Identifying and Characterizing New Expressions of Community Framing during Polarization

Hernan Sarmiento,[1,2] Felipe Bravo-Marquez,[1,2,3] Eduardo Graells-Garrido,[4] Barbara Poblete[1,2,3]

[1]Department of Computer Science, University of Chile
[2]Millennium Institute for Foundational Research on Data (IMFD), Santiago, Chile
[3]National Center for Artificial Intelligence (CENIA), Santiago, Chile
[4]Data Science Institute, Universidad del Desarrollo
{hsarmien,fbravo,bpoblete}@dcc.uchile.cl, egraells@udd.cl

Abstract

Chile experienced a series of important protests between October and December 2019. This *social unrest*, as it was called, was fueled by social inequity and radically affected the nation's status quo. A large portion of the population demanded a new Constitution and changes to the current government, whereas another part of the population rejected these social demands. This created a highly polarized scenario evidenced through online social media interactions. Analyzing controversial issues that emerge naturally from conversations in online communities can offer a more wide-scale understanding of today's political and societal discussions. Here, we analyze group polarization in social networks by studying the 2019 Chilean social unrest. Specifically, we propose an unsupervised approach for identifying and characterizing community framing (i.e., discovering and understanding polarized concepts). Our approach is based on the sequential application of community detection, topic modeling, and word embedding methods. The novelty of having an unsupervised approach is that it facilitates the performance of scalable and objective framing analyses with minimal human intervention, as it does not require prior domain or network knowledge. Using this methodology, we observe that an apparently similar conversation topic across communities can actually have completely different meanings to each group. We noted, for instance, that while an online community linked the term *gente (people)* with communism and terrorism, the other associated it with police and military oppression. In this direction, our work can help to contextualize real-world social issues in online platforms, describing how users discuss similar concepts with opposing views.

Introduction

Group polarization occurs when the tendency of individual group members is enhanced by group discussion (Sunstein 1999). It can often trigger more radical group decisions than those generated by average individuals in the group (Isenberg 1986). In recent years, polarization has been widely studied within the context of online discussions. Specifically, social media has greatly increased the volume of online exchanges among users, in particular about social and political issues.

Prior research in the area has found that online groups can tend to develop their own discourses and vocabulary around certain topics or events (Darwish et al. 2018; Demszky et al. 2019; Graells-Garrido, Baeza-Yates, and Lalmas 2020; Li, Porco, and Goldwasser 2018; Stefanov et al. 2020). Thus, developing parallel realities over time for each community. Instead, our work identifies and characterizes common themes that are discussed in communities, which are used as a proxy for *framing* analysis. In social science theory, this concept describes how people develop a particular conceptualization of an issue considering a variety of perspectives or themes, reorienting their thinking about a subject (Azjen 1980; Nelson, Oxley, and Clawson 1997; Chong and Druckman 2007).

The relevance of framing analyses lies in studying common elements that each group brings into their discussions, which can be developed by either prioritizing the issue or its attributes (McCombs and Shaw 1993). However, the fundamental challenges of framing studies are that frames are often manually selected, requiring necessary domain-specific knowledge, research, and familiarity with the data at hand (Ylä-Anttila, Eranti, and Kukkonen 2020).

Framing processes have been investigated in the context of social movements (Bitschnau, Lichtenstein, and Fähnrich 2021; Gerbaudo 2017; Rane and Salem 2012). Uncertainty during these events offers attractive opportunities to promote the creation of frames, as it often involves the disintegration of long-standing norms and beliefs that have long been taken for granted in society (Bitschnau et al. 2021). The natural structure of social media platforms has played a prominent role in facilitating communication and coordination of community framing. In the form of interpersonal communication, sites such as Facebook and Twitter have organized and disseminated information, promoting collective meaning-making processes within communities (Kavada 2020).

In order to gain deeper insight into polarization in extreme events, we identify and characterize how different groups define their reality by selecting more salient and common frames in online discussions. Motivated by the unprecedented characteristics of the social unrest that affected Chilean society and the challenges of frames identification in social media, we address the following research ques-

tions: 1) which *frames* can be found in online polarized communities during the event? 2) how do we automatically identify these *frames* in online discourses? 3) how do we measure community framing during polarization?

We propose an unsupervised approach that infers frames in online users' discussions. We apply our method to the conversations around the 2019 Chilean social unrest to quantify differences and similarities in the messages. Our method consists of the following steps: we first detect communities based on retweet interaction by evaluating multiple algorithms and metrics. We then identify common topics discussed by these communities by topic modeling, where the most prominent terms are analyzed from a polarization point of view using a joint word embedding model. The most salient property of our method compared to previous approaches is that all steps are performed in an unsupervised manner.

Our main contributions can be summarized as follows:

- We introduce the first study of polarization in social media related to the 2019 Chilean social unrest.
- We release a dataset of Twitter messages about the 2019 Chilean social unrest. These messages differ in language, location, and cultural characteristics from existing literature.
- We propose an unsupervised approach to automatically identify and characterize polarized concepts that are used as a proxy for framing analysis.

Our findings show that communities were polarized regarding users' interaction and the common terms derived from their messages. Results demonstrate that these terms displayed different quantitative and qualitative patterns between groups. For instance, terms such as *gente (people)* and *gobierno (government)* exhibit opposite meanings when we compare communities. This can contribute valuable insight to contextualize online conversations during social discussions, providing an overview of how groups express their opinions. Additionally, our method interplays both network and content methods to estimate polarization, empirically demonstrating that combining these techniques is beneficial for comprehending users' stances.

Our dataset, which is available for non-profit research purposes,[1] is a large collection of public messages in Spanish of the social unrest in Chile.

Roadmap. The paper is organized as follows: First, we review previous work on group polarization. Second, we provide a background for the 2019 Chilean social unrest and their implications. Third, we specify the dataset collected for our research. Fourth, we detail the proposed methodology and presents a quantitative and qualitative semantic analysis. Later, we discuss our main findings and implications of our work. Finally, we present our conclusions and future work.

[1] https://github.com/hsarmiento/chilean_unrest_dataset

Related Work

When discussing controversial issues, online users tend to be exposed to agreeable opinions (Barberá et al. 2015). One of the explanations of this exposure is homophily, where "individuals associate with similar ones" (Bessi et al. 2016). This phenomenon, among other factors related to media consumption, reinforces users' perceptions and blinds them from other sides of the issues under discussion (AlDayel and Magdy 2021). Then, the primary input for studying polarization is to find the several stances present in the discussion and then find groups of users with the same stance. Having this categorization helps to measure and mitigate the problems derived from polarization in social networks.

Content-based Polarization Analysis. Studies analyzing polarization in social media have relied on aligning users toward a set of specific topics or entities (Aldayel and Magdy 2019; Lai et al. 2018). The primary assumption is that communities are preliminarily identifiable based on established target topics, use common hashtags and vocabulary for label propagation, and consider a set of seed users for constructing communities. Traditional features used to determine polarization are based on extracting characteristics derived from content published by users. In this sense, authors have considered n-grams to capture the stance of the users in subjects such as abortion and gay marriage (Anand et al. 2011), and legalization of abortion and climate change (Mohammad et al. 2016).

Studies have also focused on the user's vocabulary. The hypothesis is that individuals with the same stance tend to use the same vocabulary choices to express their points of view (Darwish et al. 2020). Using an embedding approach, Benton and Dredze (2018) proposed a semi-supervised method to represent users based on their online activity. The general idea is to use the context of the users' tweets to construct author embedding and then predict the stance. Similarly, Li, Porco, and Goldwasser (2018) considered a joint embedding learning to determine users' stance using the Internet Argumentation Corpus. For each topic, the authors created individual embedding vectors, which represent pro and against stances. Considering a case study of the Turkey elections, Kutlu et al. (2019) trained an embedding vector using fastText with a skip-gram model on related tweets. The authors relied on the work of Garg et al. (2018) which demonstrated that word embeddings capture gender and racial stereotypes by comparing word vectors that are trained on different corpora to understand how a given term is defined semantically. Considering these results, the work presented by Kutlu et al. (2019) used different word vector models for each politician and stance (e.g., pro and anti-Erdogan). Using a list of known-polarized adjectives and political terms, they compared the 2,000 nearest neighbors word of each to understand the difference among low-dimensions vectors qualitatively.

Researchers have also incorporated framing analyses to understand polarization. Demszky et al. (2019) presented a study of 21 U.S. mass shooting events to measure polarization in common frames in Twitter. Considering a predefined list of Twitter accounts of U.S. Congress members and pres-

idential candidates, they applied a label propagation method to determine the users' political party. They trained a word embedding model to estimate frames, applied k-means clustering to discover common concepts, and manually assigned topics names to inspect the tweets. Finally, they computed a leave-out estimator to measure polarization between and within partisanships for each frame.

Network-based Polarization Analysis. One of the traditional approaches that have been widely utilized to infer users' stances is the retweet network. Guerrero-Solé (2017) analyzed the Catalan process towards independence on the 1,000 most retweeted users. To detect communities, they assigned a label to every edge to identify them by political orientation. They performed an iterative process by which a given user inherits a set of users' labels retweeted. Concerning Egyptian political polarization, Borge-Holthoefer et al. (2015) presented a network approach to track polarity evolution over time. Using a label propagation algorithm for detecting communities, they identified polarized groups considering an initial list of seed users for whom the partisan leaning was clear. A study of polarization in Egypt about Secularists and Islamists in different languages also considered an analysis of the retweet network (Weber, Garimella, and Batayneh 2013). Using the NodeXL and Fruchterman-Reingold communities algorithms, authors showed a polarized network describing Islamist, Secularist, and Center stances. Similar to the work of Borge-Holthoefer et al. (2015), the proposal needed a seed of users to obtain polarized communities.

Others works have also considered network features, by including these characteristics as attributes for supervised, semi-supervised and unsupervised approaches. For instance, Darwish et al. (2018) predicted online Islamophobia over time using the 2015 Paris terrorist attack in Paris as a case study. Among other features, authors considered network features such as the accounts that a user mentioned, retweeted, and replied to. Considering three different polarized events, Darwish et al. (2020) presented an unsupervised framework for detecting stance on Twitter. Their approach extracted several network characteristics such as the number of unique tweets, hashtags, and retweeted accounts with computing similarity among users. So, they applied dimensionality reduction and clustering techniques to obtain communities. Following a similar approach, Stefanov et al. (2020) identified an initial set of users' stances based on the previous methodology and then trained a classifier to determine the position of the other users. Both studies reported accuracy and f1-score values over 80% obtaining two clusters on average.

The 2019 Chilean Social Unrest

During October 2019, a series of demonstrations were initiated in Santiago, Chile's capital and largest city. Initially, the reason was the adjustment of fares for Santiago's public transport system reaching 830 Chilean pesos (US$1.20). On October 18th, high school students coordinated a fare evasion campaign, leading spontaneous takeovers of Santiago's main metro stations. This triggered open confrontations with the local police, known as *Carabineros*. The situation intensified when riots led to the destruction of retail stores and several metro stations, generating important damage to the city's infrastructure. That same night, the president of Chile, Sebastián Piñera, declared a state of emergency in the most populated regions across the country. The next day, a curfew instated in Santiago.

Protests took place in several cities with demands for changes in the Constitution and Sebastián Piñera's resignation. These have been considered the worst civil unrest in Chile since the end of the military dictatorship[2]. Human rights violations were documented during demonstrations,[3] and research has found that police repression and governmental mishandling intensified the crisis as it erupted (Somma et al. 2021; Navarro and Tromben 2019).

Demonstrators across the country differed in terms of age, gender, and social status (Durán 2020). The most iconic and popular place of manifestation was Plaza Italia square, an emblematic location in Santiago that connects several city areas with different socio-economic levels. These activities were mainly driven by young people from middle and lower income areas. They primarily demanded a fairer society based on social rights recognized and validated in a new Constitution. However, a portion of the population claimed that, instead, some reforms to the supreme law could support social demands (Henríquez 2020). Journalists and social scientists identified a polarized organization of society during the event, characterized by being *against* and *in favor* of social movements (Durán 2020; Radovic 2020; Fábrega 2020).

Dataset

We collected Twitter data covering October 19 through November 30, 2019. Initially, our data collection was based on trending terms[4] related to the event (e.g,. *#chiledesperto*, *#piñera*). Given that multiple sub-events occurred in the country during the social movement, identifying all trending topics was challenging in collecting data in terms of the bias and diversity of the content. To mitigate this issue, we complemented our data collection with messages obtained from two other systems (Peña-Araya et al. 2017; Poblete et al. 2018). Galean (Peña-Araya et al. 2017), is a platform that compiles conversations about Chilean news events from Twitter. In addition, we included data from Twicalli, a tool that constantly retrieves tweets posted from Chile using coordinates (Poblete et al. 2018).

We considered only messages published in Spanish, as it is the main language used in Chile. In order to avoid fake accounts, our study only considers user accounts that were created before the event. Hence, we removed over 64,000 users whose creation date was after October 18, 2019, including their messages (291,000).

After merging the three sources and removing duplicates by tweet id, our unified collection contained almost 30 mil-

[2]https://en.wikipedia.org/wiki/2019-2022_Chilean_protests

[3]Human Rights Watch report available at https://www.hrw.org/news/2019/11/26/chile-police-reforms-needed-wake-protests

[4]Full list of keywords available in our repository.

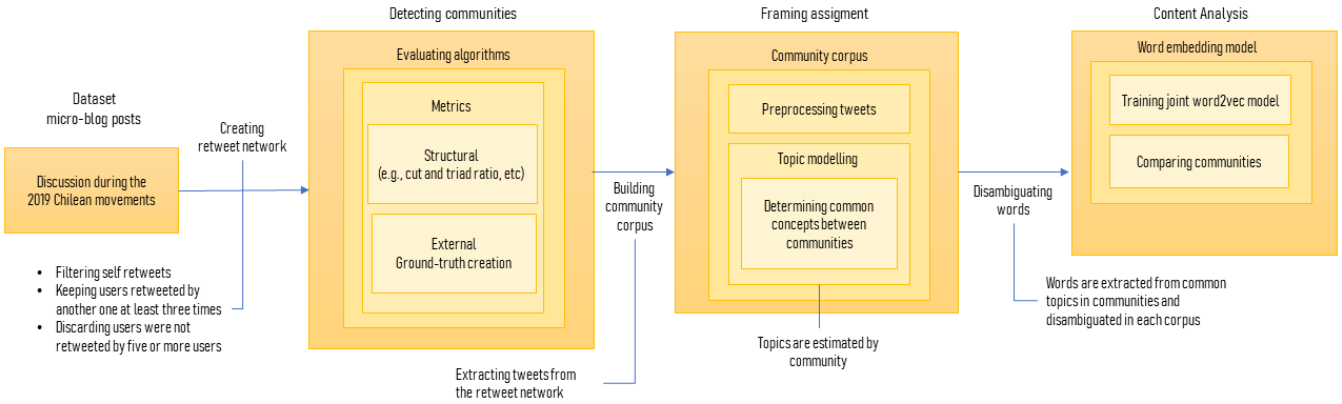

Figure 1: An overview of our proposed methodology.

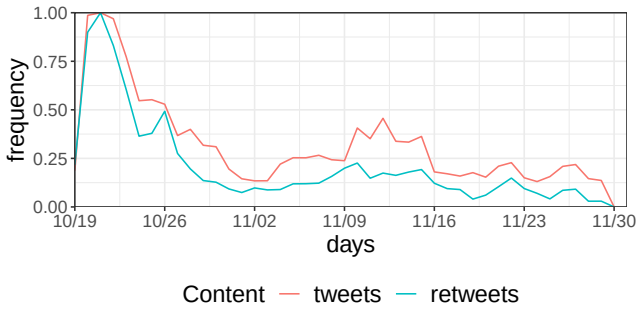

Figure 2: Max-min normalization of daily tweet and retweet frequencies.

lion messages shared by 2.1 million unique users, with over 7 million tweets and over 22.5 million retweets. Figure 2 shows the normalized frequency by day for tweets and retweets. The frequency was much higher on the first days of the event and increased slightly between November 11 and 16. This last spike was due to a referendum discussion for a new Constitution that took place in those days.

Proposed Approach

In this paper, we interpret *frames* as concepts discussed by different communities around a common object (or event), which in our case is the 2019 Chilean social unrest. We hypothesize that in these events, where communities tend to polarize towards these frames in social media, an appropriate combination of lexical and network analysis tools should allow us to analyze this framing automatically. In this sense, we consider that common topics emerging from each community conversation can be used as a proxy for *frames*. Different previous studies have explored this approach of identifying topics (and their most salient concepts) automatically for framing analysis (Pashakhin 2016; Ylä-Anttila, Eranti, and Kukkonen 2018, 2020).

Figure 1 shows a general overview of our proposed methodology. We first identify groups of users who are likely to perceive the same frames differently. For this purpose, we considered community detection methods that are evaluated in the retweet network. Then, we determine topics discussed by these communities using topic modeling techniques. Mapping these topics between the different communities allows us to automatically determine our frames (i.e., the concepts relevant to all other communities). Next, we extract the most salient words of these common topics to establish the target words that represent the framing. Finally, we train a joint word embedding model to project the meaning of these frames (represented by our target words) into a unified semantic space. So, it enables us to quantify the framing by calculating vector operations between the same concept for different communities.

Detecting Communities

We rely on the concepts of echo chambers, which states that opinions or beliefs stay communities created by like-minded people who reinforce and endorse views of each other. We inferred communities considering that the retweet network with a clustered structure could represent different opinions and points of view. Initially, our retweet network comprised over 2.1 million users (nodes) and 22.5 million interactions (directed edges). We filtered the network based on three conditions to reduce the noise by small communities or weak connections among users. We removed self retweets (a user that retweets itself). We kept users retweeted by another one at least three times. We discarded users that were not retweeted by five or more users. After filtering, our network contained 220,118 users and almost 900,000 connections.

To detect communities, we considered five commonly used community detection algorithms implemented on the *cdlib* library (Rossetti, Milli, and Cazabet 2019): Eigenvector, Greedy, Infomap, Louvain, and Stochastic Block Model. We chose the Stochastic Block Model after executing each method several times and evaluating four classes of scoring functions described by Yang and Leskovec (2015). Details about experiments and results are in our repository.

Figure 3 shows a hive diagram of the retweet network, where nodes (users) are colored according to their community. We sorted users by in-degree and out-degree values in each community to represent those who re-shared content

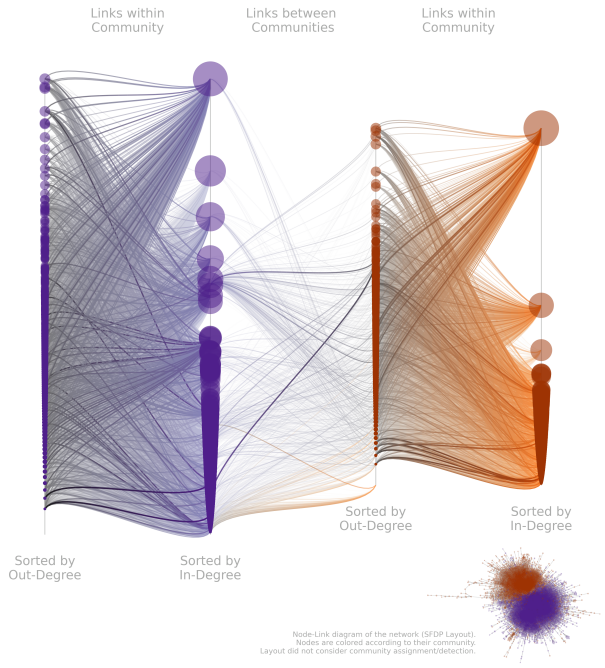

Figure 3: Hive diagram of the detected communities using the Stochastic Block Model results. Orange and purple (right and left) represent against and in favor communities, respectively. The figure shows users sorted by in-degree and out-degree values in each community.

from others (retweet action) and those who posted a message (tweet action). Our results revealed that the retweet interaction mainly occurred among users of the same community, where 92% of the user connections were made by users within the same group.

We evaluated the quality of the network community based on a ground-truth sample. For this task, we deemed into two types of users described as *in favor* and *against* the social movement. We randomly labeled 2,100 accounts that represent 1,110 and 990 in favor and against users, respectively. Our results display that the Stochastic Block Model obtained a f1-score, precision, and recall of 0.777, 0.7986, and 0.7574, respectively.

Framing Assignment

Domain knowledge is an essential aspect of analyzing group polarization on social media. It allows studying different points of view by choosing specific themes or concepts to compare two or more communities. The process that people develop a particular conceptualization of an issue or reorient their opinions about a matter is described as *framing*. In general, this requires expertise and familiarity with the matter, challenging and demanding, especially for unexpected and dynamic events. To deal with this challenge, we aim to identify topics discussed in the tweets of both communities, which can be used as a proxy for framing analysis. Creating two corpora consisting of the tweets shared by the groups, we extract topics independently by the community and compare their similarities based on the terms they compose. These terms or *target words* will be later used for comparing conversations between users' groups. Our target words correspond to the most representative terms for the topics that are jointly discussed by all groups.

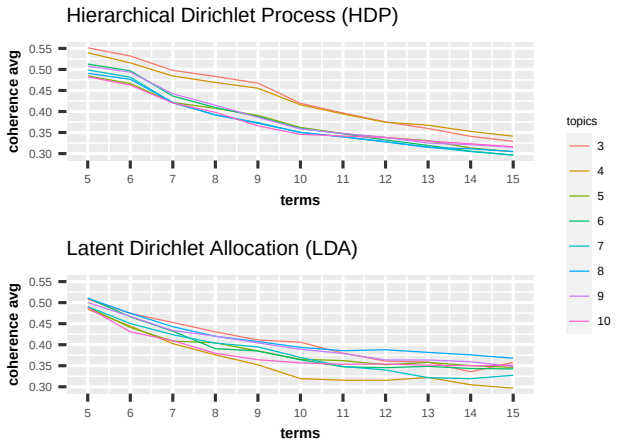

Figure 4: Average coherence for the HDP and LDA algorithms between communities.

We estimated the topics using two well-known topic modeling algorithms: the Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process (HDP). We considered the messages of those 23,864 users in our retweet network that publish at least a tweet. On the one hand, the against community comprised 10,485 users and 611,976 messages. On the other hand, the in favor community contained 13,379 users and 945,832 messages. We applied commonly used pre-processing steps to the text, such as removing accents, URLs, Spanish stopwords, hashtags, user mentions, punctuation, and numbers, and converting to lowercase. Nonetheless, we kept maintained numbers in the text (e.g., p1ñ3r4) and transformed emoticons and emojis[5] to plain text to avoid encoding problems. Additionally, we applied a phrase detection model [6] to automatically extract multi-word expressions instead of using traditional n-grams. Finally, given that our interest is in users, we group messages by the user and concatenate them in one string. Hence, we trained our topic models considering each user as a document.

We ran both algorithms with default parameters and extracted the topmost probable words in each topic in a range of 5 and 15 terms. In the case of the LDA method, we trained the model using the number of topics parameter between 3 and 10. For the HDP model, the algorithm did not require to set the number of topics to train the model. However, when extracting the topics, we considered the same number as the LDA method for a fair evaluation of both models. We used the coherence metric to evaluate the quality of the topic models, computed on each community's resulting top-

[5]We used the emoji library available at https://github.com/carpedm20/emoji

[6]We used the Gensim implementation available at https://radimrehurek.com/gensim/models/phrases.html

ics. Figure 4 shows the performance of each model where we note that the HDP model shows a higher coherence value than the LDA model. In our experiments, we chose the HDP model with $topics = 3$ and $terms = 5$ which represents the best performance.

After extracting topics in each community corpus using the HDP method, we identified what similar themes were discussed by both communities and then extracted the most salient concepts that characterize group conversations. Using the overlapping similarity with a threshold of 0.8, we obtained 6 terms that described common themes between groups' conversations. These *target words* obtained from the common topics are the following: *chile, dictadura (dictatorship), gobierno (government), gente (people), piñera* and *venezuela*.

Content Analysis among Communities

To understand community framing in polarized discussions, we studied how different the *target words* are in groups. We treated the same target words from different communities as different lexical units, but forced them to reside in the same semantic space. To do this, we created a joint word vector model in which the *target words* are disambiguated according to the community they appear in. This means that the target word $word_i$ will be renamed in each corpus with a prefix *cj* to identify in which j community the word appears. For instance, the target word *piñera* was renamed as *c1_piñera* and *c2_piñera* in $corpus1$ and $corpus2$ respectively. Thus, we can apply vector operations on these words (e.g., similarity, neighborhood) to measure community framing and polarization.

Concatenating both community corpora, we trained a word2vec model using the skip-gram negative sampling method (Mikolov et al. 2013) and the following parameters: $window\_size = 7$, $epoch = 15$, $vector\_size = 100$ and $min\_freq = 5$. Considering our word embedding model, we first computed the Euclidean distance and the cosine similarity of a pair of disambiguated target word vectors. Using as example the target word *chile*, we estimated both metrics as the $sim\_cosine(c1\_chile, c2\_chile)$ and $eucl\_dist(c1\_chile, c2\_chile)$, where both terms were represented as word vectors in our model.

Table 1 shows the results of computing these metrics. Our results exhibit that the terms *gente (people)* and *gobierno (government)* obtained the lowest cosine similarity values between 0.56 and 0.58. In contrast, the words *piñera* and *chile* had the highest similarities reaching values close to 0.8. Inspecting their Euclidean distances, an exciting observation highlights that *chile* achieved the lowest value. At the same time, the rest of the three mentioned terms obtained more than double the first one. These initial findings suggest that communities could be aligned to a much similar meaning between groups regarding the themes derived from the target word *chile*. And communities could be more polarized around the frames *gente (people)* and *gobierno (government)*, this is, they could diverge to different subjects.

We validated our results using a null model by considering that the communities' content was randomly assigned in each group. We trained this model by randomly swapping half of the content between both communities' corpora and then disambiguating the target words in each. Next, we merged both corpora, trained a joint vector model with the same parameters as the original model, and computed the cosine similarity and Euclidean distance using this null model. Table 1 includes the average results of the null model by swapping the corpus and training a new word vector model 100 times. We noted that both cosine similarities and euclidean distances were utterly different from the original. Hence, our results suggest that despite the vocabulary in both original and null models being the same, our method relies on correctly disambiguating target words in each corpus.

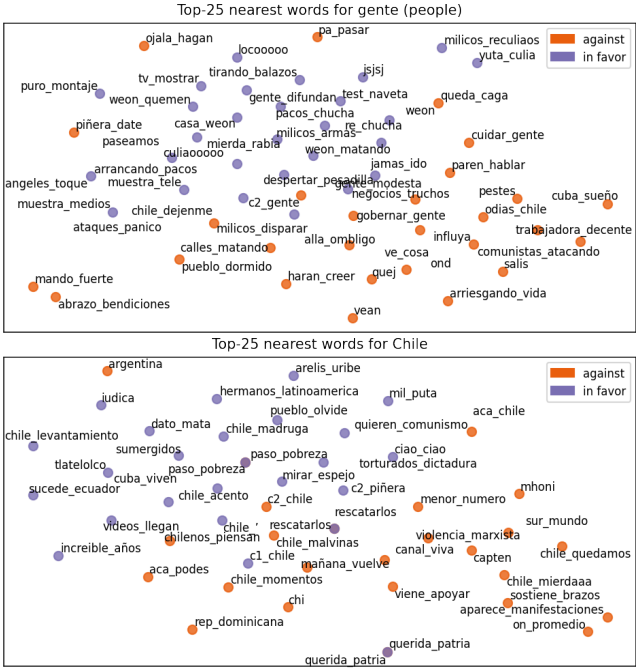

Figure 5: Two dimensional word vector representations of top-30 nearest words for the target words *gente (people)* and *chile*. Orange and purple points represent the nearest words against and in favor of communities, respectively.

We estimated the top nearest terms for each target word to gain more insights into the mentioned differences in the target words. We then visualized them, applying a dimensionality reduction using the T-SNE (T-distributed Stochastic Neighbor Embedding) algorithm. Figure 5 shows two examples of the top-25 nearest terms for the target words *gente (people)* and *chile*, which had the lowest and highest cosine similarities respectively [7]. Our results show that, for both target words, nearest words have a similar distribution in a 2-dimension space, where communities can be visually identifiable. However, we noted differences in the number of overlapping words that both communities had depending on the target word analyzed. The word *gente (people)* did not share any term between communities, while the *chile* had three

[7] The complete visualization of all target words is published in our repository.

	Cosine similarity		Euclidean distance	
term	original	null model	original	null model
gente	0.5624	0.7223 ± 0.033	2.5073	2.2915 ± 0.186
gobierno	0.5867	0.7669 ± 0.009	2.9752	2.3310 ± 0.085
venezuela	0.6023	0.7959 ± 0.013	3.4270	2.4905 ± 0.111
dictadura	0.6949	0.7799 ± 0.020	2.7852	2.4231 ± 0.118
piñera	0.7677	0.8428 ± 0.011	2.8218	2.0483 ± 0.087
chile	0.8043	0.9050 ± 0.008	1.2672	0.8948 ± 0.045

Table 1: Cosine similarity and Euclidean distance for target words between communities in our word embedding model. Rows are sorted by cosine similarity.

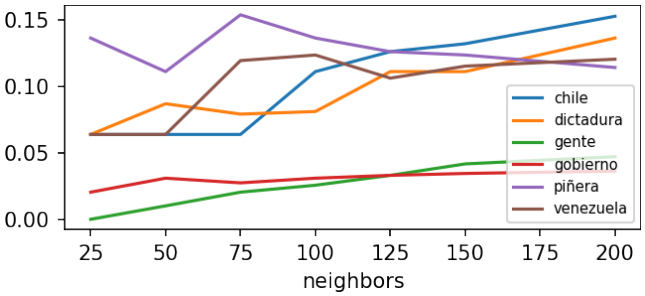

Figure 6: Jaccard index for the k-nearest terms found between communities for each target word.

common terms. Expanding the analysis to a higher number of top nearest terms for these target words, we noticed a similar pattern in which the number of common words increases depending on how close the cosine similarity was. Figure 6 shows the Jaccard index for a different number of k-nearest terms, revealing that the target words *gente (people)* and *gobierno (government)*, which had the lowest cosine similarities, slightly suffer a change in the number of common nearest terms when the neighbors' words increase. In contrast, the other target words have more than four times the common nearest words compared to the two previous ones.

To understand what themes or concepts surround the target words by the community, we quantitatively inspect the top nearest terms to find differences or similarities in the group discussions. Like the previous analyses, we focused on those target words where we found the highest and lowest cosine similarity values. We identified for the target word *gente (people)* that most of the terms found in the *against* community were linked to communism attacks associated with crime, street vandalism and terrorism (See Table 2). Several statements of this community declared that, if social demands were met, Chile would turn into Venezuela or Cuba (Fuentes 2020). They also claimed that the social unrest is not due to a grassroots movement, but rather to Cuban-Venezuelan infiltrated agents. Hence, most of the conversations tended to oppose communist governments. In contrast, the *in favor* community highlighted words that allude to violence and repression by the police and the military. During the social unrest, both institutions were accused of human rights violations during curfew hours (News 2019). These actions were reported in social networks with multiple messages and images. This community tended to refer to the police and military in a colloquial and derogatory manner using the terms (*paco* and *milico*) followed by insults. In addition, several terms also mentioned criticism of press coverage and unease with the government's actions (See Table 2).

For the target word *government (gobierno)*, we noted that the nearest terms for the against community referred to protecting the police and normalizing the situation because of the demonstrations. Oppositely, the in favor group linked their nearest terms with criticizing the president and the deployment of the military to the streets. This last situation alluded to memories lived in Chile 30 years ago, where the military took the streets before the 1973 Chilean coup d'état[8].

Regarding the target word *chile*, we found fewer differences between communities than the previous words (in terms of words with the highest cosine similarity). For instance, the against community mentioned Marxist violence, while the in favor group discussed torture and dictatorship. However, as Figure 6 shows, we found several common nearest words between communities related to the president of Chile and expressions of support related to the country.

Finally, we consider a low dimension analysis to quantify target word differences in communities. We followed a similar proposal presented by Sweeney and Najafian (2019) that measures fairness in word embeddings via the relative negative sentiment. The general idea is that words can be projected from the embedding space into a sentiment probability by training a logistic regression on some pre-labeled words (a set of 80 positive and negative emojis in our case (Wang et al. 2018)). Then, the probability of negative sentiment for some sensitive words (the target words of each community in our case) is normalized to a probability distribution. Subsequently, the negative sentiment distribution of each community is compared to a uniform distribution using the Kullback-Leibler (KL) divergence as a measure of bias. This allows us to quantify the magnitude of polarization for the target words of each community under the assumption that a neutral community should be closer to a uniform distribution.

Figure 7 shows the estimated sentiment probability of the target words for each community. Our results suggest that, in general, a same target word can display different sentiment

[8]https://es.wikipedia.org/wiki/Golpe_de_Estado_en_Chile_de_1973

Target word	Against	In favor
gente (people)	comunistas_atacando (communism_attacking), comunistas_protestas (communism_protests), romper_quemar (smash_burn)	milicos_disparando (military_shooting), golpeando_disparando (smashing_shooting), pacos_pegando (carabineros_smashing), noticieros_culiaos (fucking_news), piñera_payaso (piñera_clown)
gobierno (government)	respeto_instituciones (respect_institutions), proteger_carabineros (protect_carabineros), normalizar_situacion	piñera_cobardia (piñera_cowardice), autocritica_gobierno (self-criticism_government), ejercito_salir (army_takesstreet)
chile	aparece_manifestaciones (appears_manifestations), violencia_marxista (marxist_violence), pdte_piñera (president_piñera), chile_mierdaaa, querida_patria (dear_country)	chile_levantamiento (chile_ uprising), tanques_calles (tanks_streets), pdte_piñera (president_piñera), chile_mierdaaa, querida_patria (dear_country)

Table 2: Examples of nearest terms for target words divided into against and in favor communities.

probability in each group. We also noted that the term *piñera* was unique in the sense that it obtained a similar polarity in both communities. Observing the polarities in detail, we obtained an unexpected result for the target word *dictadura (dictatorship)*, where the in favor community showed a higher positive level than the against group. This result can be attributed to the inability of word embeddings to discriminate between the different meanings of a word. Thus, "dictatorship", referring to the Chilean (right-wing) dictatorship of Pinochet, and the Venezuelan (left-wing) dictatorship of Maduro, could be being conflated.

Regarding the KL scores, we compared the negative distributions of each group with a uniform distribution using the Kullback-Leibler divergence (KL) (see the bottom image in Figure 7). We obtained values of $KL_{against} = 0.0215$ and $KL_{infavor} = 0.0498$ for the against and in favor communities respectively. As noted, the value of $KL_{against}$ doubles the value of $KL_{infavor}$, representing a significant difference in the amount of information necessary to encode and transmit from one distribution to another. Hence, our results suggest that target words exhibit different polarities in both communities and that the in favor community shows more intense sentiment states.

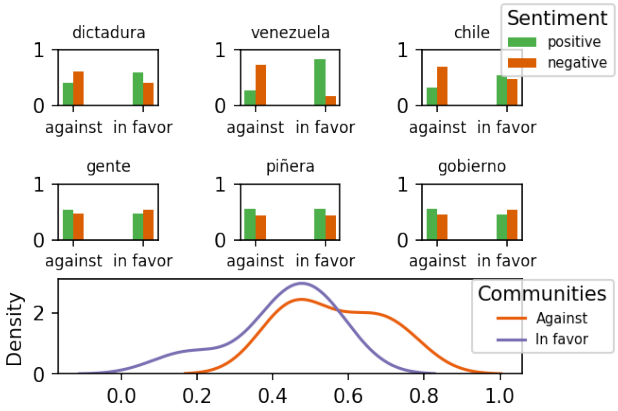

Figure 7: Sentiment probability of the target words. The top image shows the community's positive and negative probabilities of each target word. The bottom picture displays the probability density distribution of the negative sentiment.

Discussion

We have presented a polarization analysis using as a case study the 2019 Chilean social unrest. We aimed to identify users with particular stances and to understand how different they were based on social media data. Unlike previous works that relied on supervised methods to identify communities and manually check differences and similarities in the content, we provide a fully automated and unsupervised methodology to deal with these challenges. Our results showed that we correctly assigned users' stances with an f1-score of 0.77. Although our results were slightly lower compared to the state-of-the-art (f1-score of 0.80 on average), our method compensates by providing disentangled meaning around specific topics. Our analysis encompasses both network and content methods, diminishing the existing gap of these techniques for analyzing polarization. This suggests that user interaction and content are closer than expected, bringing additional insight into the understanding of polarized users.

Our method considered a framing assignment process to automatically find commonly discussed concepts among the communities. The benefit of this automatic selection was that it did not require prior domain knowledge expertise about the themes under study, especially when events seem powerfully dynamic and unexpected. For Twitter-based research, this could complement the use of hashtags for polarization studies, particularly when these cannot be identified for two or more communities, as well as suggesting common keywords that can be compared across groups.

Our approach to estimate topics that are jointly discussed by several groups can contribute to real-time message col-

lection for this type of events. Currently, data retrieval is performed by considering a set of initial keywords related to the target event. For dynamic analysis, this requires updating keywords based on trending topics and hashtags, which are usually estimated by global frequencies. However, this approach has difficulties in representing topics related to minority communities. In this sense, our method overcomes this drawback because the salient topics are estimated at the community level.

Our research could benefit political and social scientists in understanding how dynamic conversations on Twitter show insight into high-impact events in the real world. We observed that our studied frames, and the semantic around them, are still present in the ongoing discussion. For instance, the country's perception (Chile), institutional violence and human rights, the high adherence to manifestations, and the reactions to government actions. In addition, our study could exhibit future social connectedness and political behavior for forthcoming events. Chileans voted to draft a new Constitution a year after the movements. Results showed that 78.28% favored a new Constitution, while 21.72% rejected the change. The option "reject a new Constitution" won mainly in the three wealthiest municipalities, historically associated with right-wing electoral strongholds. This electoral analysis references the political and economic elite contrary to the reforms promoted after the 2019 social movements (Velásquez 2020). Although our method did not determine social and economic status for users, our results provide insights that communities in Twitter can reveal existing polarized groups about specific topics. Measuring and contextualizing polarization helps researchers complement traditional methods such as surveys or opinion polls, displaying the general overview of the population during social discussions around themes in a low-cost and real-time manner.

Conclusion

We present in this paper an analysis of group polarization related to the online discussions around the 2019 Chilean social unrest. The analysis is mainly motivated by the characteristics of the event, such as messages' language, event location, the impact on the Chilean society, and polarized groups in the country. Our results exemplified how frames were perceived in online communities, creating parallel realities regarding the same issue. In the case of the against community, conversations flowed around the influence of communism and violent groups in the movement, as well as the demonstration of respect for institutions such as Carabineros and the Chilean Armed Forces. In contrast, the in favor community highlighted the human rights violations inflicted by the two institutions mentioned above and focused on criticizing the government's handling of the situation in general. The implication of this analysis can accelerate understanding of current societal problems and their expressions, which are often measured by peer review questionnaires to a limited portion of citizens.

For future work, our methodology could track frames' evolution over time and measure their lifetime, especially when sub-events determine the agenda of the social unrest. Additionally, we will test our methodology in other datasets that differ in locations and languages. Furthermore, we will study more embedding models and representations. For example, introducing contextualized sentence embeddings to model tweets instead of terms.

Acknowledgments

This work has been funded by ANID FONDECYT Project 1191604. In addition, FB is partially funded by ANID FONDECYT grant 11200290 and U-Inicia VID Project UI-004/20. HS is supported by ANID / Scholarship Program / DOCTORADO BECAS CHILE/2020 - 21201101. We also acknowledge the support of ANID - Millennium Science Initiative Program - Code ICN17_002 and the National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

Aldayel, A.; and Magdy, W. 2019. Assessing sentiment of the expressed stance on social media. In *International Conference on Social Informatics*, 277–286. Springer.

AlDayel, A.; and Magdy, W. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4): 102597.

Anand, P.; Walker, M.; Abbott, R.; Tree, J. E. F.; Bowmani, R.; and Minor, M. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, 1–9.

Azjen, I. 1980. Understanding attitudes and predicting social behavior. *Englewood Cliffs*.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10): 1531–1542.

Benton, A.; and Dredze, M. 2018. Using author embeddings to improve tweet stance classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, 184–194.

Bessi, A.; Petroni, F.; Del Vicario, M.; Zollo, F.; Anagnostopoulos, A.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2016. Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics*, 225(10): 2047–2059.

Bitschnau, M.; Ader, L.; Ruedin, D.; and D'Amato, G. 2021. Politicising immigration in times of crisis: empirical evidence from Switzerland. *Journal of Ethnic and Migration Studies*, 47(17): 3864–3890.

Bitschnau, M.; Lichtenstein, D.; and Fähnrich, B. 2021. The "refugee crisis" as an opportunity structure for right-wing populist social movements: The case of PEGIDA. *Studies in Communication Sciences*, 1–13.

Borge-Holthoefer, J.; Magdy, W.; Darwish, K.; and Weber, I. 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*, 700–711.

Chong, D.; and Druckman, J. N. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10: 103–126.

Darwish, K.; Magdy, W.; Rahimi, A.; Baldwin, T.; and Abokhodair, N. 2018. Predicting online islamophopic behavior after #parisattacks. *The Journal of Web Science*, 4.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised user stance detection on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 141–152.

Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Shapiro, J.; Gentzkow, M.; and Jurafsky, D. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2970–3005.

Durán, M. F. G. 2020. *Estallido social y una nueva Constitución para Chile*. Lom Ediciones.

Fuentes, B. 2020. El Tridente del Fascismo se asoma sobre Chile. https://www.revistadefrente.cl/el-tridente-del-fascismo-se-asoma-sobre-chile/. Accessed: 2021-08-31.

Fábrega, J. 2020. ¿Despertó Chile? No todavía. https://www.ciperchile.cl/2019/10/24/desperto-chile-no-todavia/. Accessed: 2021-08-31.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Gerbaudo, P. 2017. The indignant citizen: Anti-austerity movements in southern Europe and the anti-oligarchic reclaiming of citizenship. *Social movement studies*, 16(1): 36–50.

Graells-Garrido, E.; Baeza-Yates, R.; and Lalmas, M. 2020. Every colour you are: Stance prediction and turnaround in controversial issues. In *12th ACM Conference on Web Science*, 174–183.

Guerrero-Solé, F. 2017. Community detection in political discussions on Twitter: An application of the retweet overlap network method to the Catalan process toward independence. *Social science computer review*, 35(2): 244–261.

Henríquez, M. 2020. La actual Constitución no es compatible con las demandas sociales. https://www.ciperchile.cl/2020/02/04/la-actual-constitucion-no-es-compatible-con-las-demandas-sociales/. Accessed: 2021-08-31.

Isenberg, D. J. 1986. Group polarization: A critical review and meta-analysis. *Journal of personality and social psychology*, 50(6): 1141.

Kavada, A. 2020. Creating the collective: Social media, the Occupy Movement and its constitution as a collective actor. In *Protest Technologies and Media Revolutions*. Emerald Publishing Limited.

Kutlu, M.; Darwish, K.; Bayrak, C.; Rashed, A.; and Elsayed, T. 2019. Embedding-based qualitative analysis of polarization in Turkey. *arXiv preprint arXiv:1909.10213*.

Lai, M.; Patti, V.; Ruffo, G.; and Rosso, P. 2018. Stance evolution and Twitter interactions in an Italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, 15–27. Springer.

Li, C.; Porco, A.; and Goldwasser, D. 2018. Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3728–3739.

McCombs, M. E.; and Shaw, D. L. 1993. The evolution of agenda-setting research: Twenty-five years in the marketplace of ideas. *Journal of communication*, 43(2): 58–67.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 31–41.

Navarro, F.; and Tromben, C. 2019. " Estamos en guerra contra un enemigo poderoso, implacable": los discursos de Sebastián Piñera y la revuelta popular en Chile. *Literatura y lingüística*, (40): 295–324.

Nelson, T. E.; Oxley, Z. M.; and Clawson, R. A. 1997. Toward a psychology of framing effects. *Political behavior*, 19(3): 221–246.

News, B. 2019. Protestas en Chile: La tortura, los malos tratos en comisarías y la violencia con connotación sexual son preocupantes. https://www.bbc.com/mundo/noticias-america-latina-50178678. Accessed: 2021-10-20.

Pashakhin, S. 2016. Topic modeling for frame analysis of news media. *Proceedings of the AINL FRUCT*, 103–105.

Peña-Araya, V.; Quezada, M.; Poblete, B.; and Parra, D. 2017. Gaining historical and international relations insights from social media: spatio-temporal real-world news analysis using Twitter. *EPJ Data Science*, 6: 1–35.

Poblete, B.; Guzmán, J.; Maldonado, J.; and Tobar, F. 2018. Robust detection of extreme events using Twitter: Worldwide earthquake monitoring. *IEEE Transactions on Multimedia*, 20(10): 2551–2561.

Radovic, P. 2020. ¿Qué tan polarizados estamos? https://www.latercera.com/la-tercera-domingo/noticia/que-tan-polarizados-estamos/EKIQPDVZTZAQFK2VTYK5GZBUIU/. Accessed: 2021-08-31.

Rane, H.; and Salem, S. 2012. Social media, social movements and the diffusion of ideas in the Arab uprisings. *Journal of international communication*, 18(1): 97–111.

Rossetti, G.; Milli, L.; and Cazabet, R. 2019. CDLIB: a Python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1): 1–26.

Somma, N. M.; Bargsted, M.; Disi Pavlic, R.; and Medel, R. M. 2021. No water in the oasis: the Chilean Spring of 2019–2020. *Social Movement Studies*, 20(4): 495–502.

Stefanov, P.; Darwish, K.; Atanasov, A.; and Nakov, P. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 527–537.

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

Sweeney, C.; and Najafian, M. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1662–1667.

Velásquez, F. 2020. El Apruebo en las comunas donde ganó el Rechazo, reflexionando sobre los privilegios. https://interferencia.cl/articulos/el-apruebo-en-las-

comunas-donde-gano-el-rechazo-reflexionando-sobre-los-privilegios. Accessed: 2021-08-31.

Wang, J.; Feng, Y.; Naghizade, E.; Rashidi, L.; Lim, K. H.; and Lee, K. 2018. Happiness is a choice: sentiment and activity-aware location recommendation. In *Companion Proceedings of the The Web Conference 2018*, 1401–1405.

Weber, I.; Garimella, V. R. K.; and Batayneh, A. 2013. Secular vs. Islamist polarization in Egypt on Twitter. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, 290–297.

Yang, J.; and Leskovec, J. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1): 181–213.

Ylä-Anttila, T.; Eranti, V.; and Kukkonen, A. 2018. Topic modeling as a method for frame analysis: Data mining the climate change debate in India and the USA.

Ylä-Anttila, T.; Eranti, V.; and Kukkonen, A. 2020. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media and Communication*, 17427665211023984.