, but the solution" are labeled as *Republican* tweets. Tweets labeled as "Anti" Ukraine for the Republican Chamber in Table 3, are actually the combination of all the stances focusing on "Russian military advances", "claiming that US aid to Ukraine is excessive", "blaming the war on Biden administration's policies", "criticizing Zelenskyy", "complaining about the rate of Ukrainian refugee intake", etc. which are the alternative to the democratic stances focusing on "Ukrainian military advances", "asking for more US/NATO aids to

Ukraine", "empathizing with Ukrainian victims of war", etc. SXSW is not included in Table 3 as it is not a politically polarized topic to begin with.

Table 3. Stances of sampled tweets for each Chamber. The rate of alignment of tweets' stances with the hypothetical stance of a Chamber shows the accuracy of the network clustering method.

Topic	Chamber	Hypothetical Stance	Sample Size	N Pro	N Anti	Alignment	N Unidentifiable
Abortion	A	Pro-Abortion (Democrat)	35	32	1	97%	2
Abortion	B	Anti-Abortion (Republican)	35	1	34	97%	0
Gun	A	Anti-Gun (Democrat)	35	0	31	100%	4
Gun	B	Pro-Gun (Republican)	35	29	1	97%	5
Ukraine	A	Pro-Ukraine (Democrat)	35	21	0	100%	14
Ukraine	B	Anti-Ukraine (Republican)	35	2	25	93%	8
Overall			210	-	-	97.3%	-

We now look at the entire retweet network. Figure 5 shows the retweet network, visualized by Forced Atlas 2 [39], on the top and the user embeddings on the bottom. As the main communities within the SXSW retweet network lacked sufficient separability, given the non-controversial nature of the topic, the Forced Atlas 2 algorithm depicted it as a unified circular atlas. In contrast, the three controversial topics manifested as two distinct circles, showcasing their discernible independence.

User embeddings are projected into 2D using UMAP and color-coded based on the corresponding retweet network (Chamber) they have participated in. The more separable the blue and red data points are, the more polarized the Chambers are. Instead, in less polarized Chamber pairs, we expect the points to be more mixed with each other.

Moreover, if the "Echo" in a "Chamber" is high, we expect to observe a higher density in its users' embeddings' 2D projection with respect to the other color-coded Chamber. This means that a more homogeneous group of people have taken the stance supported by that retweet network.

After providing a visual intuition, we apply our method (steps in Sections 4.3 and 4.4) to quantify the Echo and the Polarization of Chambers. Table 4 summarizes the values for linear separability and variance of each Chamber.

In all three controversial topics, the Chambers of the Republican stance have lower variances (higher Echo) than their Democrat counterpart (column *Var* in Table 4). Among the three controversial topics, the Chambers of the gun-control topic have the lowest variance and the highest separability from each other in comparison to other topics, whereas the exact opposite has happened for the war in Ukraine. This not only shows a higher level of polarization for the gun-control discussion and a lower polarization for the Russo-Ukraine war but also a positive relationship between the level of Echo and the polarization in online discussions. As anticipated, the sole non-controversial topic, SXSW, exhibited the least polarization and the greatest user diversity, reinforcing the robustness of our methodology. However, even though it registers as comparatively low, the observed separability for SXSW is not negligible. This raises the possibility that a non-political

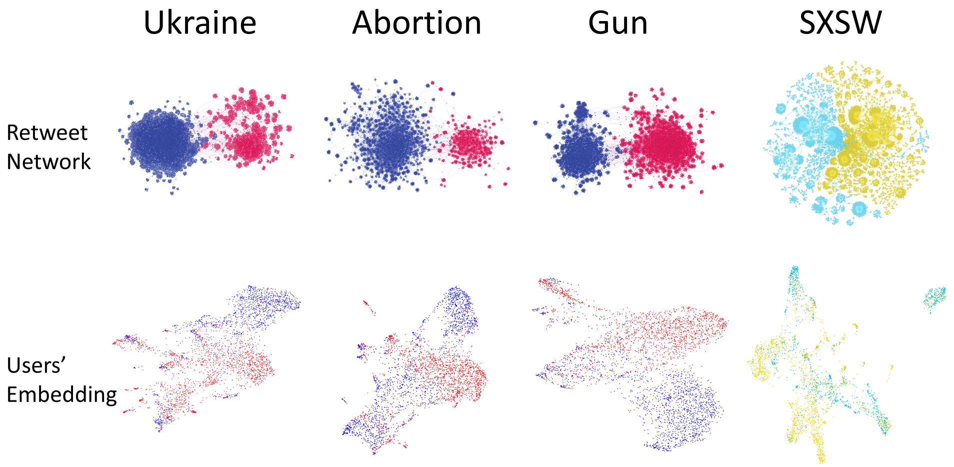

Fig. 5. Comparison of retweet networks vs 2D projection of user-embeddings. The red and blue points represent the users that had attended Conservative and Democrat Chambers in the corresponding events.

source of polarization could underlie the observed user separation. Further exploration of such instances is elaborated in Section 8.1.3.

Furthermore, Table 5 depicts the heat map of user separability between chambers across topics. As we fix A and B as the Democrat and Republican Chambers in all the topics, in case our user embedding method holds sufficiently meaningful features, our hypothesis would be to see a lower separability among the users of the same-letter Chambers (i.e., A vs A, B vs B) and higher separability among users of cross-letter Chambers (i.e., A vs B, B vs A). This hypothesis seems to hold, as the separability is 86-93% for all cross-letter Chambers while it falls to 69-80% when comparing two Chambers with similar letter codes. The minimum separability is 50%, which represents the accuracy of a classifier when the labels are random (i.e., in this case, identical: Abortion Chamber A vs Abortion Chamber A again).

For the Ukraine case, we observe a higher user separability for same-letter Chambers with the other two topics rather than Gun vs Abortion (e.g. Ukraine's Chamber B is more separable from Abortion's Chamber B – 80%, than Gun's Chamber B from Abortion's Chamber B – 69%). This further supports, as already discussed before, that the users in the Russo-Ukraine war case are more diverse and its Chambers are less likely to be divided into purely Democrat and purely Republican users.

Again, our goal is to *compare* the level of separability by comparing the performance of the classifier, not building a classifier to separate the users. However, a byproduct of this observation is to further approve the efficiency of our user embedding approach by the high accuracy obtained for separating the classes. Using our user embeddings as features, a simple linear classifier is not only able to classify Democrat vs Republican users (Section 5), but also cases like Pro-Abortion Democrats vs Anti-Gun Democrats. We find that our novel user-embedding approach has the potential to be used for future user-classification tasks.

7.3 Comparison with Supervised Baseline

This section aims at comparing our newly proposed method with existing baselines. Unfortunately, when it comes to the field of Echo Chambers and online Polarization, there is no labeled golden standard of these qualities that tells how topics are polarized and which ones are more polarized

Table 4. Summary of results for every Chamber of every topic. Columns beginning with "Separability:" for Chamber A refers to its users' separability from its twin Chamber (B) on the *same topic*, vice versa.

Topic	**Chamber**	**Affiliation**	**Var×10^5**	**Separability: SVM Accuracy**	**Separability: SVM Mean-Conf**	**Sample Tweet**
Abortion	A	Democrat	7.5 ± 0.3	89%	0.50	Nobody's life has ever been saved by preventing an abortion.
Abortion	B	Republican	5.5 ± 0.4	89%	0.50	So pro abortion protestors are protesting in cities they can still get abortions?
Gun	A	Democrat	5.8 ± 0.3	92%	0.56	Denmark has tragically experienced another mass shooting.
Gun	B	Republican	4.8 ± 0.3	92%	0.56	Sign the petition against gun control.
Ukraine	A	Democrat	7.6 ± 0.4	86%	0.48	DO YOU NOW GET IT WHY UKRAINE NEEDS ALL WEAPONS THE WORLD CAN GIVE?
Ukraine	B	Republican	6.4 ± 0.3	86%	0.48	#Washington created the fascist regime in #Ukraine... (truncated)
SXSW	A	Non-Political (Affiliation 1)	15.0 ± 0.7	82%	0.45	See you next year #sxsw. My eyes are bleeding but was a blast
SXSW	B	Non-Political (Affiliation 2)	19.6 ± 0.6	82%	0.45	Nice blog from our #Sxsw panel... (truncated)

Table 5. Levels of user separability per pair of Chambers across all the topics. Chamber A is the Democrat and Chamber B is the Republican retweet cluster.

		Chamber A			Chamber B		
		Abortion	Gun	Ukraine	Abortion	Gun	Ukraine
A	Abortion	50%	76%	80%	89%	91%	90%
	Gun	76%	50%	77%	91%	92%	93%
	Ukraine	80%	77%	50%	89%	91%	86%
B	Abortion	89%	91%	89%	50%	69%	80%
	Gun	91%	92%	91%	69%	50%	78%
	Ukraine	90%	93%	86%	80%	78%	50%

than others [31]. This makes it difficult to judge how our method performs with respect to existing works as there is no clear definition of accuracy in this domain. We address this challenge by replicating existing methods over well-established polarized topics. In particular, we chose *Abortion* and *Gun-Control* as topics where we expect a high level of polarization. On the contrary, we chose the *Ukraine* war as a topic where we expect to see lower polarized discussion in the context of the US political sphere — where our tweets come from.

We next compare the results of prior approaches over the topics. In particular, we replicate Garimella et al. [30] method of measuring user's polarity as it is vastly adopted by other scholars. As in Garimella's work, we calculate users' polarity/ideology based on the average polarity of content they had shared online as the baseline. Note that the notion of "user polarity" in [30] is the supervised equivalent of "user embeddings" in our own approach. In particular, we obtain content polarities by forming a labeled dataset of online news sources and Twitter accounts annotated as left-leaning, right-leaning, and centric. We generate this annotated dataset by combining the latest database of AllSides[7] and MediaBiasFactCheck[8] with the labeled dataset of congresspeople and senators in Section 5. Then, for each user u in the dataset, we consider the set of tweets P_u posted by u that contain links to news organizations of known political leaning ln or retweets made from the labeled politician or news accounts on Twitter. We then associate each tweet/retweet $t \in P_u$ with leaning $\ell(t) = ln$. The user polarity $p(u)$ of user u is then defined as the average political leaning over P_u [30]:

$$p(u) = \frac{\sum_{t \in P_u} \ell(t)}{|P_u|}. \tag{3}$$

The value of user polarity ranges between -1 and 1. For users who regularly share content from left-leaning sources, the user polarity is closer to -1, while for those who share content from right-leaning sources, it is closer to +1.

We restrict our comparison to the user-ideology estimation part as the later steps of Garimella's work (e.g., calculating "consumption polarity") require full access to the follower/following networks on Twitter which is no longer accessible via Twitter API.[9] After measuring the user polarity, we proceed to measure both effects with the new supervised foundation of user ideology as our baseline using the definition of Echo and Polarization in Section 3.

Figure 6 shows the distribution of user polarity across each of the Chambers of the baseline. The blue (red) curves represent the distribution of users who showed up in Democratic (Republican) Chambers for each topic (the retweet networks that were promoting Democrats' stances for each topic). The level of flatness of each distribution represents the diversity of sets of users from the entire political spectrum that has appeared in that Chamber [30]. The flatter the distribution of a Chamber, the lower the Echo of voice. Moreover, a high overlap between the distributions of two Chambers of a topic would represent a lower political polarization in the online conversation around that topic. Similar to our results (cf. Figure 5 and Table 4), we see there is an overlap between the distribution of users in the Democratic Chamber and the Republican Chamber in the case of the Russo-Ukrainian war. On the contrary, for "Abortion" and "Gun-Control", Chambers have minimal overlap as in our results, showing a higher level of polarization in those topics. In other words, only right-wing (left-wing) users — ones with positive (negative) polarity scores — had taken Republican (Democratic) stances.

[7] https://www.allsides.com/media-bias
[8] https://mediabiasfactcheck.com/
[9] https://twittercommunity.com/t/starting-february-9-twitter-will-no-longer-support-free-access-to-the-twitter-api/184611

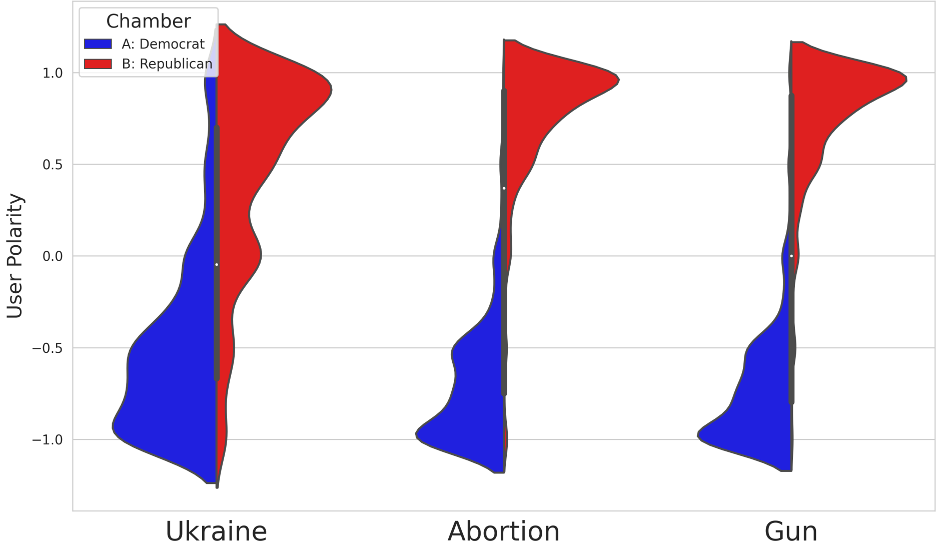

Fig. 6. Users political ideology (polarity) distribution across each Chamber of each topic. Negative values manifest left-leaning ideology and positive values manifest right-leaning ideology.

Table 6. Replication of Table 4 with Supervised Baseline.

Topic	Chamber	Affiliation	Var (Inverse of Echo)	Partisan Stance Rate (Polarization)
Abortion	A	Democrat	0.13 ± 0.02	95.9%
Abortion	B	Republican	0.13 ± 0.02	95.9%
Gun	A	Democrat	0.13 ± 0.02	96.3%
Gun	B	Republican	0.12 ± 0.02	96.3%
Ukraine	A	Democrat	0.24 ± 0.02	85.9%
Ukraine	B	Republican	0.28 ± 0.03	85.9%

We next quantify the level of Echo and Polarization per topic. To compute the baseline, we quantify the Echo by also leveraging the variance of user polarity per topic. For Polarization, we measure the percentage of partisan stances; the rate of users who supported the stances that were aligned with their original political leaning (e.g., the number of left-leaning users who took a pro-abortion stance, and vice versa, divided by the total number of users). The higher the percentage of partisan stances on a topic, the higher would be the topic's polarization. Table 6 shows the baseline results.

We make the following observations when comparing the baseline with our results in Table 4. First, the baseline's results are aligned with our method in terms of polarization judging by the correlation between the *separability* of our approach and the *partisan stance rate* of the baseline (cf. last columns in Tables 4 and 6). In particular, our results show that the Russo-Ukrainian war is the least polarized topic, and Gun control is the most polarized one. For the case of SXSW, the measurement was inapplicable as the Chambers were not initially classified as Democrat or Republican and we also did not find any sufficient number of political references in their tweets. While the results we obtain detecting polarization are comparable with the baseline, we note that our approach is unsupervised and it does not suffer the burden of the labeling process as in the baseline.

Second, we see that the Echo in the Ukraine chambers is the highest in both the baseline and our method as indicated by the "Var" column in Tables 4 and 6. However, we note that the Echo in the chambers of the Abortion and Gun topics in the baseline are not significantly different from one another as opposed to what was expected. Recall that Chambers with Democratic stances preserve higher diversity of users (lower Echo). Instead, our method is able to detect differences in terms of diversity in Democratic and Republican Chambers. We attribute the difference to a limitation of the baseline in measuring the ideology as a one-dimensional pre-defined political spectrum as we discuss in Section 8.1.3. Notably, a transformer-based user-embedding method can represent all sorts of semantic qualities produced by users that can be attributed to the user's political ideology, dialect, gender, etc. manifested in his/her produced content online. Therefore, our results are more aligned with the real-world statistics showing that Democrats are more ethnically diverse when compared to Republicans [53].

8 Discussion

We now discuss our key findings as well as limitations and future work.

8.1 Key Findings

8.1.1 Quantifying Diversity. Leveraging state-of-the-art language models, this paper proposed an intuitive, computationally cheap, and unsupervised approach for quantifying Echo-Chambers and existing polarization phenomenons. The generalizability of our metric enabled us to compare these effects across four topics. The results show that the highest polarization has happened among the Gun-Control topic's Chambers and the lowest for SXSW, the only non-controversial topic of the analysis, followed by the War on Ukraine. Moreover, we showed that the diversity of users in all three controversial topics of our analysis is lower for the Republican stances (e.g., Anti-Abortion) than the Democratic ones (e.g., Pro-Abortion) on the same topic. Pew Research Center had previously confirmed a greater representation of Democrats on Twitter [17]. What our observation adds to the polls is that the users with democratic stances are not only represented higher on Twitter in terms of number but also in terms of diversity.

We discovered that the hashtag "#SXSW", the only non-partisan hashtag of the analysis, expectedly, has the highest diversity of users among the hashtags. Then, among the partisan hashtags, "#StandWithUkraine" has the highest diversity of users. This can mean that manifesting support for Ukraine has been prevalent among people of more diverse sets of ideologies, or/and demographics, or/and etc.

In a scenario where users are mainly located in the US, this could be related to the phenomenon of "Rally Round the Flag" as in political science [33, 45, 51]. Otherwise, this high diversity can hint to the higher variety of user locations in Ukraine supporters, suggesting a higher global involvement with the topic, in comparison to the domestic issues in the analysis (i.e., *gun* and *abortion*).

The term refers to the notion that when a major national conflict takes place, the American people are likely to set aside their disagreements with the incumbent president's policies or performance in office to demonstrate a united front to the international community [8]. Although the high amount of user embedding diversity for "#StandWithUkraine" and Ukraine-related Chambers in Section 7.2 confirms it, the higher similarity (lower linear separability) of the users of the hashtag to Democrats than the Republicans tells that the rally had possibly happened among hard-core Democrats and non-political users, leaving some hard-core Republicans out.

In a related vein, Bailon et al. [35] investigated the extent to which Facebook enabled an *asymmetrical* ideological segregation in political news consumption during the 2020 US presidential election. They found that Conservatives were more likely to be exposed to ideologically homogeneous information than liberals. Combining these findings with our results which show that the homogeneity of user *embeddings*, which is higher for Republicans in our findings, and the homogeneity of users' *news consumption*, which is also higher for Conservatives according to Bailon et al., we can hypothesize that there can be a meaningful causal relationship between the two phenomena.

8.1.2 User Embedding. We embedded users by averaging the sentence embeddings of their tweets. Averaging embeddings have previously been applied to word embeddings to generate an embedding for a sentence [4]. However, to our knowledge, it has not been applied to multiple *sentence embeddings* to represent authors as in our work. As the words of a sentence are elements that are sequentially dependent on each other, their order should preferably be taken into account in an ideal NLP model. However, we posit that averaging would perform better when we are dealing with embeddings of tweets that are the *independent* elements of the user's mindset. Thus, the order would barely mean much in this case. Therefore, we expect that averaging independent sentences' (tweets') embeddings would return meaningful results. Moreover, there is a statistical justification for averaging the embeddings due to the "blessing of dimensionality." Since exponential numbers of embeddings are almost orthogonal in high dimensions, two random sets of embeddings are very unlikely to have similar averages [19].

8.1.3 Quantifying Polarization. It is worth noting that while quantifying the polarization across Chambers using embedding separability, what we measure is the separability of users' *discourse* across Chambers. Yet, understanding the underlying source of discourse separability requires further analyses. As we embed the users utilizing sentence transformers, the encoded features for every user are black boxes that have stored the online semantic behavior of a user. This means that we are not investigating the aspects on which the discourse of the users is polarized. The timeline generated by users can be influenced by his/her sociopolitical leaning, economic leaning, socioeconomic status, gender, age, personality type, geographical location, language variety, etc. Our metric can nevertheless show a high rate of user separability for two Chambers of a non-controversial topic if, for instance, the Chambers are formed based on the local follow-network in different locations and each location's dialect or daily concerns can distinguish its users from other locations.

In this paper, we applied the metric to pairs of Chambers that are known to be different on the basis of political stance on a topic (e.g. pro-gun vs. anti-gun retweet networks) and verified this by sampling a few of the tweets from the retweet network of every Chamber. In such cases, every sort of hidden encoded feature causing a difference between the users of the two clusters is translated as an underlying source of "political" polarization. For instance, if all the women are pro-choice in Chamber A, and all the men are pro-life in Chamber B, the abortion topic is polarized on gender. Alternatively, if most of the southerners in the US are pro-gun and most of the northerners are anti-gun, the Gun-control topic is polarized on geolocation.

Most of the possibly embedded features of users mentioned above can be measured as continuous variables. For instance, sociopolitical or economic views can be anywhere between alt-right to alt-left, and socioeconomic status can be a number anywhere from 0$ to 1M$+ per year). Also, demographic features such as age, gender [47], and ethnicity [56] are considered continuous spectrums of values in recent social science literature. This will make the concept of linear separability a more meaningful metric for such variables, as they will be converted into numbers embedded in a continuous 768D space and separated by a hyperplane. For possible cases of non-continuous features, although the SVM mean confidence interval would be a less meaningful metric as it relies on the distance to the separating hyperplane, the accuracy of the SVM classifier would cover the level of non-continuous divide (e.g. a hypothetical binary division in 1D would be separated by a vertical line in 0.5, yet the distance to that vertical line, which corresponds to SVM's confidence interval, would not yield a meaningful result).

8.2 Comparison with Previous Approaches

Our approach marks a departure from traditional methodologies utilized in prior works, notably those pioneered by Garimella et al., Pablo Barbera, and others [10, 18, 29]. The core idea of previous Echo Chamber measurement approaches centered around establishing correlations between the political leaning of the content the online user is exposed to or believes in, and the political leaning of contents they produce on specific topics. This correlation served as a key metric for evaluating the degree of polarization (i.e., in more controversial/polarized topics, there is a higher correlation between what users consume in general and what they produce on that topic).

User's exposure or user's general belief is typically modeled by the political leaning of the user's neighborhood [18] which is estimated from follow networks representing the connections users have with each other. The leaning of content exposure is determined by examining either the political affiliations of users in Twitter's follow-network (i.e. if user A follows Donald Trump, their score leans more toward conservatism) or by assessing the latent space position of users within this network [10]. In our work, this element is replaced by unsupervised transformers applied to the timelines of users.

The **leaning of produced content** has been traditionally calculated by counting pre-labeled political sources or examining retweets from political figures with predefined leanings. For instance, referencing/retweeting a source like Fox News on the topic of abortion will increase the conservative score of a user on that topic.

We list several advantages and disadvantages of our model when compared to the described previous approaches.

8.2.1 Advantages:

(1) **Availability of Data:** Given the evolving landscape of social media privacy policies, especially regarding the collection of follower data, our method is less vulnerable to the current social media policy restrictions. Notably, since Twitter's reform, the complete following or followers list of users is no longer visible. This trend can also spread to other social networks in the future. Our focus on the minimal amount of open-source timeline data remains a viable alternative.

(2) **Unsupervised Nature:** The reliance of the previous method on pre-labeled political sources makes them not only reliant on expensive crowd-sourcing but also less robust to the fluid nature of political landscape changes and the migration of users to new platforms. For example, as there is evidence of mass migration of users from Twitter to Mastodon [38], an analysis of polarization in a new social media like Mastodon requires new labeling of political sources and celebrities in that platform. Yet, the unsupervised nature of our approach which is based on the embedded features of the timeline, is robust to such changes.

(3) **Multi-Dimensional Understanding of Polarization:** As the foundation of previous approaches is based on sources labeled as politically left or right their understanding of polarization would be limited to political polarization exclusively; and only the left and right duality in political polarization which is not the only type of political divide [32], especially in non-western countries [1, 65]. For instance, religious divisions are more pronounced in nations that have embraced secularization and possess a heritage tied to Catholicism, indicating a heightened polarization influenced by religious passion within secular societies [58]. As sentence transformers in our approach embed various sorts of semantic information produced

by users, the measured polarization in our approach can encapsulate multi-dimensional sorts of polarizations.

8.2.2 Disadvantages:

(1) **Unspecified Source for Polarization:** In scenarios where the primary aim revolves around measuring polarization in classic conservative versus democrat dimensions, the previous methodologies provide more definitive insights into the political sources driving polarization. Unlike these approaches, our method operates as a black-box in determining the specific sources or dimensions contributing to polarization. In Section 8.3, we discuss two approaches to addressing this limitation.

(2) **Less Granularity:** The overlap of content consumption and production in previous approaches offers polarization scores at the individual user level. In contrast, our method evaluates polarization holistically by assigning an overall score to the polarization between two Chambers by looking at the overall separability of their users. However, this limitation can nevertheless be mitigated by examining the distance of users' embeddings from the support vectors' hyperplane in the SVM classifier that separates two Chambers.

8.3 Limitations & Future Work

Our method offers systematic — and unsupervised — insights into the polarization of different Web communities, which led to the key findings presented above. However, as computational social science research that aims to bridge between the *quantitative* domain of computational methods and the partly *qualitative* domain of social sciences, our approach is subject to some assumptions and limitations.

One of the limitations is the absence of an objective ground truth that tells which topic is more polarized or subject to the Echo-Chamber effect with respect to other controversial topics. This limitation is shared with previous work [31] that mentions the intuitiveness of evaluation based on the labeling that a topic is controversial/polarized. The alternative to such methodological assumptions is to hand-label/survey thousands of users [31]. We nevertheless evaluate the core of our method in Section 5 with *ground truth* of congress-people and senators who are labeled as Republican or Democrat, and we show that our method can successfully distinguish between them.

We further evaluated other intermediate steps like the network clustering step by manually labeling a random sample in Section 7.2, and compared our method with a well-established baseline in Section 7.3 showing significant improvements when compared to existing methods.

Future work can utilize our user embedding approach for any task related to user classification (e.g., gender classification and bot detection). In this paper, we embedded the users merely based on their 200 recent tweets. When using Twitter's official API to gather user data, each API response includes 200 tweets per page. As our main focus in this paper was less on reporting an intensive measurement and more on introducing and testing our proposed method, we limited the scraping to 200 tweets per user to remove the need for pagination and make the collection process less time-consuming and complex. This served as a preliminary analysis, which yielded a sufficient amount of accuracy to manifest the separability between users, both in the case of congresspeople and users in different Chambers. Moreover, given the evolving landscape of stringent data access policies, exemplified by the recent measures implemented by Elon Musk on Twitter [10], which are indicative of an industry trend likely to restrict extensive online data accessibility, our demonstration of an approach that is reliant on smaller data subsets aligns with the need for approaches less dependent on data quantity.

[10] https://techhq.com/2023/07/why-has-twitter-introduced-rate-limits/

The scope of this study was limited to quantifying the amount of Echo inside Chambers and polarization across the Chambers. However, the underlying source of the polarizations can be multidimensional, rooting in variations in sociopolitical views [32], economic views, socio-economic statuses, geographic locations, linguistic differences, etc. A potential future direction is to analyze the source of polarization between Chambers by investigating various semantic features in users' timelines and profiles. Instead of a single embedding per user, we can create separate embeddings for different aspects, such as political views and language preferences. These separate embeddings can help us better understand why and how users become separated within chambers.

A more sophisticated approach in natural language processing involves unraveling the opaque semantic features embedded by sentence-transformer models through Explainable AI techniques [60]. By deciphering the semantic meaning associated with each element in the approximately 700-dimensional vectors, we gain the capability to discern the specific semantic features contributing to the separation between two data points that have been semantically embedded. For instance, if we can identify that elements 1, 52, and 401 encapsulate the semantics of political views in texts, while elements 5, 203, and 628 pertain to accent-related features, we can utilize the coefficients derived from classifiers like SVM to elucidate the underlying source of separation. If an SVM classifier assigns high coefficients to elements 1, 52, and 401 for two chambers, it signifies that the polarization between them is rooted in the political views of the users. Similarly, heightened coefficients for accent-related elements in the embedding vector would indicate accent-related features as the source of polarization.

Data & Code Statement

For reproducibility and to facilitate future research on the topic, we release our entire code and anonymized data on GitHub at https://github.com/vahidthegreat/transformer-based-echo-chamber-detection.

Ethical Considerations

Our research is meant to help social scientists, offering a quantified perspective of the Echo Chamber effect, and for online moderators and policy-makers to track and mitigate online polarization and radicalization. Our dataset does not contain any private information. We do not publish author names, IDs, or any information that could be used to identify individuals to respect the privacy of Twitter users. The final results are fully replicable as we open-source our tool, and share anonymized data and the methods we have used to collect it.

Authors' Contributions

In abidance with the ACM Policy on Authorship,[11] all the authors have actively contributed to the project. The first author designed the pipelines and carried out the NLP analysis of the work. The second author carried out the network analysis section of the project. The third author contributed to the data collection and parts of the writing phase. The fourth, fifth, and sixth authors supervised the project and contributed with their guidance on designing the methodology and while reporting the results, which includes making several rounds of revisions in the writing phase.

[11] https://www.acm.org/publications/policies/new-acm-policy-on-authorship#:~:text=Anyone%20listed%20as%20author%20on,information%20is%20given%20to%20ACM.

Acknowledgements

This project was funded by TED2021-132900A-I00, from the Spanish Ministry of Science and Innovation, with funds from MCIN/AEI/10.13039/501100011033, and the European Union-NextGenerationEU/-PRTR; and supported by UKRI through REPHRAIN (EP/V011189/1), the UK's Research Centre on Privacy, Harm Reduction and Adversarial Influence online, as part of its PROM project. G. Suarez-Tangil has been appointed as 2019 Ramon y Cajal fellow (RYC-2020-029401-I) funded by MCIN/AEI/10.13039/-501100011033 and ESF Investing in your future.

References

[1] A. Abramowitz. 2010. The Disappearing Center: Engaged Citizens, Polarization, and American Democracy. (2010). https://doi.org/10.5860/choice.48-1737

[2] Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J Silva, and Bryon C Wallace. 2017. Quantifying mental health from social media with neural user embeddings. In *Machine Learning for Healthcare Conference*. PMLR, 306–321.

[3] Silvio Amir, Byron C Wallace, Hao Lyu, Paula Carvalho, and Mario J Silva. 2016. Modelling Context with User Embeddings for Sarcasm Detection in Social Media. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 167–177.

[4] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *ICLR*.

[5] Mohammad Ayatollahi Tabaar and A.Kadir Yildirim. 2020. Religious Parties and Ideological Change: A Comparison of Iran and Turkey. *Political Science Quarterly* 135, 4 (08 2020), 697–723. https://doi.org/10.1002/polq.13097 arXiv:https://academic.oup.com/psq/article-pdf/135/4/697/48808715/psquar_135_4_697.pdf

[6] Albena Azmanova. 2011. After the Left–Right (Dis)continuum: Globalization and the Remaking of Europe's Ideological Geography. *International Political Sociology* 5, 4 (12 2011), 384–407. https://doi.org/10.1111/j.1749-5687.2011.00141.x

[7] Ozan Aşık. 2024. Ideology, Polarization, and News Culture: The Secular-Islamist Tension in Turkish Journalism. *The International Journal of Press/Politics* 29, 2 (2024), 530–547. https://doi.org/10.1177/19401612221132716 arXiv:https://doi.org/10.1177/19401612221132716

[8] William D. Baker and John R. Oneal. 2001. Patriotism or Opinion Leadership?: The Nature and Origins of the "Rally 'Round the Flag" Effect. *Journal of Conflict Resolution* 45, 5 (2001), 661–687. https://doi.org/10.1177/0022002701045005006 arXiv:https://doi.org/10.1177/0022002701045005006

[9] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132. https://doi.org/10.1126/science.aaa1160

[10] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science* 26, 10 (2015), 1531–1542. https://doi.org/10.1177/0956797615594620 arXiv:https://doi.org/10.1177/0956797615594620 PMID: 26297377.

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[12] Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and Network Dynamics Behind Egyptian Political Polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 700–711. https://doi.org/10.1145/2675133.2675163

[13] Emanuele Brugnoli, Matteo Cinelli, W. Quattrociocchi, and Antonio Scala. 2019. Recursive patterns in online echo chambers. *Scientific Reports* 9 (2019). https://doi.org/10.1038/s41598-019-56191-7

[14] Axel Bruns. 2019. It's not the technology, stupid: How the 'Echo Chamber' and 'Filter Bubble' metaphors have failed us.

[15] Axel Bruns. 2021. Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In *Hate speech and polarization in participatory society*. Routledge, 33–48.

[16] Fernando H. Calderón, Li-Kai Cheng, Ming-Jen Lin, Yen-Hao Huang, and Yi-Shin Chen. 2019. Content-Based Echo Chamber Detection on Social Media Platforms, In 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2019-08). *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 597–600. https://doi.org/10.1145/3341161.3343689

[17] Pew Research Center. 2020. *Differences in How Democrats and Republicans Behave on Twitter.* https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/

[18] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The Echo Chamber Effect on Social Media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118. https://doi.org/10.1073/pnas.2023301118

[19] Ben Coleman. 2020. *Why is it Okay to Average Embeddings?* https://randorithms.com/2020/11/17/Adding-Embeddings.html

[20] Mauro Coletto, Venkata Rama Kiran Garimella, A. Gionis, and Claudio Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Soc. Networks Media* 3-4 (2017), 22–31. https://api.semanticscholar.org/CorpusID:54300115

[21] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication* 64, 2 (2014), 317–332. https://doi.org/10.1111/jcom.12084

[22] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of twitter users. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 192–199.

[23] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. 2020. Falling into the echo chamber: the Italian vaccination debate on Twitter. In *Proceedings of the International AAAI conference on web and social media*, Vol. 14. 130–140.

[24] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. The Spreading of Misinformation Online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559. https://doi.org/10.1073/pnas.1517441113

[25] Tao Ding, Warren K Bickel, and Shimei Pan. 2017. Multi-view unsupervised user feature embedding for social media-based substance use prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2275–2284.

[26] Tao Ding, Warren K Bickel, and Shimei Pan. 2018. Predicting delay discounting from social media likes with unsupervised feature learning. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 254–257.

[27] Joan-María Esteban and Debraj Ray. 1994. On the Measurement of Polarization. *Econometrica* 62, 4 (1994), 819–851. http://www.jstor.org/stable/2951734

[28] Y. Gao, F. Liu, and L. Gao. 2023. Echo chamber effects on short video platforms. *Sci Rep* 13 (2023), 6282. https://doi.org/10.1038/s41598-023-33370-1

[29] Kiran Garimella et al. 2018. Polarization on social media. (2018).

[30] Kiran Garimella, Gianmarco Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. *WWW '18: Proceedings of the 2018 World Wide Web Conference*, 913–922. https://doi.org/10.1145/3178876.3186139

[31] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1, 1 (2018), 1–27.

[32] Vahid Ghafouri, Vibhor Agarwal, Yong Zhang, Nishanth Sastry, Jose Such, and Guillermo Suarez-Tangil. 2023. AI in the Gray: Exploring Moderation Policies in Dialogic Large Language Models vs. Human Answers in Controversial Topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 556–565. https://doi.org/10.1145/3583780.3614777

[33] Vahid Ghafouri, Babak RezaeeDaryakenari, and Nihat Kasap. 2020. *Who rallies around the flag? Analyzing the impact of foreign interventions on nations' political stance using social media data.* Master's Thesis. Sabancı University. https://risc01.sabanciuniv.edu/record=b2473816 [Thesis].

[34] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. 2021. Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey. (08 2021).

[35] Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. Asymmetric ideological segregation in exposure to political news on Facebook. *Science* 381, 6656 (2023), 392–398. https://doi.org/10.1126/science.ade7138 arXiv:https://www.science.org/doi/pdf/10.1126/science.ade7138

[36] Kamile Grusauskaite, Luca Carbone, Jaron Harambam, and Stef Aupers. 2023. Debating (in) echo chambers: How culture shapes communication in conspiracy theory networks on YouTube. *New Media & Society* 0, 0 (2023), 14614448231162585. https://doi.org/10.1177/14614448231162585 arXiv:https://doi.org/10.1177/14614448231162585

[37] Jie Gu, Feng Wang, Qinghui Sun, Zhiquan Ye, Xiaoxiao Xu, Jingmin Chen, and Jun Zhang. 2021. Exploiting behavioral consistence for universal user representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4063–4071.

[38] Jiahui He, Haris Bin Zia, Ignacio Castro, Aravindh Raman, Nishanth Sastry, and Gareth Tyson. 2023. Flocking to Mastodon: Tracking the Great Twitter Migration. In *Proceedings of the 2023 ACM on Internet Measurement Conference* (Montreal QC, Canada) *(IMC '23)*. Association for Computing Machinery, New York, NY, USA, 111–123. https://doi.org/10.1145/3618257.3624819

[39] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLOS ONE* 9, 6 (06 2014), 1–12. https://doi.org/10.1371/journal.pone.0098679

[40] Julie Jiang, Xiang Ren, Emilio Ferrara, et al. 2021. Social media polarization and echo chambers in the context of COVID-19: Case study. *JMIRx med* 2, 3 (2021), e29570.

[41] Mansooreh Karami, Ahmadreza Mosallanezhad, Paras Sheth, and Huan Liu. 2022. Estimating Topic Exposure for Under-Represented Users on Social Media. *arXiv preprint arXiv:2208.03796* (2022).

[42] Mansooreh Karami, Tahora H Nazer, and Huan Liu. 2021. Profiling Fake News Spreaders on Social Media through Psychological and Motivational Factors. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. 225–230.

[43] Joseph T Klapper. 1960. The effects of mass communication. (1960).

[44] Sefa Şahin Koç, Mert Özer, İsmail Hakkı Toroslu, Hasan Davulcu, and Jeremy Jordan. 2018. Triadic Co-Clustering of Users, Issues and Sentiments in Political Tweets. *Expert Systems with Applications* 100 (2018), 79–94. https://doi.org/10.1016/j.eswa.2018.01.043

[45] Natalie Koch. 2023. The problem with rallying around the (Ukrainian) flag. *Space and Polity* 0, 0 (2023), 1–5. https://doi.org/10.1080/13562576.2023.2223129

[46] Yubo Kou, Yong Ming Kow, Xinning Gui, and Waikuen Cheng. 2017. One Social Movement, Two Social Media Sites: A Comparative Study of Public Discourses. *Comput. Supported Coop. Work* 26, 4–6 (dec 2017), 807–836. https://doi.org/10.1007/s10606-017-9284-y

[47] SJ Langer. 2022. Gender is a complex number and the case for trans phantoms. *Studies in Gender and Sexuality* 23, 2 (2022), 136–145.

[48] Q. Vera Liao and Wai-Tat Fu. 2014. Can You Hear Me Now? Mitigating the Echo Chamber Effect by Source Position Indicators. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 184–196. https://doi.org/10.1145/2531602.2531711

[49] Leland McInnes and John Healy. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (02 2018).

[50] Virginia Morini, Laura Pollacci, and Giulio Rossetti. 2021. Toward a Standard Approach for Echo Chamber Detection: Reddit Case Study. *Applied Sciences* 11, 12 (2021), 5390. Issue 12. https://doi.org/10.3390/app11125390

[51] John E. Mueller. 1970. Presidential Popularity from Truman to Johnson. *American Political Science Review* 64, 1 (1970), 18–34. https://doi.org/10.2307/1955610

[52] Martin Müller and Marcel Salathé. 2020. Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic. *CoRR* abs/2012.02197 (2020). arXiv:2012.02197 https://arxiv.org/abs/2012.02197

[53] Frank Newport. 2013. Democrats Racially Diverse; Republicans Mostly White. Online post. https://news.gallup.com/poll/160373/democrats-racially-diverse-republicans-mostly-white.aspx Accessed on July 5th, 2023.

[54] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[55] Shimei Pan and Tao Ding. 2019. Social media-based user embedding: A literature review. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (2019).

[56] Jan Nederveen Pieterse. 1997. Deconstructing/reconstructing ethnicity. *Nations and Nationalism* 3, 3 (1997), 365–395.

[57] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 729–740.

[58] Egbert Ribberink, P. Achterberg, and D. Houtman. 2018. Religious polarization: contesting religion in secularized Western European countries. *Journal of Contemporary Religion* 33 (2018), 209 – 227. https://doi.org/10.1080/13537903.2018.1469262

[59] A Ross Arguedas, C Robertson, R Fletcher, and R Nielsen. 2022. *Echo chambers, filter bubbles, and polarisation: a literature review.* Technical Report.

[60] Sara Salamat, Negar Arabzadeh, Shirin Seyedsalehi, Amin Bigdeli, Morteza Zihayat, and Ebrahim Bagheri. 2023. Neural Disentanglement of Query Difficulty and Semantics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 4264–4268. https://doi.org/10.1145/3583780.3615189

[61] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2017. Anatomy of News Consumption on Facebook. *Proceedings of the National Academy of Sciences* 114, 12 (2017), 3035–3039. https://doi.org/10.1073/pnas.1617052114

[62] Bryan C. Semaan, Scott P. Robertson, Sara Douglas, and Misa Maruyama. 2014. Social Media Supporting Political Deliberation across Multiple Public Spheres: Towards Depolarization. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1409–1421. https://doi.org/10.1145/2531602.2531605

[63] Kai Shu, Amrita Bhattacharjee, Faisal Alatawi, Tahora H Nazer, Kaize Ding, Mansooreh Karami, and Huan Liu. 2020. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 6 (2020), e1385.

[64] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[65] D. Slater and A. Arugay. 2018. Polarizing Figures: Executive Power and Institutional Conflict in Asian Democracies. *American Behavioral Scientist* 62 (2018), 106 – 92. https://doi.org/10.1177/0002764218759577

[66] Mingfei Sun, Xiaoyue Ma, and Yudi Huo. 2022. Does Social Media Users' Interaction Influence the Formation of Echo Chambers? Social Network Analysis Based on Vaccine Video Comments on YouTube. *International Journal of Environmental Research and Public Health* 19, 23 (2022). https://doi.org/10.3390/ijerph192315869

[67] J. Treviranus and S. Hockema. 2009. The value of the unpopular: Counteracting the popularity echo-chamber on the Web. *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)* (2009), 603–608. https://doi.org/10.1109/TIC-STH.2009.5444430

[68] P. Törnberg. 2018. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS ONE* 13 (2018). https://doi.org/10.1371/journal.pone.0203958

[69] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and Fake News: Early Warning of Potential Misinformation Targets. *ACM Transactions on the Web* 13, 2 (2019), 10:1–10:22. https://doi.org/10.1145/3316809

[70] Giacomo Villa, Gabriella Pasi, and Marco Viviani. 2021. Echo Chamber Detection and Analysis. *Social Network Analysis and Mining* 11, 1 (2021), 78. https://doi.org/10.1007/s13278-021-00779-3

[71] Austin Horng-En Wang, Yao-Yuan Yeh, Charles K.S. Wu, and Fang-Yu Chen. 2023. Why Does Taiwan Identity Decline? *Journal of Asian and African Studies* (2023), 00219096231168068. https://doi.org/10.1177/00219096231168068

[72] Dandan Wang and Yuxing Qian. 2021. Echo Chamber Effect in Rumor Rebuttal Discussions About COVID-19 in China: Social Media Content and Network Analysis Study. *Journal of Medical Internet Research* 23 (2021). https://doi.org/10.2196/27009

[73] Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. 2017. Community preserving network embedding. In *Thirty-first AAAI conference on artificial intelligence.*

[74] Zhen Zhang, Hongxia Yang, Jiajun Bu, Sheng Zhou, Pinggang Yu, Jianwei Zhang, Martin Ester, and Can Wang. 2018. ANRL: attributed network representation learning via deep neural networks.. In *Ijcai*, Vol. 18. 3155–3161.

Received 18 July 2023; revised 15 April 2024; accepted 27 June 2024