Unsupervised Framing Analysis for Social Media Discourse in Polarizing Events

HERNAN SARMIENTO*, Millennium Institute for Foundational Research on Data, IMFD, Chile
RICARDO CÓRDOVA, Department of Computer Science, University of Chile, Chile
JORGE ORTIZ, Department of Computer Science, University of Chile, Chile
FELIPE BRAVO-MARQUEZ, Department of Computer Science, University of Chile, Chile, National Center for Artificial Intelligence, CENIA, Chile, and Millennium Institute for Foundational Research on Data, IMFD, Chile
MARCELO SANTOS, School of Communications, Universidad Diego Portales, Chile and Millennium Nucleus to Study Politics, Public Opinion and Media in Chile, MEPOP, Chile
SEBASTIÁN VALENZUELA, School of Communications, Pontificia Universidad Católica de Chile, Chile, Millennium Nucleus on Digital Inequalities and Opportunities, NUDOS, Chile, and Millennium Institute for Foundational Research on Data, IMFD, Chile

This study investigates the concept of *frames* in the realm of online polarization, with a focus on social media platforms. The research extends the understanding of how frames—emerging, complex, and often subtle concepts—become prominent in online conversations that are polarized. The study proposes a comprehensive methodology for identifying and characterizing these frames, integrating machine learning techniques, network analysis algorithms, and natural language processing tools. This method aims for generalizability across multiple platforms and types of user engagement. Two novel metrics, *homogeneity* and *relevancy* are introduced for the rigorous evaluation of identified frame candidates.

Grounded in several foundational presumptions, including the role of topics and multi-word expressions in framing, the study sheds light on how frames emerge and gain significance within digital communities. The research questions explored include the methods for identifying frames, the variability and significance of these frames, and the effectiveness of different computational techniques in this context.

To validate the approach, we present a case study of the 2021 Chilean presidential election, using data from both Twitter and WhatsApp platforms. This real-world application allows for the examination of how frames fluctuate in response to events and the specific mechanisms of platforms. Overall, the study makes several key contributions to the field, offering new insights and methodologies for analyzing the complexities of online polarization. It serves as groundwork for future research on the dynamics of online communities, especially those associated with distinctly polarized events.

Authors' Contact Information: Hernan Sarmiento, hernan.sarmiento@imfd.cl, Millennium Institute for Foundational Research on Data, IMFD, Santiago, Región Metropolitana, Chile; Ricardo Córdova, rcordova@dcc.uchile.cl, Department of Computer Science, University of Chile, Santiago, Región Metropolitana, Chile; Jorge Ortiz, jlortiz@dcc.uchile.cl, Department of Computer Science, University of Chile, Santiago, Región Metropolitana, Chile; Felipe Bravo-Marquez, fbravo@dcc.uchile.cl, Department of Computer Science, University of Chile, Santiago, Región Metropolitana, Chile and National Center for Artificial Intelligence, CENIA, Santiago, Región Metropolitana, Chile and Millennium Institute for Foundational Research on Data, IMFD, Santiago, Región Metropolitana, Chile; Marcelo Santos, marcelo.santos@udp.cl, School of Communications, Universidad Diego Portales, Santiago, Región Metropolitana, Chile and Millennium Nucleus to Study Politics, Public Opinion and Media in Chile, MEPOP, Santiago, Región Metropolitana, Chile; Sebastián Valenzuela, savalenz@uc.cl, School of Communications, Pontificia Universidad Católica de Chile, Santiago, Región Metropolitana, Chile and Millennium Nucleus on Digital Inequalities and Opportunities, NUDOS, Santiago, Región Metropolitana, Chile and Millennium Institute for Foundational Research on Data, IMFD, Santiago, Región Metropolitana, Chile.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

CCS Concepts: • **Information systems** → **Social networks**; • **Applied computing** → *Sociology*; • **Computing methodologies** → *Lexical semantics.*

Additional Key Words and Phrases: Framing Analysis, Polarization, Social Network Analysis

ACM Reference Format:

Hernan Sarmiento, Ricardo Córdova, Jorge Ortiz, Felipe Bravo-Marquez, Marcelo Santos, and Sebastián Valenzuela. 2024. Unsupervised Framing Analysis for Social Media Discourse in Polarizing Events. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 43 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The digital age has fundamentally transformed public discourse. Social media platforms, in particular, have become central hubs for the exchange, amplification, and fragmentation of information. While these platforms offer spaces for diverse dialogue, they also enable the proliferation of misinformation and the deepening of societal divisions. Polarization—the process by which social or political groups split into opposing sub-groups with contrasting positions, goals, and viewpoints—has been a growing concern in this context, as fewer individuals remain neutral or hold intermediate stances [29, 53, 68, 82, 87].

In recent years, research on polarization has evolved beyond simplistic explanations, such as 'filter bubbles' and 'echo chambers' [21, 47, 64], towards more nuanced approaches. One prominent explanation revolves around *framing*: the way particular interpretations of an issue, or *frames*, shape individuals' perceptions, attitudes, and behaviors [26, 36, 44, 84, 95].

In online environments, frames often emerge and spread through user-generated content, shaping collective interpretations of events and driving polarized opinions [92]. As users selectively share content, specific frames gain prominence, creating locally homogeneous views within groups, often reinforcing existing partisan perspectives [7]. Over time, as like-minded content becomes dominant in these groups, perceptions of polarization intensify [71].

To better understand these dynamics, various methods—ranging from machine learning and network analysis to content analysis—have been applied to study frames and their influence on polarization [7, 31, 33, 41, 70, 81, 91]. Yet, despite significant progress, current approaches often rely on domain expertise to pre-select relevant topics for analysis, which may limit their generalizability across platforms and contexts. There is a pressing need for methodologies that can identify frames automatically, with minimal prior knowledge and human intervention, and that are adaptable to different forms of content and modes of interaction.

This study addresses these gaps by proposing a systematic, generalizable methodology for identifying and characterizing frames in social media. Our approach leverages machine learning, network analysis, and natural language processing (NLP) techniques to uncover frames in polarized communities.

We begin by assuming that users can be grouped into distinct, non-overlapping communities based on their engagement levels on platforms such as Twitter and WhatsApp. Once these communities are delineated, we extract potential frames from user-generated content. The identification of these frames is grounded in the following key assumptions:

- Topics can serve as a proxy for frames [49, 50, 78, 101, 102].
- Topics are composed of multi-word expressions (MWEs), or collocations, which are phrases that frequently co-occur, contributing to the semantic structure of frames.
- The prominence of MWEs within communities signals important frame candidates.
- A multi-word expression must appear within at least one topic to qualify as a frame candidate.

We rigorously evaluate our proposed frame identification methodology across diverse scenarios, employing different topic modeling algorithms, a range of topic numbers, various text representations, and two key metrics—*homogeneity* and *relevancy*. These metrics assess the semantic distinctiveness of the identified frame candidates and their relevance to the ongoing debate.

To demonstrate the efficacy and limitations of our approach, we present a case study on the 2021 Chilean presidential election, a highly polarized political event marked by the stark ideological divide between Gabriel Boric (left-wing) and José Antonio Kast (right-wing). Using data from public discussions on X/Twitter and WhatsApp groups supporting different candidates, we identify and analyze key frames shaping polarization within these digital communities.

1.1 Research Questions

This study seeks to understand how *frames* circulating on social platforms can be identified, characterized, and evaluated. The specific research questions guiding this work are:

(1) How can machine learning techniques, network analysis algorithms, and natural language processing tools be utilized to identify and characterize *frames* within digital platforms?
(2) How do *frames* on digital platforms fluctuate in response to real-world events, platform-specific mechanisms, and the characteristics of user-generated content?
(3) How can we assess the significance and variability of frames?
(4) What impact do different topic modeling algorithms, topic numbers, and text representations have on the effectiveness of the proposed frame identification methodology?
(5) In the context of the 2021 Chilean Presidential Elections, how can this frame identification methodology be applied to highlight key *frames* that mattered to communities across Twitter and WhatsApp platforms?

1.2 Contribution

This research advances our understanding of *frames*—dynamic, emergent, and often covert constructs prevalent in polarized online communities. Our key contributions are as follows:

(1) We propose a comprehensive methodology for identifying and characterizing *frames* across multiple platforms. By integrating machine learning, network analysis, and natural language processing (NLP), our approach offers a generalizable framework for studying frames in polarized contexts.
(2) We account for platform-specific attributes, such as user integration within communities and modes of engagement. This enables us to derive nuanced insights into the dynamics of polarization from diverse social media platforms.
(3) We introduce two novel metrics—*homogeneity* and *relevancy*—to evaluate the quality of the identified *frames*. These metrics measure semantic divergence and relevance to the debate, offering a rigorous tool for assessing frame significance within polarized discussions.
(4) Our methodology is grounded in key assumptions about the role of topics and multi-word expressions (MWEs) in framing analysis. By leveraging these foundational concepts, we systematically identify frames and map their emergence and evolution within digital communities.
(5) We demonstrate the practical applicability of our approach through a case study on the 2021 Chilean Presidential Elections. This real-world example validates our methodology and highlights how frames shift in response to events, platform mechanisms, and user interactions.

(6) Our analysis spans two platforms, Twitter and WhatsApp, illustrating how polarization unfolds differently across them. By identifying key *frames* within each platform, we provide a comprehensive view of polarization across varying content forms and interaction modalities.

1.3 Ethical Statement

The data collected through WhatsApp groups are highly sensitive due to their political and semi-private nature. We therefore immediately anonymized sensible parts of the data (WhatsApp IDs), and all multimedia, such as images, audio, and videos, were deleted. Furthermore, the analysis performed for both Twitter and WhatsApp content only considered processed text. This means that there was no human intervention when reading and analyzing each message individually. Instead, only patterns at the topic level (topic modeling) and word embeddings were observed by the authors. This methodology ensures that individual user identities are protected and that the analysis remains focused on general patterns and trends rather than specific user content [51]. Finally, we will release the dataset only in a case-by-case scenario to avoid making individuals identifiable in the content of messages or by cross-referencing with other datasets or any future system that may harm these individuals.

2 Literature Review

This section reviews the existing studies on two significant topics: the analysis of framing in communication and the phenomenon of polarization in social media. The first subsection, *Framing Analysis*, examines the various methodologies and theoretical approaches used to understand how information is structured to influence public perception. The second subsection, *Polarization in Social Media*, investigates how social media platforms contribute to the intensification of polarized opinions and the mechanisms through which they shape public discourse.

2.1 Framing Analysis

Framing has been a central concept in multiple disciplines, including communication [61], political science [54], psychology [93], and sociology [38]. While these fields often present divergent definitions and methods, framing is widely recognized as a constructivist concept, emphasizing the role of language in shaping the social construction of reality through interaction between individuals and groups [15].

To use a popular definition among communication and political science scholars, Entman [36] states that *"to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation."* Thus, a frame can function both as a process—how frames are constructed and communicated—and as an outcome, representing the interpretative structure embedded within a text [44]. Frames help individuals interpret events, guiding their perceptions and behaviors accordingly.

A key distinction in framing analysis is the differentiation between "frame-building" and "frame-setting" [32]. Frame-building refers to how frames are constructed in media, while frame-setting addresses how these frames influence audience perceptions. This dual perspective offers insights into how media content is shaped and interpreted.

Frames are instrumental in creating structured meaning from ambiguous or complex realities, functioning as "interpretative packages" [38]. In political communication, frames are frequently deployed as strategic tools to shape public opinion or influence the interpretation of events [60]. Framing can occur at both a superficial level, where frames are applied directly to specific issues, and at a deeper level, where cultural and social frames evolve more gradually through long-term interactions.

Framing is also closely related to the second level of agenda setting, often referred to as attribute agenda setting [69]. While both framing and attribute agenda setting explore how media influences audience perceptions, framing is considered a more complex construct, focusing not only on how an object is presented but also on the underlying interpretative processes.

Traditionally, framing has been identified through manual content analysis, a method effective for capturing nuanced frames but limited by scale and subjectivity. As computational methods gain traction, automated approaches to framing analysis are increasingly used to process large datasets [39]. Matthes [66] highlights the potential of automated techniques to mitigate the limitations of manual coding, though challenges related to precision and interpretability remain.

Ali and Hassan [3] provide an extensive overview of computational framing methods, including supervised learning, unsupervised approaches like topic modeling, and advanced neural networks. They emphasize the scalability of these methods but acknowledge persistent challenges such as frame ambiguity and dataset annotation.

Supervised machine learning has shown promise in large-scale frame identification [22], and text embeddings have proven valuable for framing classification by capturing semantic relationships between words [58]. Despite these advances, most framing studies—whether manual or computational—focus on identifying either generic frames, common across topics, or thematic frames, specific to particular issues [18, 85].

Given our focus on a specific issue, we align our analysis with thematic frames. Recent research has proposed using topic modeling as a proxy for framing analysis, suggesting that the co-occurrence of words in texts can serve as indicators of how themes are framed [49, 101, 102]. As Ylä-Anttila et al. [102] notes, word-use patterns emerging from topic modeling algorithms can approximate framing patterns by highlighting how certain words are used to discuss a theme.

One of the key challenges in computational framing analysis is selecting word clusters that not only co-occur frequently but also capture the essence of a frame. This process can be resource-intensive and requires careful consideration to ensure meaningful frames are identified.

Our study addresses these challenges by incorporating Network-Activated Frames (NAF) [6], which conceptualizes framing as a dynamic process occurring within social networks. NAF posits that users frame events by sharing content within their networks, and these frames become more resonant as they align with the dominant values of the community. As users are exposed to only a subset of the broader discourse, frames tend to be locally homogeneous, amplifying certain messages while suppressing others that do not align with the community's prevailing viewpoints.

This study introduces a novel approach to framing analysis that addresses several limitations of existing computational methods. Specifically, we:

(1) Conceptualize frames as context-specific constructs emerging within distinct user communities, rather than generalized clusters of terms across entire datasets. By utilizing the NAF approach, we detect frames that resonate within particular communities based on engagement patterns.
(2) Automate the frame identification process with minimal reliance on domain knowledge. Unlike traditional methods that require pre-selected topics, our approach introduces quantitative metrics for automatic frame extraction, assuming that frames are salient concepts within a community's discourse.
(3) Expand the representation of frames by capturing them as multi-word expressions (MWEs). This allows for richer, more nuanced frame identification, providing a more accurate depiction of how frames function within polarized digital environments.

2.2 Polarization in social media

In the midst of heated debates on contentious topics, it is common for online participants to come across viewpoints mirroring their own [8]. This occurrence, combined with other factors linked to the consumption of media, strengthens users' pre-existing beliefs and restricts their exposure to differing standpoints on the debated issues [2, 16]. Therefore, a vital part of analyzing polarization involves recognizing the diverse positions within the discussion, followed by pinpointing clusters of users sharing similar viewpoints. This classification assists in assessing and tackling the issues that arise due to polarization within social networks.

Research that examines polarization on social media platforms often proceeds by classifying users based on their alignment with specific topics or entities [1]. It is typically assumed that such communities can be preliminarily recognized by their focus on certain subjects, their shared use of specific hashtags and vocabulary, and the inclusion of a group of initial users to establish these communities. However, the process of manually assigning labels beforehand can be resource-intensive, particularly in large networks. It demands significant time investment, impacts the distribution of stances in the dataset, influences the quality of inter-annotator consistency due to the necessity of topic expertise, and is hampered by the absence of definitive datasets, among other challenges.

The information in Table 1 demonstrates that numerous methods, strategies, and case studies have been examined up to this point. The literature has investigated both content-based and network-based approaches, focusing on online content (such as tweets) and the relationships between objects (like users), respectively. Furthermore, various tasks employing distinct machine learning techniques (for instance, supervised learning) have been suggested, which will be elaborated upon in the subsequent text.

There is an emphasis on examining the vocabulary used by users. Underpinning this focus is the hypothesis that individuals sharing a stance are likely to express their opinions using similar language. This idea was supported by research that found individuals taking a similar stance (e.g., opposition to abortion) frequently use analogous vocabulary to articulate their views [1, 5, 57, 73].

Authors have employed embedding learning strategies to determine users' stances by integrating both textual content and user interactions [14, 62]. In the work of Garg et al. [40], their findings revealed that word embeddings capture gender and racial stereotypes when word vectors trained on diverse corpora are compared to understand the semantic definition of a given term. Based on the results of the previous work, Kutlu et al. [59] implemented various word vector models for each political figure and stance (e.g., supportive or opposed to Erdogan). They utilized a catalog of known polarized adjectives and political terms, and compared each term's 2000 nearest neighbors' words to qualitatively understand the differences among low-dimensional vectors.

Meraz and Papacharissi [71] examine how both elites and non-elites negotiate the flow of information to establish dominant frames, focusing on the interaction between gatekeepers, who manage information dissemination, and gatewatchers, who challenge this control. The research highlights the networked environment of Twitter, showing how frames are influenced by network effects, hashtags, and interaction markers such as retweets (RT) and mentions. The study underscores the fluid, iterative processes of networked framing, where frames are persistently revised, rearticulated, and dispersed by both crowds and elites, demonstrating the dynamic mechanisms driving the propagation of dominant frames in social media.

Demszky et al. [33] 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 presidential candidates, they applied a label propagation method to determine the users' political party. They trained a word embedding

Article	Task	Approach	ML technique	Dataset/event
Aldayel and Magdy [1]	Relationship between stance and sentiment	Content-based	None (statistics analysis)	SemEval stance dataset
Meraz and Papacharissi [71]	Framing analysis	Both content-based and network-based approaches	None (network analysis and hashtags frequency)	Protests in Cairo, Egypt
Anand et al. [5]	Stance classification	Content-based. N-grams, LIWC and grammatical features	Supervised	Topics from Convenceme.net
Mohammad et al. [73]	Stance detection and classification	Content-based. N-grams	Supervised	SemEval stance dataset
Klebanov et al. [57]	Stance classification	Content-based. Vocabulary selection	Supervised	Multiple debate corpora
Benton and Dredze [14]	Stance classification	Content-based. Word embeddings	Supervised	SemEval stance dataset
Borge-Holthoefer et al. [19]	Polarization analysis	Both content-based and network-based approaches.	Supervised	Egyptian political sphere
Weber et al. [98]	Polarization analysis	Network-based. Retweet network analysis based on a set of seed users	Supervised	Egyptian political sphere
Coletto et al. [27]	Polarization analysis	Network-based. Retweet network analysis based on graph patterns	Supervised	Twitter controversial pages
Garimella et al. [41]	Polarization analysis	Both content-based and network-based approaches.	Semi-supervised	Twitter controversial pages
Demszky et al. [33]	Polarization and framing analysis	Both content-based and network-based approaches.	Semi-supervised	U.S mass shootings
Guerrero-Solé [48]	Polarization analysis	Network-based. Retweet network analysis based on a set of seed users	Unsupervised	The Catalan Referendum for Independence
Darwish et al. [31]	Stance detection	Both content-based and network-based approaches	Unsupervised	Twitter political discussions
Kutlu et al. [59]	Polarization analysis	Content-based. Word embeddings	Unsupervised	2018 Turkey elections
Sarmiento et al. [81]	Polarization and framing analysis	Both content-based and network-based approaches.	Unsupervised	The 2019 Chilean unrest

Table 1. Summarization of content-based and network-based approaches in the literature addressing polarization analysis and stance detection problems. The table is sorted by ML technique.

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.

In addition to characterize users based on the content published by them, articles in this field have also studied network relationships that emerge from diverse objects in social media. Social media platforms yield a rich interaction

structure wherein users communicate and connect with others in several ways. For example, sharing ideas from other users [19, 48], replying to messages, following others [43], and using similar specific keywords in their content [31].

In a study examining polarization in Egypt concerning Secular and Islamist perspectives across multiple languages, an analysis of the retweet network was conducted by Weber et al. [98]. Using a seed of users and employing NodeXL and Fruchterman-Reingold community algorithms, the authors revealed a polarized network representing Islamist, Secularist, and Center positions. Darwish et al. [30] investigated stance detection of online Islamophobia, using the 2015 Paris terrorist attack as a case study. Among various features, the authors took into account network characteristics such as the accounts that a user mentioned, retweeted, and responded to.

Darwish et al. [31] introduced an unsupervised framework for detecting stance on Twitter, considering three polarized events. This approach extracted multiple network features, including the number of unique tweets, hashtags, and retweeted accounts, while also computing user similarity. Subsequently, they applied dimensionality reduction and clustering techniques to identify communities. Similarly, Sarmiento et al. [81] proposed a slightly human-supervised approach to framing analysis. Considering the 2019 Chilean protests as a case study, they approached frame identification in polarized communities by selecting those topics that are almost similar across communities. Although the authors demonstrated substantial differences and similarities in the use of various concepts, their calculation lacks an in-depth analysis of the quality and homogeneity of the selected frames.

Other works have included additional structures, such as reply and follower graphs. Coletto et al. [27] studied controversial topics in Twitter, considering a motif-based approach that enriches traditional graph features (i.e., network structure and temporal characteristics) to predict if a conversation thread is controversial or not. Garimella et al. [41] proposed a graph-based three-stage pipeline to quantifying controversy in social media, which involves the creation of a conversation graph about a topic, identifying potential sides of the controversy, and measuring the amount of controversy based on the structure of the graph. In these two mentioned articles, both works require a seed of initial keywords (or topics) to analyzed controversial themes. To understand long-term polarization effects in Twitter, Garimella and Weber [43] analyzed the increasing of US political polarization over the last eight years. Their analysis relied on re-constructing retweet, followers and shared hashtags networks among others. The authors claimed that polarization increased, depending on the measure, between 10% and 20%.

Differentiation from prior work. Our method provides a comprehensive approach that leverages both social network analysis and content analysis to offer a detailed analysis of thematic frames that emerge in social media during polarized events. This dual approach ensures that we capture the complexities of both the social interactions and the content driving the discourse, making it more effective than single-method approaches.

Various works, as described in Table 1, have considered this dual-method approach. However, they often rely on supervised learning, which requires a set of seed objects [98](e.g., known users), preconceived ideas about what topics may be important over time in online discussions [41], and in several cases, pre-trained models with past data, which may not integrate the current context of conversations.

Considering these limitations and the challenges of analyzing social media platforms given that the discourse can rapidly evolve and new topics can emerge unpredictably, our unsupervised method offers significant advantages. By not relying on predefined seeds or past models, our approach remains adaptable and responsive to real-time changes in the discourse. This flexibility is crucial for accurately capturing the dynamic nature of online discussions and providing timely insights into the frames and issues that matter most to polarized communities.

Furthermore, we introduce novel metrics for rigorous evaluations. Existing methods often lack rigorous metrics for evaluating identified frames, relying on qualitative assessments and basic statistical techniques [59]. Our approach introduces two novel metrics, homogeneity and relevancy, enabling a more rigorous and quantifiable evaluation. These metrics ensure that the detected frames are both relevant to polarized communities and internally consistent.

Additionally, many of the efforts presented in previous studies focus on a single social media platform or a specific type of user engagement, limiting the generalizability of their findings [31, 33]. Such a limited focus can lead to conclusions that do not hold true across different platforms or contexts of online engagement. Our methodology, validated on both Twitter and WhatsApp during the 2021 Chilean Presidential Elections, aims for applicability across multiple platforms. By applying our methodology to different social media platforms, we ensure that our findings are applicable to various online environments.

Lastly, our framing perspective contrasts with that of Meraz and Papacharissi [71]. Whereas their work examines frame negotiation between elites and non-elites in networked settings, our approach explores organic frame emergence driven by community dynamics. Our focus lies in understanding how influential users and viral content shape frame prominence within communities, highlighting the covert and complex aspects of frames in polarized environments.

3 Proposed Methodology

The issue of polarization is a common phenomenon across social media platforms, especially evident in areas such as political discussions, debates on contentious issues like same-sex marriage and abortion, and more recently, crisis events such as COVID-19 and public protests. For the purpose of our research, we characterize polarized events as unique occurrences that unfold within a specific timeframe and location, where two or more groups of users, with contrasting views, can be recognized on a particular online platform. For instance, online discussions related to presidential elections including users from opposites parties, messages posted during demonstrations and protests that generated different points of view in people, among others. In this scenario, the users' stances can be discerned based on their interactions with other elements (e.g., users) and the content they engage with.

In our work, we aim to identify and characterize *frames*, which are emerging concepts that are relevant in a polarized discussion across communities within a particular platform. Figure 1 depicts a general overview of our proposed framework in order to capture these emerging concepts discussed in a polarized event. Our data sources include active users who share content on an online platform and engage with other entities, such as additional users. This engagement enables us to categorize these users into communities, the nature of which can fluctuate based on each platform's unique characteristics. For instance, it might be implicitly inferred based on the nature of user interactions, such as following or responding to other users, as is common on platforms like Twitter. Alternatively, community affiliation might be explicit, manifested when users integrate into specific groups like those found on Telegram or Whatsapp.

As Sarmiento et al. [81] demonstrated in their work, mapping users to communities allows for the identification of complex and unique relationships between content and communities, revealing that the vocabulary and context of a specific community around a given topic are crucial to the analysis of the framework. Once the communities are identified, we can begin analyzing conversations and salient concepts that are important for the discussion.

Our model is formally delineated as such: Let us consider a designated social media platform P_i (e.g., Twitter or Whatsapp). This platform is comprised of $u \in U_{P_i}$ active users. Each user u is assigned to a community $C_{P_i,j}$, wherein j serves as a marker for one of the discernible communities extant on P_i. This community $C_{P_i,j}$ is symbolized as a compilation of documents D, detailing the content shared by each user $u \in U_{P_i}$. We postulate that a document $d \in D$ may be characterized either as an individual message (such as a tweet or a Whatsapp message), or a concatenated

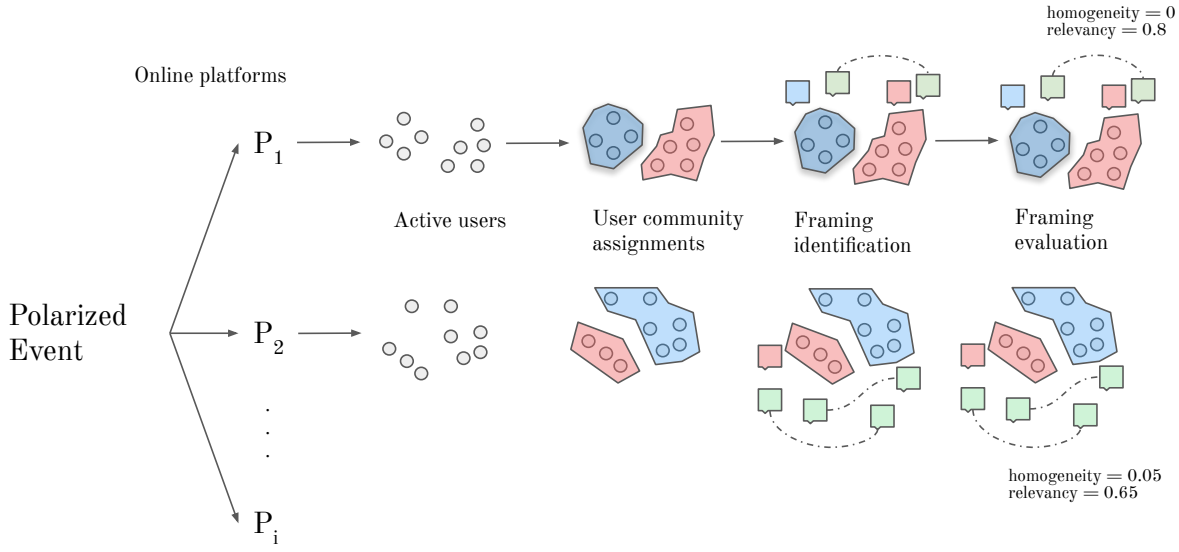

Fig. 1. Given a polarized event, we start by collecting data from various online platforms where the event is discussed. This data is then filtered to identify active users and assign them to specific communities based on their interactions. Subsequently, framing identification is conducted to detect and analyze terms that represent different frames within these communities. Finally, the identified frames are evaluated based on their homogeneity and relevance.

Notation	Description
P_i	A designated social media platform (e.g., Twitter or WhatsApp).
U_{P_i}	Active users u on platform P_i.
$C_{P_i,j}$	Community j on platform P_i.
D	Collection of documents shared by users in $C_{P_i,j}$.
d	An individual document within D.
K	Number of discrete topics within a community.
T_j	A topic within the document collection.
w_k	A multi-word expression or term.
FC	A finite set of frame candidates described by w multi-word expressions.
$P(w_k\|D,T_j)$	Probability of multi-word expression w_k in topic T_j given document collection D.
$mTP(w_k,D)$	Maximum inter-topic probability of w_k in D.
$minCommunityProb(w_k)$	Minimum inter-community probability of w_k.
$R(w_k\|D,T_j)$	Ranking of w_k in topic T_j given document collection D.
$mTR(w_k,D)$	Minimum inter-topic ranking of w_k across all topics in D.
$maxCommunityRank(w_k)$	Maximum inter-community ranking of w_k.
$homogeneity_\alpha(WV)$	Proportion of word vectors in WV with cosine similarity $\geq \alpha$.
$relevancy_\beta(FC)$	Proportion of top β frame candidates relevant in the discussion.

Table 2. Summary of mathematical notations used in our proposed methodology.

compilation of all messages posted by a user (such as the totality of a user's tweets). We subsequently categorize documents D associated with $C_{P_i,j}$ into K discrete topics, representing the most noticeable themes within a community. Using topic modeling methods to calculate these K topics, we acquire the probability $P(w_k|D,T_l)$ allocated for a

multi-word expression (or collocation) w_k in documents D to be associated with topic T_l. In our work, these multi-word expressions represent emerging, complex and compound terms that can be extracted from the documents. Given that topics can differ between communities according to content and relevance, we determine that multi-word expressions must be salient in at least one of its topics in each corpus. Thus, they can be interpreted as frame candidates, describing important concepts that naturally emerge from the conversations.

In the following sections, we describe each step of our framework, involving from the assignment of users into communities until the evaluation of the frame candidates. For a better comprehension of all the notations used in the following sections, we summarize them as shown in Table 2.

3.1 User Community Assignments

Social media platforms foster participation and communal activities among users. The concept of participation, as depicted in literature, is multifaceted, ranging from a political-ideological construct closely tied to power dynamics, to simply acting together and communicating without any political roles or power implications [25, 55, 65]. Participation transforms passive audiences into active participants or users, and this participation can be divided into explicit and implicit forms [83, 96]. Explicit participation is intimately connected with the production of user-generated content, while implicit participation involves maintaining connections and a sense of togetherness, rather than content production. While both explicit and implicit participation relate to actions that a user exhibits on a platform, they do not necessarily represent the manner in which they engage in a specific community according to their preferences.

In this section, we describe the various ways a user might interact within a distinct community on a certain kind of social media platform. We hypothesize that users may become part of a community on such platforms, either explicitly or implicitly. We define implicit engagement as a situation where users demonstrate their interests on platforms by utilizing specific common keywords in messages, such as hashtags, responding or re-sharing others' content, and following certain users. In general, implicit engagement is observed in platforms where users cannot affiliate to a particular group or community. This kind of interaction is often seen on platforms like Twitter or Instagram, where users tailor their content and interactions to reflect their individual preferences. Conversely, explicit engagement is identified by actions where a user actively joins a community through either public or private URLs. Platforms like Whatsapp or Reddit exemplify this kind of engagement, allowing users to join groups that align with their interests.

Considering the descriptions for implicit and explicit engagement, we next define the procedures to assign users into communities depending on the type of engagement found in online platforms.

3.1.1 Implicit engagement. Consider a platform, denoted as P_i, which encompasses an ensemble of active users, represented as $u \in U_{P_i}$. Unlike other platforms, P_i does not innately categorize its users into specific communities. Nonetheless, based on the principle of homophily —where individuals tend to associate with those similar to themselves— we can gather them into communities predicated on their distinct interactions.

We calculate user community affiliations by examining a graph of user interactions, designated as G_U, among users U. The interpretation of this interaction graph G_U can vary according to the unique traits of the platform P_i. Examples of such interpretations could include a network mapping followers and followees, a reply or retweet network, among others. Employing this interaction graph G_U, we can utilize a variety of methods to algorithmically derive $C_{P_i,j}$ communities, characterized by the $u \in U_{P_i}$ users. We then use community detection algorithms to obtain the assignment of each user $u \in U_{P_i}$ to a community $C_{P_i,j}$

In order to illustrate this scenario, we consider the Twitter platform as an example. On this platform, U users shared their status via text, visual, and audio-visual content. Furthermore, they interact with others in form of re-sharing content, replying messages, following users, among other similar activities. However, due to the intrinsic structure of Twitter, users are not facilitated to affiliate with a specific group or community. On the contrary, users tend to converge based on shared interests, denoted by the usage of mutual keywords or hashtags in their communications, redistribution of content resonating with their preferences, and so on. Consequently, by developing an interaction network G and implementing community detection algorithms, we could ascertain a multitude of user groups that demonstrate parallel interests.

3.1.2 Explicit engagement. Consider a platform, denoted as P_i, which encompasses an ensemble of active users, represented as $u \in U_{P_i}$. In this scenario, a community $C_{P_i,j}$ consists of a set of users $u \in U_{P_i}$, which each user u has joined to a specific public or private group by following its interests. This assignment allows us to accurately obtain the membership of each group without additional computational methods.

Literature analyzing platforms, where users explicitly join to a community, has shown that a common challenge is to primarily determine groups or communities that will be analyzed [42, 46]. This challenge is associated given the extensive number of groups that may initially be identified, which is extensive in time and domain-specific knowledge. One of the approaches to deal with this, it is to develop a *snowball* approach [9, 42]. This approach consists in manually collecting an initial set of groups in a platform, considered as *seed groups*. Next, the procedure checks whether a new *shared group* has been posted by users in the seed group. If this *shared group* does not appear in the collection, it adds to the sample. As previous works in this field, we considered that each *shared group* belongs to the same community (e.g., political affiliation or stance) as the original *seed group*.

To illustrate this scenario, we examine the features of WhatsApp. This platform allows users to engage in one-to-one communication through chats, as well as participate in public or private groups for many-to-many communication. When it comes to public groups, it is assumed that users join based on their shared interests related to the group's theme. Consequently, the content shared within the group aligns closely with the stance of its members. We refer to publicly accessible groups as *seed groups*, while any new, unseen groups that are introduced within the seed group are termed *shared groups*.

3.2 Framing Identification

In this section, we explore two approaches for identifying frames in online content. Our primary motivation for pursuing this study is to develop a nuanced understanding of how topics are framed in online discourse, which can have far-reaching implications for public opinion and policy decisions.

As discussed in Section 2.1, there exists a body of work suggesting that the outputs of topic modeling algorithms can serve as indicators of various methods for discussing a particular theme [102]. The core idea is that each identified topic consists of clusters of frequently co-occurring words within documents, which are distinct from other topics, and which in turn represent operationalized frames that capture the key themes of the discussion.

However, one of the challenges lies in selecting groups of words that not only appear together frequently but also effectively capture the essence of a specific frame within the discussion. This process is computationally intensive, requiring substantial resources to analyze large volumes of data and identify these specific and meaningful word combinations.

In contrast to previous research that views topics as collections of individual words extracted from content, we argue that topics are composed of a series of frames, which can be described as collocations. According to Benson [13], collocations are defined as word combinations that occur together more frequently than by chance. In Natural Language Processing (NLP), the task of collocation extraction aim to identify in a corpus complex lexical items, often characterized as unpredictable, idiosyncratic, holistic and mutually selective [86]. This perspective implies that more intricate multi-word expressions or collocations[1] may better characterize a topic, thus providing semantic enrichment for identifying the frame. For example, a person's full name, or a compound of previously unseen complex terms, may emerge as a significant concept or frame within a topic.

We describe the framing identification process as follows: Consider a collection of documents $d_1, d_2, \ldots, d_n \in D$, where D is the content shared by $u \in U_{P_i}$ active users belong to the community $C_{P_i,j}$ in a platform P_i. Our goal is to identify a finite number of frame candidates FC, which are described by w multi-word expressions (or collocations) relevant in a discussion in at least one theme across different communities.

Each collection of documents D can be stratified into K topics, which represents the most salient themes in a community. Using topic modeling techniques to compute these K topics, we derive probabilities to understand the relevance of multi-word expressions in a given topic. Specifically, the probability $P(w_k|T_j)$ of a multi-word expression w_k pertaining to topic T_j is calculated. It is imperative to clarify that D, the collection of documents upon which the model is trained, does not constitute a random variable in this context. Therefore, we denote the relationship with the collection of documents D in our probabilities as $P(w_k|D, T_j)$. This allow us to indicate that these probabilities are conditional on the documents D being fixed. Finally, given that these K topics may be composed of a different number of w multi-word expressions and several degrees of relevance for a community, we determine that frame candidates FC are salient w multi-word expressions in at least one of its topics in each community. We next describe two approaches to compute frame candidates that are discussed in online conversations using topic models.

3.2.1 Probabilistic-based approach. This approach is based on the assumption that the probability $P(w_k|D, T_j)$ of a multi-word expression w_k to pertain in topic T_j is directly related with its relevance in the collection of documents D. Therefore, the model quantifies the relevance of a multi-word expression w_k within a specific topic T_j by calculating its conditional probability based on a fixed collection of documents D.

To understand the relevance of a term across multiples topics in community documents, we define the *maximum inter topic probability* (mTP) for a multi-word expression w_k in D as the maximum (or higher) probability obtained for w_k across all topics in D. Formally, this can be defined as follows:

$$mTP(w_k, D) = \max_{T_j \in D} \begin{cases} P(w_k|D, T_j) & \text{if } w_k \in T_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

By computing the mTP score for all multi-word expressions in all documents, we obtained the relevance of them in each community document. Given our interest is to determine if a multi-word expression w_k can be considered part of the frame candidates FC, we estimate the *minimum inter community probability* $minCommunityProb(w_k)$ as the minimum topic probability for all communities. This expression determines the worst community-level probability between all best-topic level probabilities for w_k. The $minCommunityProb(w_k)$ can be formally defined as follows:

$$minCommunityProb(w_k) = \min_{D} mTP(w_k, D) \tag{2}$$

[1]Throughout the rest of the document, multi-word expressions and collocations are used interchangeably

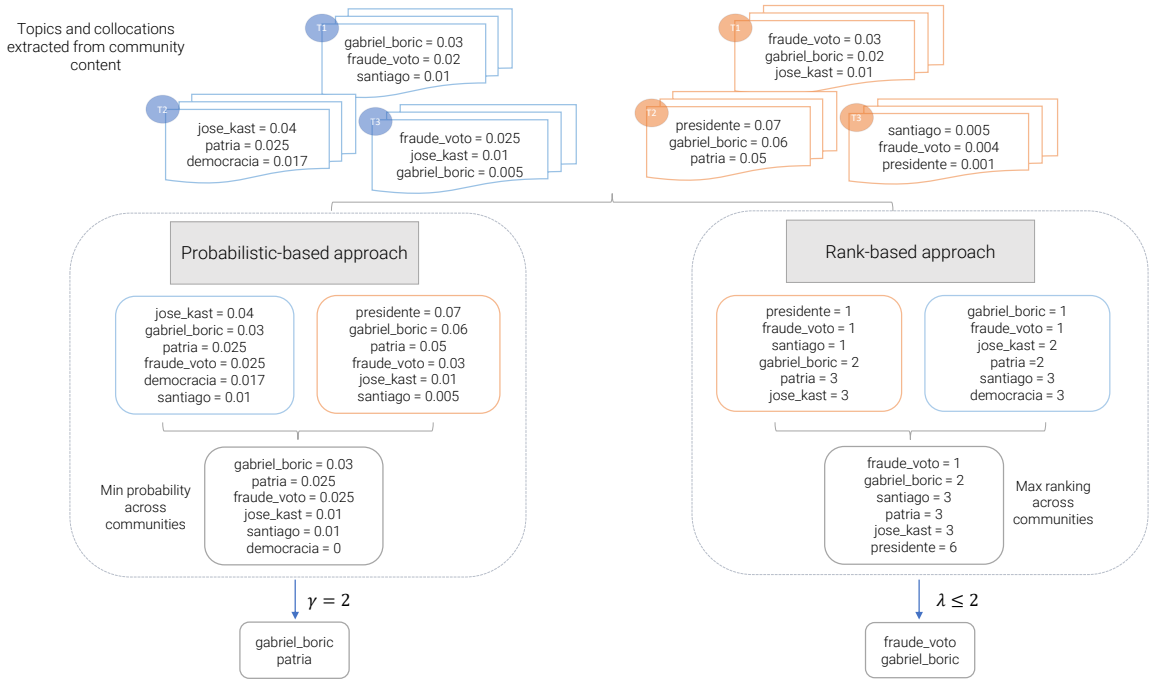

Fig. 2. An example of the framing identification approaches based on topics extracted from two distinct communities (represented in blue and orange). Topics are characterized through a series of word collocations and associated probabilities. The resulting frame candidates are determined by the specific approach employed and its hyper-parameters.

Finally, the set of frame candidates FC are obtained by selecting top γ multi-word expressions w_k sorted decreasingly by their $minCommunityProb$, where γ is tuned to obtain a desired number of frames (e.g., 10). In the case of ties in the probability values of two or more multi-word expressions, one of them is chosen randomly.

Figure 2 illustrates an example of the probabilistic-based approach. In this example, $j = 3$ topics are estimated using the document corpus from each community. It is observed that the multi-word expression *gabriel_boric* appears in two different topics within each community (blue and orange document sets), and it shows varying probabilities across topics. To apply the probabilistic-based approach, we first identify the highest probability for every multi-word expression within each community; this corresponds to mTP as defined in Equation 1. Subsequently, we calculate the $minCommunityProb$ between the communities and select the term *gabriel_boric* as a *frame candidate*. This is because it ranks among the top-2 most probable terms in the communities when $\gamma = 2$. Similarly, the term *patria (homeland)* is also chosen as a frame candidate using this method.

3.2.2 Rank-based approach. Unlike the previous approach, the rank-based approach considers the collocation-level position in a topic T_i. We convert the probability $P(w_k|D, T_j)$ into a ranking $R(w_k|D, T_j)$, which is easier to interpret and more robust to outliers; i.e., the lowest ranking inside a topic represents w_k with the highest probability inside said topic. Hence, the transformation of $P(w_k|D, T_j)$ to a ranking metric $R(w_k|D, T_j)$ serves to enhance the interpretability of multi-word expression relevance in topic T_i, while also offering greater resilience to outliers.

We define the *minimum inter topic ranking* (*mTR*) as the minimum (or best) ranking obtained for w_k across all topics in D. This score represents the relevance of a collocation in a corpus under the assumption that relevancy in a collection of documents is to be salient in at least one of its topics by measuring its topic model word ranking. Formally, this can be mathematically defined by:

$$mTR(w_k, D) = \min_{T_j \in D} \begin{cases} R(w_k|D, T_j) & \text{if } w_k \in T_j \\ |T_j| + 1 & \text{otherwise} \end{cases} \tag{3}$$

Given that a multi-word expression might not be present in a topic T_j, we use the notation $|T_j| + 1$ as a default ranking for these terms. Assigning a rank of $|T_j| + 1$ to terms not present in a topic T_j ensures they are ranked lower (i.e., less relevant) than any term that is present in the topic.

By computing the *mTR* score for all multi-word expressions in all community content, we obtained their ranking in each community. To be a part of the frame candidates *FC*, w_k must also be relevant in at least one topic for all communities. With this goal in mind we compute the *maximum inter community ranking* $maxCommunityRank(w_k)$ as the maximum (or worse) minimum topic ranking for all communities. This score can be interpreted as the worst community-level ranking between all best-topic level rankings for a multi-word expression. Therefore, to obtain a good position (low ranking) here, w_k needs to have a good ranking in at least one topic for all communities. The $maxCommunityRank(w_k)$ can be formally defined by:

$$maxCommunityRank(w_k) = \max_{D} mTR(w_k, D) \tag{4}$$

The final frame candidates *FC* are obtained by selecting all multi-word expressions in which $maxCommunityRank(w_k) \leq \lambda$ where λ is tuned to obtain a desired number of frame candidates(e.g., 10). In the case of ties in the ranking values of two or more multi-word expressions, one of them is chosen randomly.

Figure 2 illustrates an example of a rank-based approach, in which the same topics, collocations, and associated probabilities are extracted from document set D, as in the probabilistic-based approach. Unlike the probabilistic method, the rank-based strategy focuses on the position (or ranking) of each collocation within the extracted topics from a given corpus. Consequently, a collocation with a high probability might not occupy a top ranking within a community if other collocations are better positioned within a particular topic. When converting each probability into a ranking, we note that the term *fraude_voto (vote_fraud)* achieves a $mTR = 1$ ranking in both communities under study. Thus, when calculating *maxCommunityRank*, this term is selected as a *frame candidate* because it has the highest ranking (i.e., the lowest numerical value) across communities. Although the probability of this collocation is considerably lower compared to others across various topics, it still secures one of the top ranks when we compute both the *mTR* and the *maxCommunityRank*. Therefore, the collocation qualifies as a frame candidate when $\lambda \leq 2$.

3.3 Framing Evaluation

This section aims to assess the effectiveness of approaches for choosing frame candidates. In general, we focus on the evaluation of two main tasks (see Figure 3). First, we determine which topics best represent the content of each community. We find the best settings considering various topics model algorithms, number of topics and document representations that maximize the coherence score. This coherence score described how well a topic is supported by a referenced corpus.

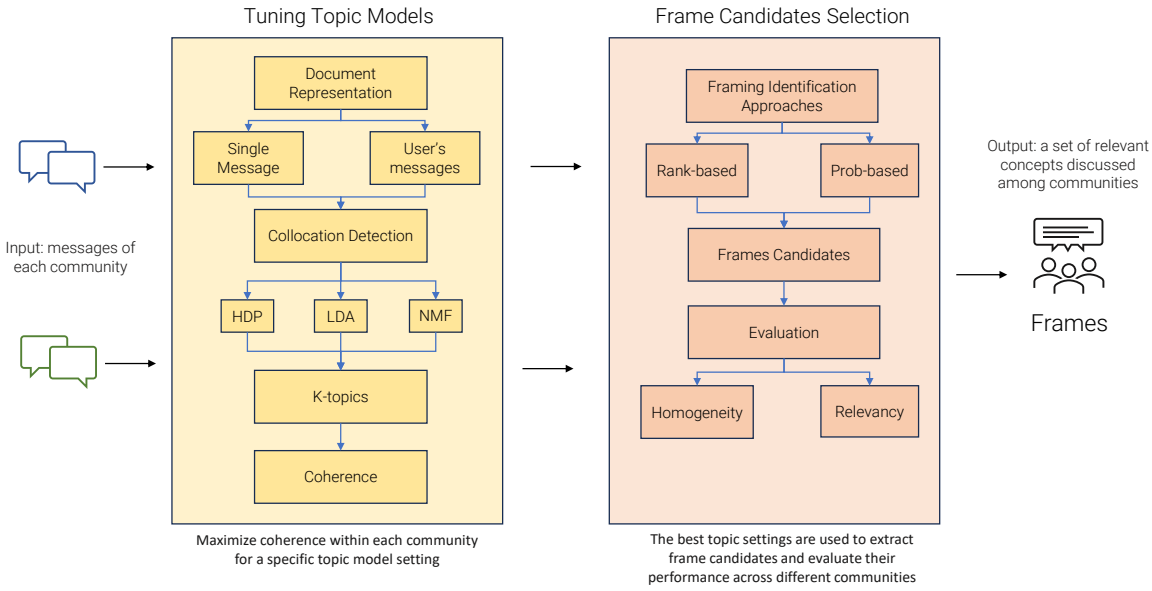

Fig. 3. A general overview of the proposed framing evaluation. This consists on two phases: tuning topic models and the selection of frame candidates.

We address the second part of the evaluation by determining the quality of the frames candidates. They are estimated based on two proposed approaches presented in Section 3.2. Additionally, we propose two metrics that allow us to semantically quantify the variety and importance of the frame candidates, namely homogeneity and relevancy.

3.3.1 Tuning Topic Models. To extract topics from each community content, we use three topic modeling algorithms: the Hierarchical Dirichlet Process (HDP) [97], Latent Dirichlet Allocation (LDA) [52] and Non-Negative Matrix Factorization (NMF) [103]. For each combination of a document representation and topic modelling, we tune them varying the number of topics that maximize the coherence score. Finally, we choose the best number of topics of each model, extract frame candidates *FC* using the two framing identification approaches, and compare them using two proposed metrics: homogeneity and relevancy scores.

We evaluated every topic model - derived from the combination of a topic modeling algorithm and the document representation - in a range of $K \in [2, 50]$ topics composed of $1,000$ words and chose that which maximizes the coherence score. To illustrate this scenario, we may choose the *LDA* algorithm and a document representation by user, for which the coherence score is computed multiples times by varying the number of topics. Hence, to determine the best number of topics for every model, we select the one that obtains the highest value.

3.3.2 Frames Candidates Selection. We extract and evaluate the set of frames candidates *FC* from the topic models by using the two framing identification approaches. We obtained 12 configurations, which represent different topic modeling algorithms, document representations and framing identification approaches. For instance, we consider the LDA algorithm, all messages aggregated by users for the document representation and the rank-based approach for framing identification. Finally, we determine the quality of the frame candidates by proposing two scores that measure semantically how varied the words are (called homogeneity) and the relevance of them (named relevancy).

Homogeneity. We define the $homogeneity_\alpha$ as the proportion of all pairs of distinct frame candidates $(fc_i, fc_j) \in FC$ that exceed (or are equal to) a certain similarity threshold α. The *homogeneity* quantifies how much semantically different (or similar) the frame candidates are. Hence, a lower $homogeneity_\alpha$ value represents a higher diversity in the final set of frames.

To estimate this score, we first trained a word2vec model $w2v_D$ using the collection of documents D. The word2vec model allows us to represent every word as a low-dimensional space vector and compute operations over it (e.g., euclidean distance and vector similarity). With this in mind, we computed the cosine similarity $sim(v_i, v_j)$ of a pair of frame candidates $(fc_i, fc_j) \in FC$, which are represented as a pair of word vectors $(wv_i, wv_j) \in WV$, where WV represents the whole set of frame candidates vectors. Finally, we estimate the proportion of every distinct pair of word vectors (wv_i, wv_j) that obtains a $sim(wv_i, wv_j) \geq \alpha$ in the word2vec model. We formally defined the *homogeneity* as follows:

$$homogeneity_\alpha(WV) = \frac{\sum_{(wv_i, wv_j) \in WV, i<j} \begin{cases} 1 & \text{if } sim(wv_i, wv_j) \geq \alpha \\ 0 & \text{otherwise} \end{cases}}{\frac{|WV|(|WV|-1)}{2}} \tag{5}$$

Relevancy. We define the $relevancy_\beta$ as the proportion of β frame candidates that are relevant in the discussion. Therefore, a higher *relevancy* represents that the set of frame candidates are more significant across communities. We estimated this score by calculating the number of extracted frame candidates FC that appear in a ground truth collection of labeled terms. The accuracy and reliability of this ground truth collection were crucial, and manual labeling played a significant role in this validation process. We categorized labels as *relevant* and *not relevant* for the discussion. To create the ground truth collection, we selected each topic model's top 30 most probable collocations and assigned them one of the mentioned labels. This set of collocations was composed of 206 unique multi-word expressions.

We hired five annotators who were familiar with the event covered by the data to perform the labeling task. Their domain expertise ensured that the frames selected as ground truth were accurately validated, reflecting a detailed understanding of the context and content. Manual labeling was essential in distinguishing between relevant and non-relevant frames, providing a solid foundation for our ground truth dataset. This process not only enhanced the precision of our relevancy measurements but also ensured that our frame identification was contextually accurate and meaningful.

To evaluate agreement among the annotators, we used majority voting and subsequently calculated the Fleiss' kappa score κ to assess the reliability of the agreement. Unlike other kappa statistics such as Cohen's kappa, which are applicable only for evaluating agreement between two raters, Fleiss' kappa is suitable for assessing three or more annotators on either ordinal or nominal (categorical) scale data. As a result, we determined a Fleiss' kappa score of $\kappa = 0.4879$, which can be considered as moderate agreement, given the number of annotators involved in the labeling process. This score underscores the importance of the manual labeling effort in ensuring that the selected frames are reliably and consistently categorized.

To calculate the $relevancy_\beta$, we selected the top β frame candidates of each topic model and estimated the proportion of them that appears as *relevant* in our ground truth collection. This top β frame candidates are selected based on criteria of each framing identification approach. In the case of the probability-based approach, we choose β frame candidates depending on highest probable collocations. For the ranked-based approach, they are selected considering the top β ranked frames.

We formally define this score as follows:

$$relevancy_{\beta}(FC) = \frac{1}{|FC|} \sum_{i=1, fc_i \in FC}^{\beta} \begin{cases} 1 & \text{if } is\_relevant(fc_i) \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where $is\_relevant(fc_i)$ checks if the frame candidate fc_i was labeled as relevant in our gold-standard collection of terms.

4 Case of study: The 2021 Chilean Presidential Election

To evaluate our proposed methodology for identifying frames in online polarized discussions, we conducted an analysis of the 2021 run-off presidential election in Chile. This election pitted the far-right candidate José Antonio Kast against the left-wing candidate Gabriel Boric.

We selected this election because it stood out as the most polarized in over 30 years. None of the centrist coalitions made it to the run-off [11]. Kast ran on a right-wing platform echoing the issue positions of Donald Trump in the USA and Jair Bolsonaro in Brazil. His program proposed abolishing Chile's Ministry of Women and Gender Equity, constructing a barrier at the northern border to halt immigration, implementing significant tax cuts, and prohibiting all abortion methods [12, 37].

In opposition to Kast was Boric, a former student leader and congressman who led a coalition of left-wing parties and movements known as the "Frente Amplio", which included the Communist Party. His campaign championed a progressive agenda that involved replacing the privately dominated health and pensions system with a public welfare system, raising corporate taxes, and adopting a comprehensive set of feminist policies [10, 35].

Considering the substantial ideological gap between these two campaigns, it is a compelling case for applying the methods outlined in this article. Furthermore, there is evidence that the Kast campaign relied on a polarizing discourse through the use of virulent language and social bots that "intoxicated" the run-off [89, 90], leading analysts to tag the rightist campaign as "dirty" [17, 76].

In order to understand this phenomenon from different social media sources, we further analyze two platforms that differ in the type of engagement they have. Such platforms are Twitter and Whatsapp, which have an implicit and explicit engagement, respectively.

4.1 Data Collection

For both Twitter and WhatsApp platforms, we retrieved data from November 03, 2021, until the runoff, which occurred on December 20, 2021. Figure 4 displays the normalized frequency of messages obtained for each social media platform. In the case of Twitter, frequency considers the original message (tweets) and the re-shared one (retweets). As the plot shows, the frequency of Twitter was highest near November 21, the day the first round was held. In contrast, the volume of Whatsapp messages appears lower before the first round. However, it increased to reach the maximum value near the second round, on December 20.

In the next sections, we provide additional explanation about the data collection process developed for each platform.

4.1.1 Twitter data extraction. We collected Twitter data using the API[2] provided by the platform. We first retrieved all messages that mentioned the candidates' user accounts. Second, we considered a set of hashtags that refer to at least one of the candidates, mention a political party, or are related to the news about the presential election. For instance, terms such as `#Kast2022`, `#BoricPresidente` and `#Presidenciales2021`. By merging both collections, we obtained

[2]https://developer.twitter.com/en/docs/twitter-api

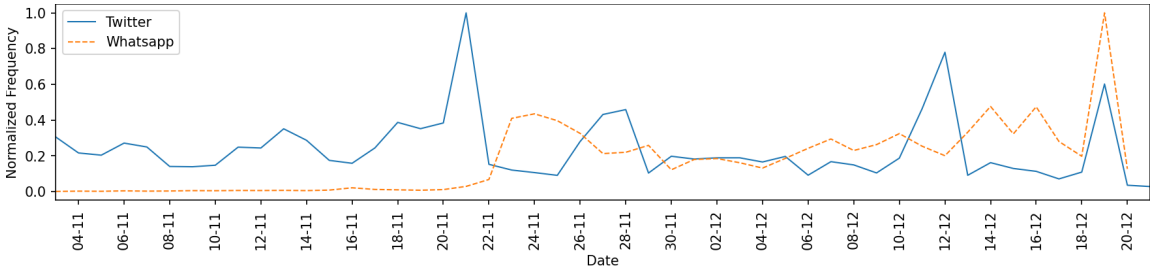

Fig. 4. Max-min normalized frequency of Twitter (tweets and retweets) and WhatsApp messages over the collection period.

a dataset comprising 6.2 million messages, of which 4.1 million and 2.1 million correspond to retweets and tweets, respectively.

4.1.2 WhatsApp data extraction. We manually collected a set of political groups - associated with one of the candidates - that can be accessed via public links shared in other social networks (e.g., Facebook and Twitter) or accessible by search engines. We refer to this set as *seed groups*. Using these *seed groups*, we applied a snowball approach similar to that of Baumgartner et al. [9]. This process checks if a new group was shared on chat and then joined it. We refer to these groups as *shared groups*.

In order to gather information from these groups, we joined partisan chat groups that were centered around the 2021 Chilean presidential election, using a singular anonymous account. We gathered public messages from the groups that this anonymous account was a part of, employing automated techniques akin to those used by Baumgartner et al. [9], to extract data from this platform.

Given that users in each group may share links to other, more specific groups, we kept those that included a profile title and description associated with a particular partisanship. For instance, the group named "*BIOBÍO UNIDO X KAST* (Biobío united for kast)" described an online community that supports the candidate J.A. Kast from a specific geographic location (Biobío state in this example). In contrast, we removed shared groups related to religious associations, buy-and-sell communities, and hobbies groups that do not contain political descriptions in their profiles. Finally, our WhatsApp dataset comprised 705,413 messages shared by 28,048 unique users in a total of 492 groups (see Figure 5).

The frequency of messages for each candidate group, as illustrated in Figure 5, followed a similar trend post the initial round, around November 21. An interesting observation about these frequencies is the considerable increase in the data volume within Boric's groups after the first round. This occurrence can be linked to the unexpected and successful results achieved by the right-wing candidate, José Antonio Kast [11]. This could have triggered a higher participation of groups associated with the Boric candidate, increasing the volume of data, and sharing new and unpublished groups.

4.2 User Community Assignments

We first identify user communities in both online platforms to analyze topics discussed among polarized groups.

4.2.1 Twitter Communities. As mentioned in Section 3.1.1, implicit community interaction can be noticed on platforms like Twitter. Since users cannot to join groups or communities, we relied on community detection algorithms to pinpoint sets of users that belong to a specific, unique group. In this context, we adopted a methodology similar to the one outlined by Sarmiento et al. [81] on identifying such communities. In their research, they utilized the user retweet network interaction to pinpoint polarized groups or political parties. In our study, the initial retweet network consisted

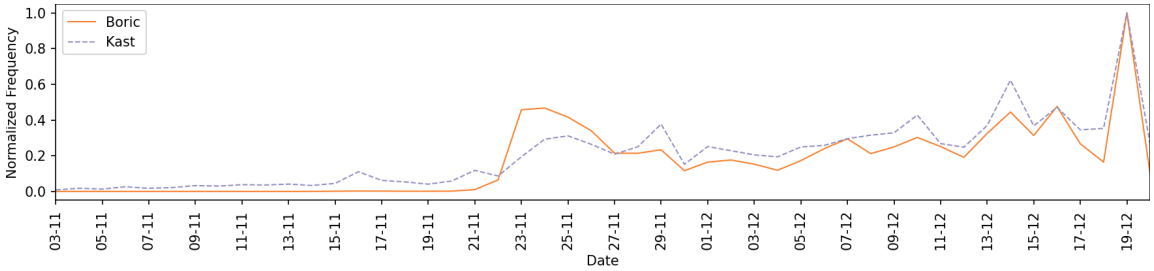

Fig. 5. Max-min normalized frequency for Gabriel Boric and José Antonio Kast in Whatsapp messages.

Community	# of messages	# of unique users
Cluster 1	246,871	4,072
Cluster 2	292,872	3,824

Table 3. Dataset distribution by Twitter communities after applying the SBM community detection algorithm.

Political affiliation	# of messages	# of unique users	# of groups
Pro-Boric	555,764	23,034	419
Pro-Kast	149,649	5,014	73

Table 4. Dataset distribution by Whatsapp communities.

of 4.1 million messages and around 251, 000 unique users. As indicated by previous research, the retweet network can offer a more in-depth understanding of the principle of homophily in social networks than mention or reply networks. This can be attributed to the fact that the retweet network tends to create more segregated communities. The latter networks can act as channels through which users encounter information and viewpoints they might not have chosen to engage with initially [28, 100].

Previous to applying community detection algorithms, the work of Sarmiento et al. [81] filtered the retweet network to remove weak connections among users (e.g., a minimum number of retweets received). Following the exact steps as the authors applied, our consolidated retweet network comprised 49, 296 users (nodes) that represent 164, 630 connections (edges) among them. Additionally, they evaluated multiple structural metrics for different community detection algorithms, obtaining that the Stochastic Block Model (SBM) had the best performance. Hence, we applied the SBM algorithm in our users retweet network.

Figure 6 shows the retweet network by displaying detected communities according the results of the community detection algorithm. Each circle represents a user and the color represents the community to which it belongs. As noted, there are two separated user clusters composed of 26, 438 and 22, 858 users.

Given that our interest is in the content published by users, we consider only those who posted messages (tweets) in our network instead of those who only re-shared (retweeted) content from other users. These users correspond to 7, 906 accounts, which were distributed in communities of 4, 072 and 3, 824 unique users, who posted 246, 871 and 292, 872 messages (tweets) respectively (see Figure **??** and Table 3).

4.2.2 Whatsapp Communities. In the case of WhatsApp, the retrieved metadata does not allow obtaining information about the users interaction. For instance, it is not possible to know if a user is on the contact list numbers of another

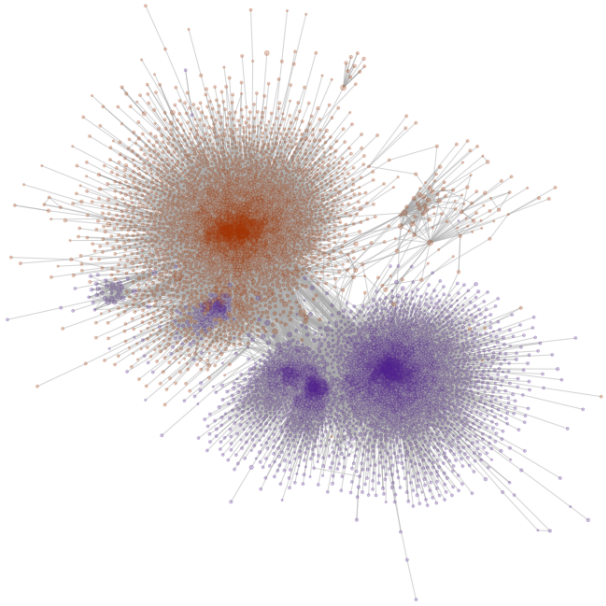

Fig. 6. Identified communities in the Twitter retweet network using the Stochastic Block Model algorithm (SBM).

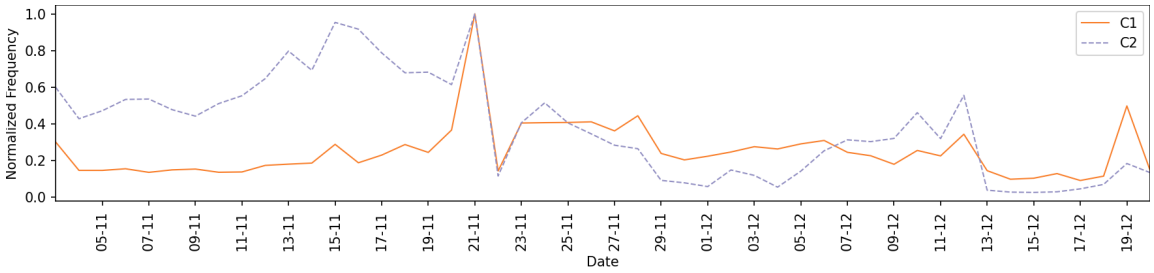

Fig. 7. Max-min normalized frequency for Twitter messages (tweets) obtained in each community.

user or who seen a specific message. Given that we obtained WhatsApp messages using a snowball approach, we considered that each *shared group* belongs to the same community (political affiliation) as the original *seed group*. As Table 4 shows, most of the collected groups in our snowball process are pro-Boric, which outnumber the pro-Kast by almost 6 times. Although the number pro-Boric groups was considerably higher, we noted that the ratio between community messages and users distribution were slower (about 3 and 4 times respectively).

4.3 Framing Evaluation

We present the evaluation of the proposed framing identification on Twitter and Whatsapp datasets. Given the diverse and unstructured nature of data on these platforms, preprocessing is essential to reduce noise and standardize text, thereby enhancing the accuracy of our framing analysis. We first preprocess text by removing accents, URLs, Spanish

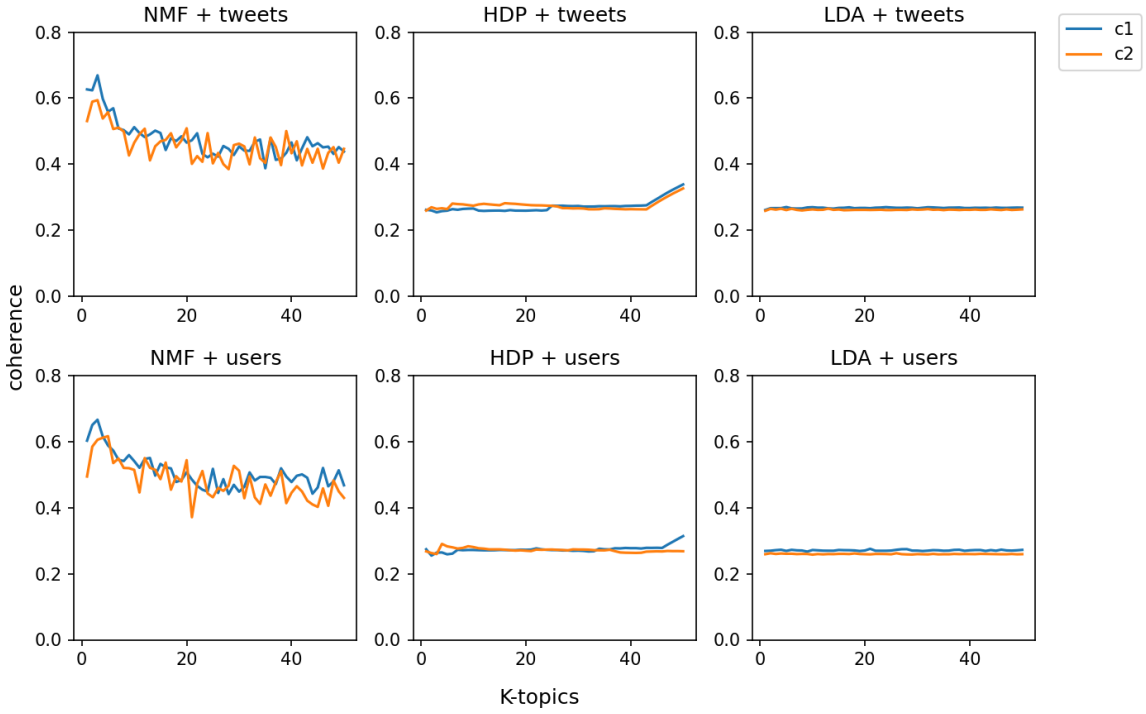

Fig. 8. Coherence scores for various topic models on Twitter, each combining a topic modeling algorithm and document representation, evaluated across 2 to 50 topics.

stopwords, hashtags, user mentions, punctuation, emojis, emoticons, and numbers and converting them to lowercase. Furthermore, we kept tokens with more than three characters, and sentences (messages in our case) with more than three tokens. Additionally, we considered that a document can be represented as an individual message or the concatenation of all messages shared by a users.

We outlined how topics are generally formed from multi-word expressions, rather than merely isolated words. In order to extract these lexical units, we utilized the phrase detection method from Gensim[3], which automatically identifies common multi-word expressions within a sequence of sentences. This model was developed based on the concept of learning phrases proposed by Mikolov et al. [72]. According to this concept, the meaning derived from a text (for example, a sentence) is not merely an aggregation of the meanings of its separate words. Consequently, the model is capable of discovering words that frequently appear in conjunction but rarely appear in other contexts. For instance, the model would identify "New York Times" as a single entity rather than as three separate words: "New," "York," and "Times."

We consider all mentioned models and settings for identifying frame candidates, as mentioned in Section 3.3. To better assess our method's effectiveness and limits, we tested it with different settings. For the scores we are calling $homogeneity_{\alpha}$ and $relevancy_{\beta}$, we tried out different values: $\alpha = [0.6, 0.7, 0.8, 0.9]$ and $\beta = [5, 10, 15, 20, 25, 30]$. This gave us a clearer picture of how our approach worked under various conditions. We next detail this task divided into each platform.

[3]https://radimrehurek.com/gensim/models/phrases.html

Algorithm	Doc. Representation	nº Topics c1	nº Topics c2	Coherence c1	Coherence c2	Avg. Topics	Avg. Coherence
HDP	tweets	50	50	0.3386	0.3265	50	0.3326
HDP	users	50	4	0.3139	0.2903	27	0.3021
LDA	tweets	5	13	0.2702	0.2654	9	0.2678
LDA	users	21	26	0.2754	0.2620	23.5	0.2687
NMF	tweets	3	3	0.6689	0.5937	3	0.6313
NMF	users	3	5	0.6662	0.6164	4	0.6413

Table 5. Optimal coherence scores for each topic model on Twitter, distinguished by their topic modeling algorithm, document representation, and ideal topic count for maximizing coherence in community documents.

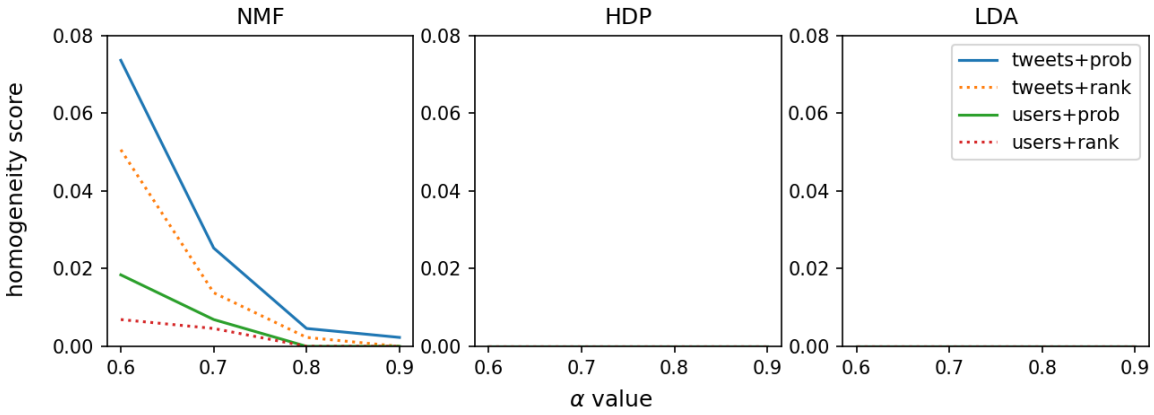

Fig. 9. Homogeneity scores of frame candidates in Twitter content using probabilistic-based and rank-based approaches, evaluated for an α range of 0.6 to 0.9. Homegeinity scores equal to zero represent higher diversity values.

4.3.1 Twitter dataset. We identified two communities comprising of 4, 072 and 3, 824 unique users, who collectively contributed 246, 871 and 291, 872 messages, respectively. We applied three topic modeling algorithms (HDP, LDA and NMF) to each set of documents - represented as individual messages and by users - associated with each community. Additionally, we evaluated the performance of the models and their setting by measuring coherence scores.

The coherence scores for each algorithm and community are depicted in Figure 8, where we found that the NMF algorithm performed better than the other algorithms, regardless of the document representation. Table 5 summarizes the best configuration for each algorithm in terms of the number of topics and average coherence obtained in the communities. We found that on average, 19 topics resulted in a coherence score of 0.4073. Moreover, the NMF algorithm outperformed the other algorithms in both levels of document representation, while LDA performed the worst. In addition, NMF achieved its highest coherence scores with fewer than 10 topics.

We next computed $homogeneity_{\alpha}$ and $relevancy_{\beta}$ scores by considering the best setting of the topic models and including the two proposed framing identification approaches for each setting. Figure 9 shows the homogeneity score divided into topic modelling algorithms in various levels of document representations and framing identification approaches. Our first observation is that the unique models that did not obtain $homogeneity = 0$ (higher homogeneity in the set of frame candidates) were those using the NMF algorithm. For the other models, we noted that they obtained the lowest score, representing that frame candidates are semantically different in the word2vec model, even for highest threshold values.

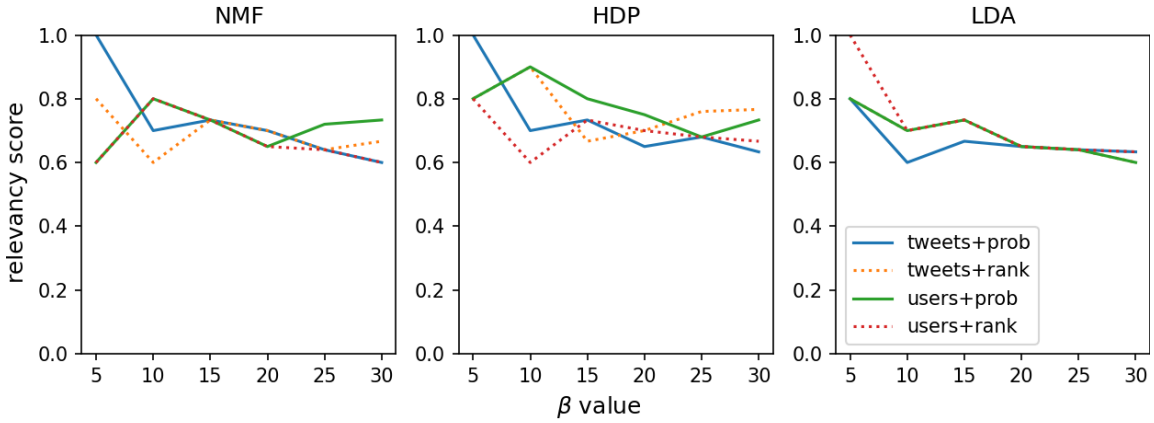

Fig. 10. Relevancy scores of frame candidates in Twitter content using probabilistic-based and rank-based approaches, evaluated for an β range of 5 to 30.

We further estimated the relevancy score (see Figure 10) in the same scenarios as previously explained. When we choose $\beta = 5$ as the number of top frames that may be relevant in our ground truth, we notice that the average relevance is about of 0.81 ($max = 1$ and $min = 0.6$). In detail, one interesting observation is that, representing every document as an individual tweet, it obtained a $relevancy_5 = 1$ for all topic modelling algorithms (HDP, LDA, and NMF). In the case of the framing identification method, the probabilistic-based approach appeared in two-third of these three best scenarios.

We further note that results change considerably for $\beta > 5$. For instance, we observe that the average relevancy decreased until 0.725 when $\beta \in [10, 15]$, while for $\beta \in [20, 25, 30]$ the average decay to approximately 0.657. Regarding the best settings, we notice that there are two topic models more consistent in terms of the maximum relevancy found across this range of values. Both mentioned models are the HDP algorithm at the level of document represenation by tweet, but with a different framing identification approach. In detail, they obtained an average relevancy score of 0.758 and 0.772 for the rank-based and probabilistic-based approaches, respectively.

Finally, by analyzing both homogeneity and relevancy scores, we suggest that the election of a specific setting may mostly depend on the number of relevant *candidates frames* found in the collection (β parameter) rather than the threshold of similarity (α parameter). On the one hand, for the three best settings with $\beta = 5$, two of them had zero values for all range of α values. These two settings considered the HDP and LDA algorithms, both representing documents as a single message, but differing in the framing identification approach. On the other hand, for the two best settings when $\beta > 5$ (both using HDP), we notice that they also obtained zero values for all the range of α values. Additionally, we note that representing documents as individual messages leads in this scenario, and the probabilistic-based approach was slightly higher than the rank-based approach.

Overall, our results suggest that representing documents as a single message had a better performance. In addittion, the HDP seems one of the best topic modelling algorithms independent of the level of document representation and framing identification approach.

4.3.2 Whatsapp dataset. Our communities were composed of 23, 034 and 5, 014 unique users that shared 555, 764 and 149, 649 messages in Pro-Boric and Pro-Kast Whatsapp groups, respectively. In addition, we follow a similar evaluation as the presented for Twitter to compute coherence score in wide range of topic model settings.

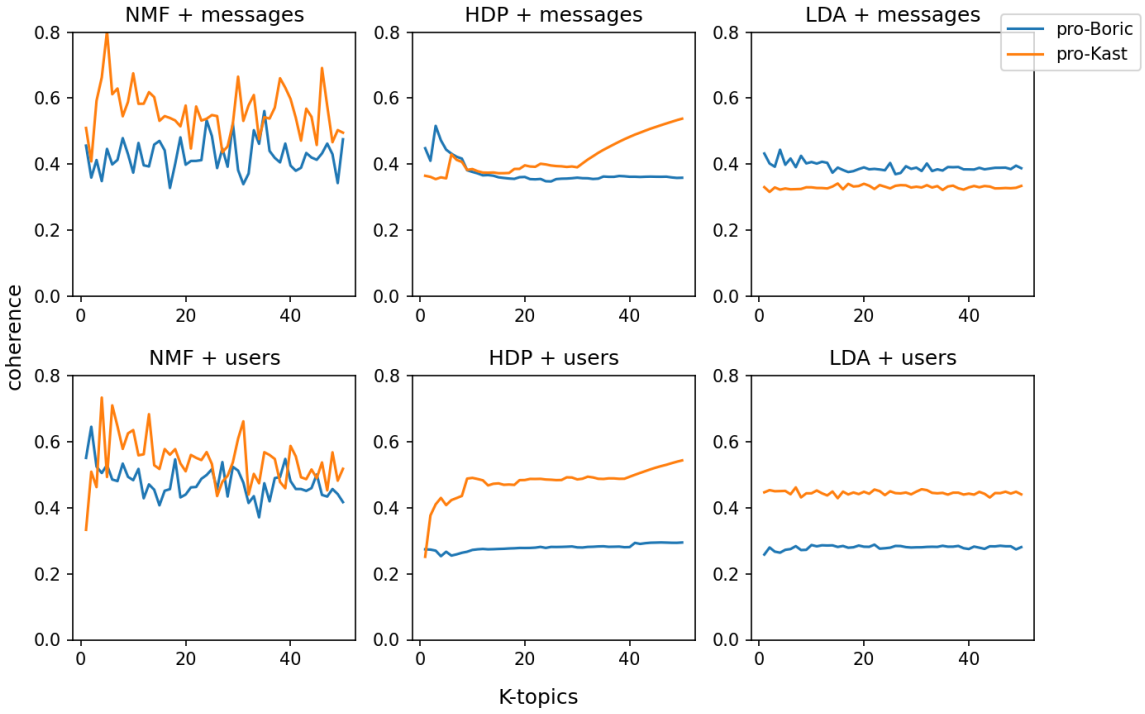

Fig. 11. Coherence scores for various topic models on Whatsapp, each combining a topic modeling algorithm and document representation, evaluated across 2 to 50 topics.

Figure 11 shows the coherence score in each Pro-Boric and Pro-Kast communities by varying the number of topics for the six evaluated models. Additionally, Table 6 shows the best performance for each model according to coherence and number of topics obtained in communities. We found that on average, 19 topics were estimated by models, resulting in a coherence score of 0.4995. For this dataset, the NMF algorithm had a better performance for both document representations than the LDA and HDP. Additionally, the NMF obtained a fewer number of topics for the optimal coherence score than the other models.

Until this part of the valuation, we notice two main general observations comparing with the evaluation in the Twitter dataset: 1) the average number of topics that obtained the best performance is almost the same. However, standard deviation was higher in Whatsapp. 2) the average coherence across models was 18% higher in Whatsapp. Furthermore, they achieved a lesser standard deviation. Hence, our results initially suggest that computed topics had a better adjustment for the content in Whatsapp than Twitter.

We next estimated $homogeneity_{\alpha}$ and $relevancy_{\beta}$ scores by considering the best setting of the topic models and the two framing identification approaches. Figure 12 shows the homogeneity score divided into topic modelling algorithms and framing identification approaches. We observe that all settings which use the NMF algorithm obtained non-zero values of homogeneity for every evaluated α. Similar to Twitter, the other topic algorithms (HDP and LDA) achieved the lowest score, representing a high homogeneity of frame candidates in the embedding model.

As Figure 13 shows, we also computed the relevancy score for every setting. We note there are no settings that identify all frames as relevant. Our results display that the best performance was achieved only in one setting, achieving

Algorithm	Doc. Representation	nº Topics c1	nº Topics c2	Coherence c1	Coherence c2	Avg. Topics	Avg. coherence
HDP	tweets	3	50	0.5159	0.5376	26.5	0.5268
HDP	users	46	50	0.2946	0.5431	48	0.4189
LDA	tweets	4	15	0.4437	0.3414	9.5	0.3926
LDA	users	22	7	0.288	0.4614	14.5	0.3747
NMF	tweets	29	7	0.4899	0.7006	18	0.5953
NMF	users	2	4	0.6449	0.7333	3	0.6891

Table 6. Optimal coherence scores for each topic model on Whatsapp, distinguished by their topic modeling algorithm, document representation, and ideal topic count for maximizing coherence in community documents. Communities c1 and c2 represent pro-Boric and pro-Kast groups, respectively.

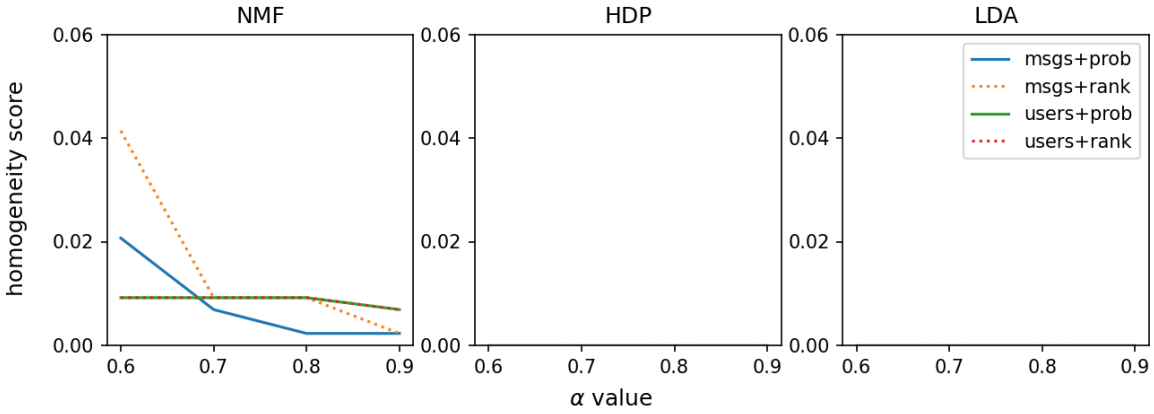

Fig. 12. Homogeneity scores of frame candidates in Whatsapp content using probabilistic-based and rank-based approaches, evaluated for an α range of 0.6 to 0.9. Homegeinity scores equal to zero represent higher diversity values.

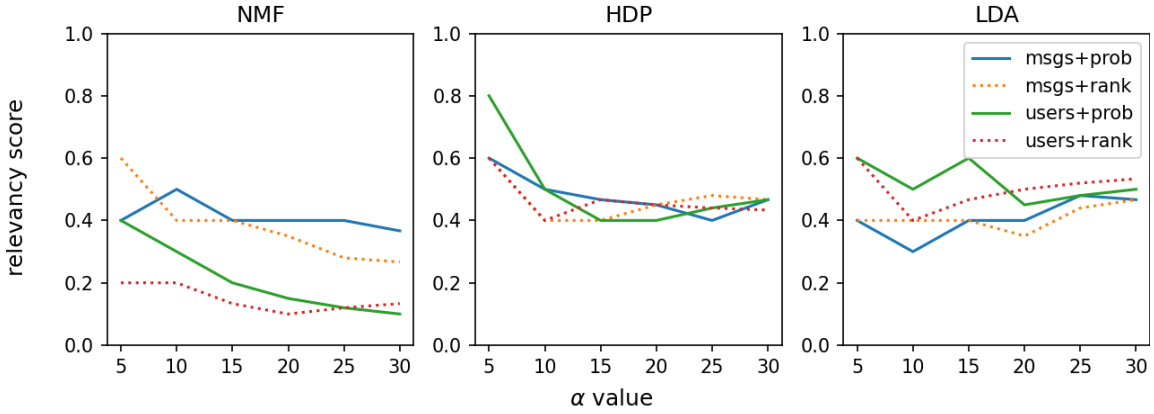

Fig. 13. Relevancy scores of frame candidates in Whatsapp content using probabilistic-based and rank-based approaches, evaluated for an β range of 5 to 30.

a relevancy of 0.8 for the smallest β value corresponding to $\beta = 5$. This setting used the HDP algorithm, the user-based text document representation, and the probabilistic approach. Additionally, inspecting other models we observe that

relevancy did not exceed the value of 0.6 for this scenario. This therefore suggests that, even for the fewest number of relevant terms, the identified frames may not be meaningful in our gold-standard collection of terms.

In addition, we computed relevancy for a range of $\beta > 5$ values, in which the results decay drastically. We notice that this score does not exceed 0.5 in most of the settings, suggesting that the identification of relevant frames is only possible for a few number of terms in this platform. In this scenario, there is no clear pattern regarding which setting maximizes relevancy with respect to the algorithm and framing identification approach.

Finally, by studying both homogeneity and relevancy, our analysis suggests that the initial selection of topic models—and their corresponding settings—can be highly influenced by the relevancy score, as there are no settings that identify all frames as relevant. Hence, maximizing its value can be crucial for identifying frame candidates. This approach ensures that the most pertinent frames are highlighted, enhancing the accuracy and effectiveness of the analysis.

Overall, according to the best scores found in the evaluation, our results suggest that representing documents by aggregating messages by user, employing the HDP algorithm, and using a probabilistic-based approach for the identification of frames maximizes the results.

4.4 Results

In this section, we analyze community framing on social media platforms. For each platform, we select the frames that demonstrated the best performance based on the metrics and settings used in the framing evaluation.

The analysis quantifies the semantic differences and similarities across communities in their use of these frames. We deem an analogous procedure as presented by Sarmiento et al. [81]. They considered that the exact frame resides in the same semantic space but treating as different lexical units. To do this, we created a joint word vector model in which the frames are disambiguated according to the community they appear in. For instance, the frame *boric* was renamed as *c1_boric* and *c2_boric* in the set of documents corresponding to communities $c1$ and $c2$, respectively. Therefore, after training the word embedding model by considering this disambiguation, the same frame can be represented as two different word vectors. Thus, we can apply vector operations on these word vectors (e.g., similarity, neighborhood) to measure community framing in our embedded model.

We further conducted a low dimension analysis to quantify community framing. We adopted an approach proposed by Sweeney and Najafian [88], which uses the relative negative sentiment to assess fairness in word embeddings. The basic premise of this approach is that words can be transformed from the embedding space into a sentiment probability through a logistic regression model trained on pre-labeled words.

Analogous to the framing evaluation, we present results of the community framing for each platform by describing the most relevant findings of them.

4.4.1 Twitter dataset. We obtained that the entire set of frame candidates was relevant ($relevancy = 1$) when $\beta = 5$. Additionally, for highest values of β, the best setting achieved an relevancy of 0.758. In this direction, we chose that setting that allows us to analyze multiple frames and demonstrate good performance for both homogeneity and relevancy scores. Thus, we consider five frames for which the HDP algorithm, the representation of a document through an individual message, and the probabilistic-based approach, obtained the best performance for $\beta = 5$.

Table 7 displays the selected frames sorted in decreasing order based on their estimated probabilities from the framing identification approach. Additionally, we have included the cosine similarity of each frame in our word2vec model. To train the word2vec model, we first concatenated both community corpora and used the skip-gram negative sampling

Frame	Max prob c1	Max prob c2	Min communities	Cosine similarity
chile	0.0055	0.0057	0.0055	0.6061
boric	0.0073	0.0052	0.0052	0.5686
kast	0.0029	0.0058	0.0029	0.5263
candidato (candidate)	0.0024	0.0027	0.0024	0.4652
presidente (president)	0.0023	0.0033	0.0023	0.3906

Table 7. Frames found in Twitter according to the results obtained from the framing identification and evaluation.

method [72]. In addition, we included the following parameters at the moment of training: *window_size = 7*, *epoch = 15*, *vector_size = 100* and *min_freq = 5*. Thus, we computed the similarity between the two disambiguated word vectors that represent the same frame, but in different communities

Our first observation is that the frames with the highest probabilities also have the highest similarities when this value is computed between the word vectors of the disambiguated frames. For instance, the top two frames *boric* and *chile* obtained cosine similarities of 0.5686 and 0.6061, respectively. On contrary, the least likely frames *candidato (candidate)* and *presidente (president)* had cosine similarities of 0.4652 and 0.3906, respectively. This first finding suggests that if a frame is most probable between communities, it may be also most similar in a common semantic space. This implies that community framing involving these types of frames might exhibit less polarization, as they display a higher frequency of shared multi-word expressions. Consequently, this could amplify semantic differences between communities.

We next obtained the top nearest terms for each disambiguated frame to gain more insights into the mentioned differences between communities. We then visualized them, applying a dimensionality reduction using the T-SNE (T-distributed Stochastic Neighbor Embedding) algorithm. Figure 14 shows two examples of the top-30 nearest terms for the frames *chile* and *presidente (president)*, which had the highest and lowest cosine similarities respectively [4]. In both cases, it is not visually possible to separate terms depending on the community in which they appear.

Given that visually we did not observe differences for the selected frames, we next focus on specific characteristics of these top-30 nearest terms when they are analyzed in each community. Here, we noticed differences in the number of overlapping words between communities for each frame. In a detailed analysis of the frame *chile*, we identified six common terms shared between communities, such as *cambios_concretos (concrete_changes)* and *proteger_democracia (protect_democracy)* (see Table 8). In addition, a closer examination of each community's data revealed that in community $c1$ the most similar terms for the frame were *cambios_concretos (concrete_changes)* and *parlamento_apoye (parliament_supports)*. Conversely, community $c2$ primarily utilized terms like *pueblo_construye (people_build)* and *nombre_democracia (name_democracy)* to refer to the frame. Additionally, we provide a few examples of the messages related to the both *presidente (president)* and *chile* frames divided into communities $c1$ and $c2$ (see Table 10 and Table 9). Overall, our findings suggest that there is no clear pattern associated with one candidate or objective discourse about the terms surrounding this frame in each community.

In the case of the frame *presidente (president)*, our results exhibit a completely different pattern in comparison with the previous analysis. First, we identify the presence of three common terms across different communities, such as *presidente_necesitamos (president_we_need)*, *boric_electo (boric_elected)* and *esperanza_justicia (hope_justice)*. In detail, our community-specific analysis revealed that the community labeled as $c1$ predominantly utilized the terms

[4]The complete visualization of all frames is published in our repository.

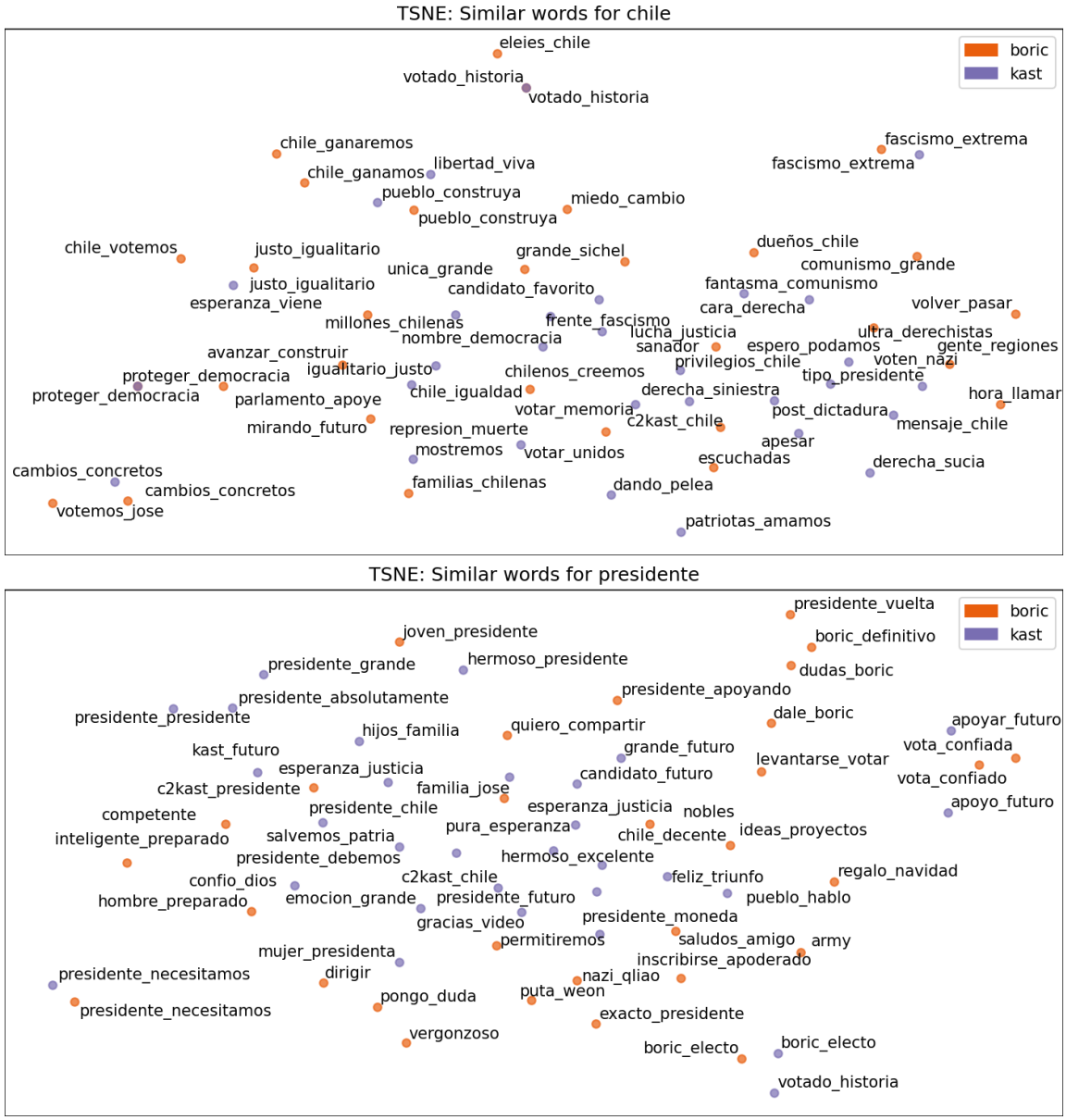

Fig. 14. Examples of frames in Twitter. The visualization presents two dimensional word vector representations of top-30 nearest words for the frames *chile* and *presidente (president)*. Orange and purple points represent the nearest words found in pro-Boric and pro-Kast communities, respectively.

joven_presidente (young_president) and *dale_boric (go_boric)*. In addition, this community exhibited terms, such as *inscribirse_apoderado (register_as_agent)*, *levantarse_votar (rise_to_vote)* and *vota_confiado*, which can be associated with a rumor about electoral fraud from other right-wing party [75]. Conversely, the community labeled as *c2* commonly used the terms *kast_futuro (kast_future)* and *familia_jose (family_of_jose)*, which can be linked with the mention of

Frame	Community c1	Community c2	Common terms between communities c1 and c2
chile	cambios_concretos (concrete_changes), parlamento_apoye (parliament_supports), chilenos_creemos (chileans_believe), proteger_democracia (protect_democracy), justo_igualitario (fair_egalitarian)	justo_igualitario (fair_egalitarian), nombre_democracia (name_of_democracy), cambios_concretos (concrete_changes), pueblo_construya (people_build), proteger_democracia (protect_democracy)	justo_igualitario (fair_egalitarian), cambios_concretos (concrete_changes), fascismo_extrema (extreme_fascism), votado_historia (voted_history), proteger_democracia (protect_democracy)
presidente (president)	joven_presidente (young_president), dale_boric (go_boric), esperanza_justicia (hope_justice), inscribirse_apoderado (register_as_proxy), levantarse_votar (rise_to_vote)	confio_dios (trust_in_god), chile_decente (decent_chile), kast_futuro (kast_future), familia_jose (family_of_Jose), salvemos_patria (save_the_homeland), esperanza_justicia (hope_justice)	presidente_necesitamos (president_we_need), boric_electo (boric_elected), esperanza_justicia (hope_justice)

Table 8. Examples of the nearest collocations for the frames *chile* and *presidente (president)* in Twitter. Table also shows common collocations found between communities

Messages c1	Messages c2
Que foto tan hermosa, dos personas en posición de poder pero que no lo anhelan para su beneficio sino para avanzar a un mejor país, más justo, igualitario y digno	Por un Chile distinto y renovado, más inclusivo, fraterno, justo, igualitario #Elecciones2021CL
What a beautiful photo, two people in positions of power who do not crave it for their own benefit but to advance towards a better country, more just, equal, and dignified #Elecciones2021CL	*For a different and renewed Chile, more inclusive, fraternal, just, and equal. #Elections2021CL*
Todo el éxito!!! El domingo todos a votar para seguir construyendo un chile más justo e igualitario.	Con la más apañadora, esperenado nuestro turno con esperanza en un Chile más justo e igualitario
All the success!!! On Sunday, everyone go vote to continue building a more just and equal Chile	*With the most supportive one, waiting our turn with hope for a more just and equal Chile*

Table 9. Examples of messages for the frame *chile* on Twitter. User mentions other than the candidates were removed. Additionally, we removed emojis and URLs from messages.

the candidate José Antonio Kast. Furthermore, we note that this community drew attention to certain terms, namely *confio_dios (trust_in_god)*, *salvemos_patria (save_our_homeland)*, and *chile_decente (decent_chile)*, which are associated with the central propaganda of the right-wing political party and the campaign of their candidate [79]. Overall, the qualitative analysis of frames by community revealed that communities $c1$ and $c2$ can be characterized as Pro-Boric and Pro-Kast, respectively.

Although our analysis focused on frames with the lowest and highest probabilities due to space constraints, the other frames showed a similar usage pattern among communities. For example, the frame *kast* had nearest terms such as *creemos_boric (we_believe_boric)* and *nazi_jose* in Pro-Boric community, while Pro-Kast community displayed

Messages c1	**Messages c2**
@user Sí, más difícil que la cresta pero peor si El innombrable es presidente. Así que con todo para lograr a #BoricPresidente para que haya más esperanza, justicia y derechos y menos odio.	No hay donde perderse, @joseantoniokast es el Presidente que necesitamos #CandidatoLlegóTuHora #AtreviDOS
@user Yes, harder than hell but worse if the unmentionable is president. So we go all out to achieve #BoricPresident so there will be more hope, justice, and rights, and less hate.	*There's no getting lost, @joseantoniokast is the President we need #CandidateYourTimeHasCome #DareTo*
Mañana a levantarse temprano e ir a votar. No importa el calor, la fila, nada, lo que importa es construir el país en que queremos vivir con nuestros hijos e hijas.	@joseantoniokast GRANDE PRESIDENTE. INSCRIBIRSE COMO APODERADO DE MESA.
Tomorrow, get up early and go vote. It doesn't matter the heat, the line, nothing, what matters is building the country we want to live in with our sons and daughters.	*@joseantoniokast GREAT PRESIDENT. REGISTER AS A POLLING AGENT.*

Table 10. Examples of messages for the frame *presidente (president)* on Twitter. User mentions other than the candidates were removed. Additionally, we removed emojis and URLs from messages.

terms like *derecha_encanta (right_love)* and *anti_comunistas (anti_communist)*. Similarly, for the frame *boric*, terms like *boric_grande (great_boric)* and *boric_excelente (excelent_boric)* appeared in Pro-Boric community, while *boric_amarillo (spineless_boric)* and *derecha_decente (right_decent)* were present in Pro-Kast community. Previous studies have approached this evaluation by labeling users based on their stance or political party and estimating various metrics, such as f1-score and accuracy [81]. However, in our analysis, we minimized human intervention during the initial phases by avoiding such evaluation.

To quantitatively comprehend community framing for a broad range of neighbor words, we estimated the ratio of common terms that are shared between communities when we analyzed a frame. We computed the Jaccard index for each frame by varying the numbers of k-neighbor terms in the range of $k \in [25, 500]$, with intervals of 25 nearest words. Figure 15 illustrates the results of this analysis, where we observe that the Jaccard index for the frame *chile* increases as the number of common terms grow. This may indicate that this frame was semantically similar between communities, indicating common discussions surrounding it. In contrast, our results show that the frame *presidente (president)* maintains its value almost constant across different values. Therefore, this may indicate that the frame *presidente (president)* was discussed with different meaning and perspectives between communities given its low and almost constant ratio of common terms.

Finally, we include a low dimensional analysis that enables us to establish fairness across communities, using a similar methodology as outlined by [88]. To achieve this, we assessed the comparative negative sentiment linked to each frame within these communities. We utilized a collection of $2,784$ Spanish words from the NRC, linked with $1,246$ positive and $1,538$ negative terms [74]. With these, we trained a logistic regression model that estimated the sentiment probability of a word vector within our joint word2vec model. Consequently, we predict the negative sentiment probability of the of the frame's word vectors in each community in each community, resulting in a probability distribution of the complete set of frames. This distribution is compared with a uniform distribution using the Kullback-Leibler (KL) divergence as a measure of bias.

Figure 16 shows the sentiment probability of each frame and the negative distribution by community. We observe that every frame obtain a distinct sentiment probability, in terms either of its intensity or its polarity. However, we

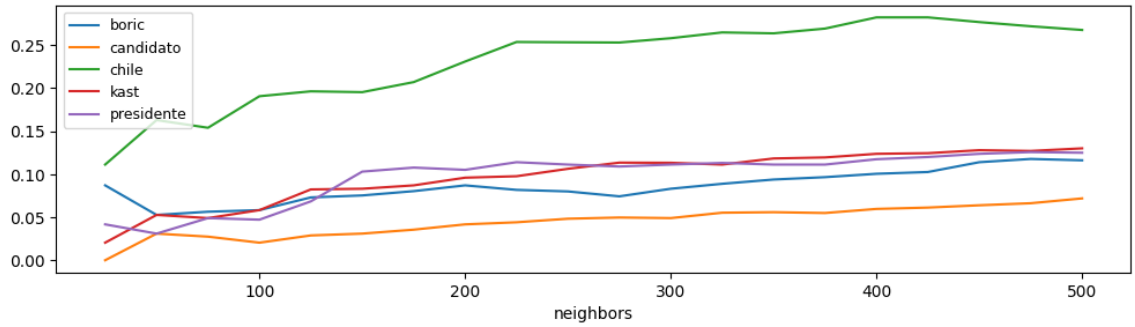

Fig. 15. Jaccard index for the k-nearest neighbors collocations found between Twitter communities for each frame. Neighbors collocations are obtained in the of $k \in [25, 500]$ with intervals of 25 neighbors. The following terms are translated from Spanish to English as follows: *candidato (candidate)* and *presidente (president)* .

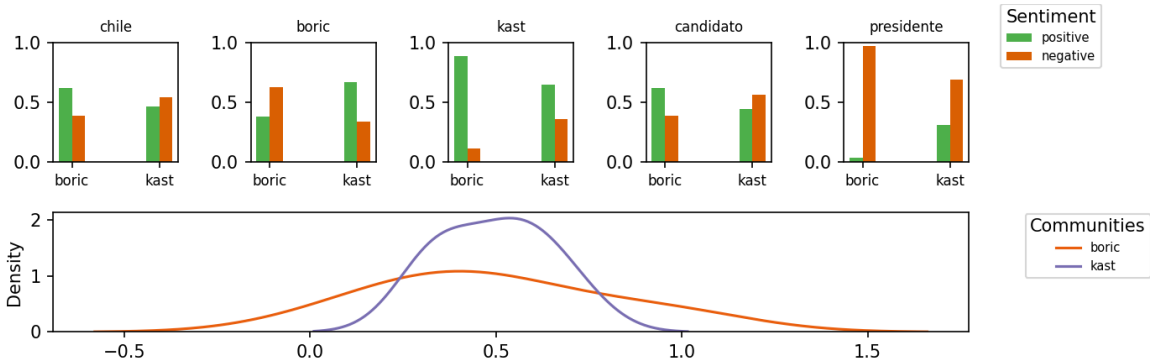

Fig. 16. Sentiment probability of the frames found in Twitter. The top image shows the community's positive and negative probabilities of each frame. The bottom picture displays the probability density distribution of the negative sentiment. The following terms are translated from Spanish to English as follows: *candidato (candidate)* and *presidente (president)*

obtained unexpected results in frames directly associated with the mention of candidate's names. For the frame *boric*, a higher negative sentiment probability appeared in the Pro-Boric community $c1$, while a higher positive sentiment probability can be noted in the Pro-Kast community $c2$. This result can be attributed to the inability of word embeddings to discriminate between the different meaning of a word, which is also documented by Sarmiento et al. [81] in their analysis of the 2019 Chilean unrest movement.

Regarding the negative probability distribution for the frames shows on the bottom in Figure 16, we visually notice differences in the shape of them. By estimating the KL divergence, we obtained values of $KL_{Boric} = 0.251$ and $KL_{Kast} = 0.053$ for Pro-Boric and Pro-Kast communities, respectively. As noted, their values are totally different in magnitude, representing a significant difference in the amount of information necessary to encode and transmit from one distribution to another. Hence, our results suggest that frames exhibit different polarities in both communities, where the Pro-Boric community shows a wider range of sentiment states, both positive and negative.

4.4.2 Whatsapp dataset. We determined that the optimal configuration included five frames, resulting in a relevancy of 0.8 and no similarity between frames at different thresholds. This configuration employed the HDP algorithm,

frame	Max prob c1	Max prob c2	Min communities	Cosine similarity
kast_kast	0.0195	0.0861	0.0195	0.8448
kast	0.0072	0.0053	0.0053	0.6154
boric	0.0033	0.0465	0.0033	0.5446
gente (people)	0.0032	0.0036	0.0032	0.3804
gabriel_boric	0.0032	0.0030	0.0030	0.2295

Table 11. Frames found in Whatsapp according to the results obtained from the framing identification and evaluation.

the representation of a document through the concatenation of messages from one user, and the probabilistic-based approach.

Table 11 displays the chosen frames sorted by their probabilities according to the chosen framing evaluation approach. Analogous to Twitter, we follow the same steps for disambiguating frames and training the word2vec model using the same settings. With this in mind, we obtained that the frames with the highest probabilities were *kast_kast* and *kast*, with corresponding cosine similarities of 0.8448 and 0.6154, respectively. Conversely, the least likely frames were *gente (people)* and *gabriel_boric*, achieving cosine similarities of 0.3804 and 0.2295, respectively. Considering these results, we observed that most of the frames were directly linked to the names of the candidates, except for *gente (people)*.

In accordance with the methodology employed for the analysis of the Twitter dataset, we conducted an estimation of the top nearest terms for each frame, which were then visualized in a two-dimensional space. Our particular focus was on the disambiguated frames exhibiting the highest similarity and dissimilarity, namely *kast_kast* and *gabriel_boric*, respectively (refer to Figure 17). The visualization of the two-dimensional space revealed a distribution of terms grouped according to their respective communities. Specifically, the majority of terms were observed to be clustered based on the communities they belonged to, as showed in Figure 17. Our results demonstrate that, while there is a distribution of terms by group, in the case of the frame *kast_kast* there exists a third group where both communities share the majority of terms.

We next focus on a deeper semantic understanding of the top-30 nearest terms found for the frames *kast_kast* and *gabriel_boric* across communities (see Table 12) Our results showed that *kast_kast* shared nine terms between communities. Examples of these common terms are *nacion_kast (kast_nation)*, *kast_presidente (kast_president)*, *chile_exista (chile_exists)* and *jamas_marxista (never_marxist)*. Our community-specific analysis showed that the Pro-Boric community predominantly used the terms *dosis_socialismo (dose_socialism)*, *saluda_pueblo (greets_people)* and *manipulando_chile (manipulating_chile)*. On contrary, we identified that, in Pro-Kast groups, nearest terms are *jamas_marxista (never_marxist)*, *familia_familia (family_family)* and *nacion_kast (nation_kast)*. Our results suggest that both communities share various terms, which even appear in most nearest words for each disambiguated frame. In fact, we noted that nearest terms cannot be differentiated depending on which community they appear because a vast portion of them (almost one-third) are shared between groups. Overall, we observed that these terms did not allow identifying a specific stance regarding how the frame was discussed in each community, which can be supported by the previous statement.

In the case of the frame *gabriel_boric*, there were not common terms between groups. In detail, our analysis for the Pro-Boric community revealed that most similar terms for the frame were *concentremos_esfuerzos (let's_focus_efforts)*, *neutrales_apoyamos (neutrals_support)* and *todxs_chilenxs (all_chileans)* These terms can be interpreted as a direct and explicit support to the candidate Boric, in which groups and users called for a joint effort independent of their party

Fig. 17. Examples of frames in Whatsapp. The visualization presents two dimensional word vector representations of top-30 nearest words for the frames *kast_kast* and *gabriel_boric*. Orange and purple points represent the nearest words for Boric and Kast communities, respectively.

politics or gender [24, 94]. Conversely, the community-specific analysis for Pro-Kast groups displayed that the most similar terms were *denuncia_candidato (denounces_candidate)* and *tapar_situacion(cover_situation)*. These words can be associated with the alleged harassment complaint involving the candidate, Gabriel Boric [23]. Overall, our results for

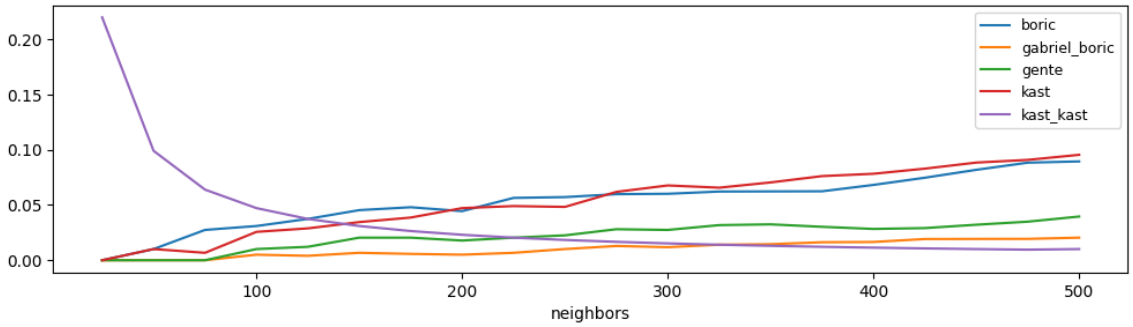

Fig. 18. Jaccard index for the k-nearest neighbors collocations found between Whatsapp communities for each frame. Neighbors collocations are obtained in the of $k \in [25, 500]$ with intervals of 25 neighbors. The following term is translated from Spanish to English as follows: *gente* (people).

frame	**Community c1**	**Community c2**	**Common terms between communities**
kast_kast	saluda_pueblo (greets_people), manipulando_chile (manipulating_chile), libertad_libertad (freedom_freedom), dosis_socialismo (dose_of_socialism), nacion_kast (nation_kast), libertador_presidente (liberator_president)	familia_familia (family_family), entiende_idea (understands_idea), saluda_pueblo (greets_people), lindo_compartir (nice_to_share), libertad_libertad (freedom_freedom)	jamas_marxista (never_Marxist), nacion_kast (nation_kast), libertador_presidente (liberator_president), kast_presidente (kast_president)
gabriel_boric	concentremos_esfuerzos (let's_focus_efforts), neutrales_apoyamos (neutrals_support), todxs_chilenxs (all_chileans), puerta_puesta (door_put), amig_intentemos (friend_let's_try), vivo_compañero (alive_comrade)	denuncia_candidato (denounces_candidate), auto_comunistas (self_communists), tapar_situacion (cover_situation), esperando_provocacion (waiting_provocation), kast_crucificado (kast_crucified), logrado_mintiendo (achieved_by_lying)	-

Table 12. Examples of the nearest terms for the frames *kast_kast* and *gabriel_boric* in Whatsapp. Table also shows common collocations found between communities.

the frame *gabriel_boric* revealed that communities tend to discuss from different point of views, which is supported by the fact that we did not find common terms in communities and the semantic differences in the use of language of them.

We expanded our study to incorporate a greater number of nearest terms. Figure 18 depicts the Jaccard index calculated for the nearest terms in each pair of disambiguated frames between communities, using the same range of

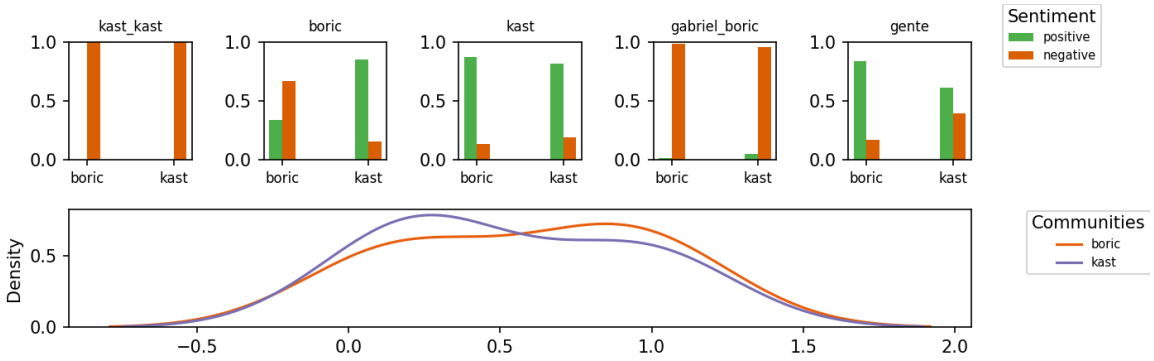

Fig. 19. Sentiment probability of the frames found in Whatsapp. The top image shows the community's positive and negative probabilities of each frame. The bottom picture displays the probability density distribution of the negative sentiment. The following term is translated from Spanish to English as follows: *gente* (people)

neighbors as in Twitter. Our findings highlight two noteworthy observations for the frames *kast_kast* and *gabriel_boric*. Firstly, we observed that the Jaccard index for the frame *kast_kast* is highest for the minimum number of evaluated neighbors (25-nearest terms), but decays exponentially as the number of neighbors increases. This suggests that discussions surrounding this frame may be highly similar when we limit the analysis to a smaller number of nearest terms. However, as the number of neighbors increases, the homogeneity of topics discussed in relation to the frame becomes more apparent. Thus, we can conclude that the similarity of discussions on *kast_kast* is context-dependent and influenced by the number of nearest terms considered in the analysis. In contrast, the Jaccard index for the frame *gabriel_boric* remains almost constant and close to zero as the number of neighbors increases. This indicates that the discussions surrounding this frame are markedly distinct in both communities, owing to differences in language usage.

Finally, we estimated the sentiment probability of the frames in the word embedding model to quantify fairness between communities. Figure 19 shows that most of frames have a similar sentiment probability, with the exception of the frame *boric* that exhibits a different pattern between communities. Furthermore, we observe that, frames which should be associated with a well-known polarity in their communities, obtained an opposite value of sentiment probability. For instance, the frame *kast_kast* exhibits only negative values in Pro-Kast groups and *kast* positive probability in Pro-Boric groups. Similar to the case of Twitter, these results also suggest the inability of word embeddings to discriminate between the different meaning of a word.

Another result worth noting is that the probability density distribution of negative sentiment in each community appears to have a similar shape, but with a different skew (see the bottom of Figure 19). Comparing each probability with a uniform distribution, we obtained KL divergences of of $KL_{Boric} = 0.345$ and $KL_{Kast} = 0.354$ for communities of Boric and Kast, respectively. Hence, these scores suggest that communities show a similar sentiment distribution across frames.

5 Discussion

We developed and evaluated a method for identifying framing that includes choosing the most likely (or top-ranked) multi-word expressions. These expressions are prominent and meaningful in the subjects conversed about among various communities. Through an extensive evaluation of diverse topic modelling algorithms, document representations,

framing identification approaches and metrics, our methodology provided a novelty procedure that require weak human intervention to select and analyze frames in polarized online discussions.

We evaluated three topic modelling algorithms which were initially fine-tuned with multiple number of topics with the goal of maximizing the coherence score. In both Twitter and Whatsapp platforms, we noted that the NMF algorithm outperformed at least 20% the other scenarios, independent of whether the representation of a document was through individual messages or concatenation of a user's content. Nevertheless, our evaluations using the homogeneity and relevancy scores showed that the NMF achieved one of the worst values, especially in the former metric. Hence, considering only coherence as a measure of best topic model selection for identifying salient themes discussed between communities, it may provoke a inaccurate selection of frames.

We proposed the identification of frames candidates based on their probability and ranking inter and intra communities. On the one hand, our results exhibited that the probabilistic-based approach obtained better performance in most of the scenarios across platforms when we evaluated relevancy. On the other hand, the ranked-based approach displayed a slightly higher performance in various scenarios when homogeneity was computed. However, we noticed that this pattern only appeared for Twitter data.

By comparing the framing evaluations between the studied platforms, we can describe various observations which can be deduced from our results. We first noted that the average number of topics that maximize coherence in each platform were similar, obtaining 19.4 and 19.9 for Twitter and Whatsapp, respectively. However, we observed that the identification of topics in Twitter was more sparse than Whatsapp, achieving a standard deviation of 18.02 and 15.87 in each social network, respectively. Secondly, our results in Tables 5 and 6 show that the average coherence seems higher in Whatsapp than Twitter, with values of 0.499 and 0.407, respectively. In addition, coherence scores covered a wider range of values in the case of Twitter, having 0.179, 0.641 and 0.267 for standard deviation, maximum and minimum values, against 0.125, 0.689 and 0.347 in Whatsapp.

We further realized that Twitter and Whatsapp exhibited differences for homogeneity and relevancy scores. Our results for homogeneity (see Fig. 9 and 12) displayed that, in both platforms, this score obtained the best performance for HDP and LDA algorithms. However, for evaluations when we considered the NMF algorithm, both platforms showed that this model generated less varied frames with respect to the semantics surrounding these terms. In addition, this effect was more pronounced in Twitter, showing a worse performance (highest values) than Whatsapp. Our analysis further demonstrated that frames computed from Twitter were almost twice as relevant as in Whatsapp. In fact, our complete evaluation for different β values exhibited average relevancy scores of 0.711 and 0.409 for Twitter and Whatsapp, respectively. This suggest that extracting significant frames from tweets may be more feasible than chat conversations.

We also discuss results for the specific frames obtained in each platform and community. Inspecting Tables 7 and 11, our first finding was that two terms (*boric* and *kast*) were exactly the same in both platforms. However, our probabilistic-based approach, which obtained the best performance for selecting frames in both platforms, exhibited a higher probability for the frame *boric* in Twitter, almost doubling its value in comparison to Whatsapp. On contrary, our results showed that the frame *kast* was twice as likely on Whatsapp than Twitter.

We made a second observation regarding the frames used in Twitter and WhatsApp, specifically concerning the similarity of the selected terms among the pro-Boric and pro-Kast communities. Upon calculating the average cosine similarity for the set of frames, we obtained values of 0.511 and 0.522 for Twitter and WhatsApp, respectively. Although these results were similar, we observed a significant difference in the standard deviation for each platform. In Twitter, the standard deviation was 0.085, whereas in WhatsApp, it was 0.234. This suggests that the content in WhatsApp may

be more fragmented among communities than on Twitter. One possible explanation for this difference is that WhatsApp is a more private platform than Twitter. This means that users may be more likely to share information with people who share their views on WhatsApp, which could lead to more homogeneous communities. Additionally, WhatsApp is a messaging platform, which means that users are more likely to have one-on-one conversations than on Twitter, where users can see and respond to messages from a wider audience. This could also lead to more fragmentation, as users may be less likely to be exposed to opposing viewpoints.

6 Conclusion

In this study, we aimed to better understand how polarization works on digital platforms by focusing on the role of frames. Our approach used a mix of machine learning, network analysis, and natural language processing to identify and assess the quality of these frames in online discussions. Our methods offer a starting point for future studies on how to pick the most relevant and uniform frames for closer study.

An important finding was how different algorithms work on different platforms. For example, the Non-negative Matrix Factorization (NMF) algorithm was good at maximizing coherence but not as effective in ensuring homogeneity. This shows that we need to use multiple criteria for evaluating the quality of frames, rather than just one.

We also noticed differences between Twitter and WhatsApp in our case study on the 2021 Chilean Presidential Elections. Frames identified from Twitter discussions were generally more relevant but less diverse in meaning. This could mean that Twitter is a better platform for finding important frames, but we should be careful not to oversimplify the issues. WhatsApp, on the other hand, showed more diversity in discussions, likely because it is a more private platform that encourages in-group conversations.

Our results also showed that some frames, like *boric* and *kast* were more common on one platform over the other. This reminds us that we need to consider multiple platforms to get a complete picture of how topics are framed online.

6.1 Limitations

Our method has limitations in both the frame identification component and the determination of their meaning across communities. These limitations are described below.

6.1.1 Limitations in the frame identification component.

(1) The method restricts the concept of a frame to multi-word expressions in textual content: this excludes the possibility of frames appearing in different formats such as images, audio, or video.
(2) Communities do not overlap: our method does not account for elements that belong to one community appearing in two or more groups simultaneously. This also applies to communities that may be reflected in a hierarchical manner.
(3) Our method is specifically designed to identify frames in polarized settings: this feature requires frames to be relevant to at least one topic discussed across different communities. It may overlook frames that are highly salient in one community but only marginally relevant in another, despite some degree of framing being present.
(4) Both proposed frame identification approaches do not consider ties when collocations obtain the same value: if two or more multi-word expressions obtain the same ranking or probability, the method may not accurately distinguish between them, potentially leading to ambiguities in frame identification and a less precise understanding of the key issues within the discourse.

6.1.2 Limitations of the meaning determination component.

(1) Static word embeddings are susceptible to the meaning conflation deficiency problem: if a frame includes a polysemous word, the representation might merge different senses of the word. This issue could be mitigated using sense embedding models or contextualized embeddings, but this would increase complexity.
(2) Our approach does not account for how frames evolve within communities on social media: frames can change over time, and without longitudinal analysis, our method may miss these shifts and trends, potentially leading to a static and incomplete representation of the discourse.

6.2 Future Work

Based on the identified limitations of our current method, future work could focus on addressing these challenges and expanding the scope of the research. The following areas are proposed for further development:

(1) Expanding frame representation to multi-modal data: research methods for recognizing frames in non-textual content such as images, audio, and video. This would allow for a more comprehensive analysis of frames across various media formats, enhancing the method's applicability in diverse communication settings [77].
(2) Incorporate overlapping community detection algorithms: This approach will enable a more nuanced analysis of frames within the complex structure of online social networks, allowing for elements to belong to multiple communities. [80, 99].
(3) Handling ties in frame ranking: Improve the frame identification algorithms to handle ties more effectively when multiple expressions have the same ranking or probability. This could include implementing a tie-breaking mechanism or a more nuanced scoring system to better differentiate and prioritize frames [63].
(4) Incorporate contextualized embeddings to overcome the limitations of static word embeddings: this approach would help in accurately capturing the different senses of polysemous words within frames, thus improving the precision of meaning determination [4, 56].
(5) Exploring embedding-based topic models as an alternative to traditional topic modeling approaches : these models can leverage the rich semantic information encoded in embeddings to provide more detailed and contextually relevant thematic categorization, potentially offering superior performance to traditional frequency-based approaches [34, 45].
(6) Utilizing modern LLMs to interpret the identified frames via prompting: explore the use of large language models (LLMs) to better understand and interpret frames. LLMs can provide nuanced insights into the context and significance of frames, potentially enhancing our method's analytic capabilities [20]. Additionally, LLMs can offer judgment and provide tools for manual labeling, thereby improving the accuracy and depth of frame analysis. [104].
(7) Enhancing the evaluation of our frames: future work will include a comparison with manual frame identification by experts [67]. This will provide additional validation by benchmarking our method against expert judgments, thereby ensuring greater relevancy and accuracy.

Acknowledgments

We acknowledge the support of ANID - Millennium Science Initiative Program - Code ICN17_002 (IMFD) and the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. Additionally, MS is funded by ANID, Fondecyt

de Iniciación, grant Nº 11230980 and Fondo de Financiamiento de Centros de Investigación en Áreas Prioritarias, grant Nº NCS2021_063.

References

[1] Abeer Aldayel and Walid Magdy. 2019. Assessing sentiment of the expressed stance on social media. In *International Conference on Social Informatics*. Springer, 277–286.

[2] Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management* 58, 4 (2021), 102597.

[3] Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 9335–9348.

[4] Asaf Amrami and Yoav Goldberg. 2018. Word Sense Induction with Neural biLM and Symmetric Patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 4860–4867. https://doi.org/10.18653/v1/D18-1523

[5] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 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.

[6] Natalia Aruguete and Ernesto Calvo. 2018. Time to# protest: Selective exposure, cascading activation, and framing in social media. *Journal of communication* 68, 3 (2018), 480–502.

[7] Natalia Arugute, Ernesto Calvo, and Tiago Ventura. 2023. Network activated frames: content sharing and perceived polarization in social media. *Journal of Communication* 73, 1 (2023), 14–24.

[8] 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.

[9] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[10] BBCNews. 2021. Gabriel Boric: en qué consiste la agenda transformadora con la que llega a la presidencia de Chile. https://www.bbc.com/mundo/noticias-america-latina-59723286. [Online; accessed 19-May-2023].

[11] BBCNews. 2021. Kast vs. Boric: las principales propuestas de los rivales más antagónicos que ha tenido Chile en las últimas décadas. https://www.bbc.com/mundo/noticias-america-latina-59383712. [Online; accessed 19-May-2023].

[12] BBCNews. 2021. Kast vs. Boric: las principales propuestas de los rivales más antagónicos que ha tenido Chile en las últimas décadas. https://www.bbc.com/mundo/noticias-america-latina-59383712. [Online; accessed 19-May-2023].

[13] Morton Benson. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography* 3, 1 (1990), 23–34.

[14] Adrian Benton and Mark Dredze. 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.

[15] Peter Berger and Thomas Luckmann. 2016. The social construction of reality. In *Social theory re-wired*. Routledge, 110–122.

[16] Alessandro Bessi, Fabio Petroni, Michela Del Vicario, Fabiana Zollo, Aris Anagnostopoulos, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics* 225, 10 (2016), 2047–2059.

[17] BioBioChile. 2021. Siches denuncia ante Servel "campaña sucia" de Kast por video de Izquierdo llamando a "hacer trampa". https://www.biobiochile.cl/noticias/nacional/chile/2021/12/17/siches-denuncia-ante-servel-campana-sucia-de-kast-por-video-de-izquierdo-llamando-a-hacer-trampa.shtml. [Online; accessed 19-May-2023].

[18] Porismita Borah. 2011. Conceptual issues in framing theory: A systematic examination of a decade's literature. *Journal of communication* 61, 2 (2011), 246–263.

[19] 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*. 700–711.

[20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[21] Axel Bruns. 2022. El Filtro burbuja (The Filter Bubble). *Revista Latinoamericana de Economía y Sociedad Digital (RLESD)* Special Issue 1 (2022).

[22] Björn Burscher, Daan Odijk, Rens Vliegenthart, Maarten De Rijke, and Claes H De Vreese. 2014. Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures* 8, 3 (2014), 190–206.

[23] Bío Bío. 2021. "Chile merece la verdad": sale a la luz testimonio sobre presunto acoso de Boric. https://www.biobiochile.cl/noticias/nacional/chile/2021/12/13/chile-merece-la-verdad-sale-a-la-luz-testimonio-sobre-presunto-acoso-de-boric.shtml. [Online; accessed 28-September-2023].

[24] Bío Bío. 2021. Michelle Bachelet se define: "Yo voy a votar por Gabriel Boric". https://www.biobiochile.cl/noticias/nacional/chile/2021/12/14/michelle-bachelet-se-define-yo-voy-a-votar-por-gabriel-boric.shtml. [Online; accessed 28-September-2023].

[25] Nico Carpentier. 2011. *Media and participation: A site of ideological-democratic struggle*. Intellect.

[26] Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.* 10 (2007), 103–126.

[27] Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. 2017. A motif-based approach for identifying controversy. In *Eleventh International AAAI Conference on Web and Social Media*.

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

[29] Gary W Cox. 1990. Centripetal and centrifugal incentives in electoral systems. *American Journal of Political Science* (1990), 903–935.

[30] Kareem Darwish, Walid Magdy, Afshin Rahimi, Timothy Baldwin, and Norah Abokhodair. 2018. Predicting online islamophopic behavior after# parisattacks. *The Journal of Web Science* 4 (2018).

[31] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 141–152.

[32] Claes H De Vreese. 2005. News framing: Theory and typology. *Information design journal* 13, 1 (2005), 51–62.

[33] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 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.

[34] Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics* 8 (2020), 439–453.

[35] ElPaís. 2021. Chile consolida el primer Gobierno feminista latinoamericano. https://elpais.com/internacional/2022-03-11/la-consolidacion-del-primer-gobierno-feminista-de-chile.html. [Online; accessed 19-May-2023].

[36] Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory* (1993), 390–397.

[37] FastCheck. 2021. Programa presidencial de José Antonio Kast tiene un apartado que se llama "Coordinación Internacional Anti-Radicales de Izquierda". https://www.fastcheck.cl/2021/10/07/programa-presidencial-de-jose-antonio-kast-tiene-un-apartado-que-se-llama-coordinacion-internacional-anti-radicales-de-izquierda-real/. [Online; accessed 19-May-2023].

[38] William A Gamson and Andre Modigliani. 1989. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology* 95, 1 (1989), 1–37.

[39] Javier García-Marín and Adolfo Calatrava. 2018. The use of supervised learning algorithms in political communication and media studies: Locating frames in the press. *Communication & Society* 31, 3 (2018), 175–188.

[40] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[41] 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.

[42] Kiran Garimella and Gareth Tyson. 2018. WhatApp Doc? A First Look at WhatsApp Public Group Data. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1. https://doi.org/10.1609/icwsm.v12i1.14989

[43] Venkata Rama Kiran Garimella and Ingmar Weber. 2017. A long-term analysis of polarization on Twitter. In *Eleventh international AAAI conference on web and social media*.

[44] Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience.* Harvard University Press.

[45] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[46] LP Gudipaty and KY Jhala. 2015. Whatsapp forensics: decryption of encrypted whatsapp databases on non rooted android devices. *Journal Information Technology & Software Engineering* 5 (2015), 2.

[47] Pedro Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. 2013. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 215–224.

[48] Frederic Guerrero-Solé. 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 (2017), 244–261.

[49] Shohreh Haddadan, Elena Cabrio, Axel J Soto, and Serena Villata. 2022. Topic Modelling and Frame Identification for Political Arguments. In *International Conference of the Italian Association for Artificial Intelligence*. Springer, 268–281.

[50] Tobias Heidenreich, Fabienne Lind, Jakob-Moritz Eberl, and Hajo G Boomgaarden. 2019. Media framing dynamics of the 'European refugee crisis': A comparative topic modelling approach. *Journal of Refugee Studies* 32, Special_Issue_1 (2019), i172–i182.

[51] Nadia Herrada Hidalgo, Marcelo Santos, and Sérgio Barbosa. 2024. Affordances-driven ethics for research on mobile instant messaging: Notes from the Global South. *Mobile Media & Communication* (2024), 20501579241247994.

[52] Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems* 23 (2010).

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

[54] Shanto Iyengar. 1994. *Is anyone responsible?: How television frames political issues.* University of Chicago Press.

[55] Henry Jenkins and Mark Deuze. 2008. Convergence culture. , 5–12 pages.

[56] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[57] Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2010. Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 conference short papers*. 253–257.

[58] Anne C Kroon, Toni van der Meer, and Rens Vliegenthart. 2022. Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research* 4, 2 (2022), 528–570.

[59] Mucahid Kutlu, Kareem Darwish, Cansin Bayrak, Ammar Rashed, and Tamer Elsayed. 2019. Embedding-based qualitative analysis of polarization in turkey. *arXiv preprint arXiv:1909.10213* (2019).

[60] George Lakoff. 2014. *The all new don't think of an elephant!: Know your values and frame the debate.* Chelsea Green Publishing.

[61] Sophie Lecheler and Claes H De Vreese. 2019. *News framing effects: Theory and practice.* Taylor & Francis.

[62] Chang Li, Aldo Porco, and Dan Goldwasser. 2018. Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3728–3739.

[63] Hang Li. 2022. *Learning to rank for information retrieval and natural language processing.* Springer Nature.

[64] Antonis Matakos, Evimaria Terzi, and Panayiotis Tsaparas. 2017. Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery* 31 (2017), 1480–1505.

[65] Janne Tapani Matikainen. 2015. Motivations for content generation in social media. *Participations: Journal of Audience and Reception Studies* (2015).

[66] Jörg Matthes. 2009. What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & mass communication quarterly* 86, 2 (2009), 349–367.

[67] Jörg Matthes and Matthias Kohring. 2008. The content analysis of media frames: Toward improving reliability and validity. *Journal of communication* 58, 2 (2008), 258–279.

[68] Nolan McCarty, Keith T Poole, and Howard Rosenthal. 2016. *Polarized America: The dance of ideology and unequal riches.* mit Press.

[69] Maxwell McCombs and Sebastian Valenzuela. 2020. *Setting the agenda: Mass media and public opinion.* John Wiley & Sons.

[70] Marcelo Mendoza, Sebastián Valenzuela, Enrique Núñez-Mussa, Fabián Padilla, Eliana Providel, Sebastián Campos, Renato Bassi, Andrea Riquelme, Valeria Aldana, and Claudia López. 2023. A Study on Information Disorders on Social Networks during the Chilean Social Outbreak and COVID-19 Pandemic. *Applied Sciences* 13, 9 (2023), 5347.

[71] Sharon Meraz and Zizi Papacharissi. 2013. Networked gatekeeping and networked framing on# Egypt. *The international journal of press/politics* 18, 2 (2013), 138–166.

[72] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[73] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 31–41.

[74] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* 29, 3 (2013), 436–465.

[75] El Mostrador. 2021. Giorgio Jackson llama a adherentes de Boric a que se inscriban como apoderados de mesa: faltan 8 mil. https://www.elmostrador.cl/elecciones-2021/2021/12/17/giorgio-jackson-llama-a-adherentes-de-boric-a-que-se-inscriban-como-apoderados-de-mesa-faltan-8-mil/. [Online; accessed 28-September-2023].

[76] El Mostrador. 2021. La propaganda sucia . https://www.elmostrador.cl/destacado/2021/12/15/la-propaganda-sucia/. [Online; accessed 19-May-2023].

[77] Kay L O'Halloran, Gautam Pal, and Minhao Jin. 2021. Multimodal approach to analysing big social and news media data. *Discourse, Context & Media* 40 (2021), 100467.

[78] Sergey Pashakhin. 2016. Topic modeling for frame analysis of news media. *Proceedings of the AINL FRUCT* (2016), 103–105.

[79] El País. 2021. José Antonio Kast, el católico de extrema derecha que seduce a Chile. https://elpais.com/chile/2023-05-14/jose-antonio-kast-el-catolico-de-extrema-derecha-que-seduce-a-chile.html. [Online; accessed 28-September-2023].

[80] Giulio Rossetti, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. 2017. Tiles: an online algorithm for community discovery in dynamic social networks. *Machine Learning* 106 (2017), 1213–1241.

[81] Hernan Sarmiento, Felipe Bravo-Marquez, Eduardo Graells-Garrido, and Barbara Poblete. 2022. Identifying and Characterizing New Expressions of Community Framing during Polarization. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 841–851.

[82] Giovanni Sartori. 2005. *Parties and party systems: A framework for analysis.* ECPR press.

[83] Mirko Tobias Schäfer. 2011. *Bastard culture! How user participation transforms cultural production.* Amsterdam University Press.

[84] Dietram A Scheufele. 1999. Framing as a theory of media effects. *Journal of communication* 49, 1 (1999), 103–122.

[85] Holli A Semetko and Patti M Valkenburg. 2000. Framing European politics: A content analysis of press and television news. *Journal of communication* 50, 2 (2000), 93–109.

[86] Violeta Seretan. 2011. *Syntax-based collocation extraction.* Springer Dordrecht.

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

[88] Chris Sweeney and Maryam Najafian. 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.

[89] TerceraDosis. 2021. Bots x Kast. https://terceradosis.cl/2021/11/17/bots-x-kast/. [Online; accessed 19-May-2023].

[90] TerceraDosis. 2021. Cómo los perfiles automatizados intoxican la presidencial. https://terceradosis.cl/2021/12/16/como-los-perfiles-automatizados-intoxican-la-presidencial-en-bots-kast-supera-ampliamente-a-boric/. [Online; accessed 19-May-2023].

[91] Chau Tong, Hyungjin Gill, Jianing Li, Sebastián Valenzuela, and Hernando Rojas. 2021. "Fake news is anything they say!"—Conceptualization and weaponization of fake news among the American public. In *What IS News?* Routledge, 153–176.

[92] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).

[93] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science* 211, 4481 (1981), 453–458.

[94] Radio UChile. 2021. Movimiento por el Agua y los Territorios llama a votar por Boric. https://radio.uchile.cl/2021/11/30/movimiento-por-el-agua-y-los-territorios-llama-a-votar-por-boric/. [Online; accessed 28-September-2023].

[95] Sebastián Valenzuela, Yonghwan Kim, and Homero Gil de Zúñiga. 2012. Social networks that matter: Exploring the role of political discussion for online political participation. *International journal of public opinion research* 24, 2 (2012), 163–184.

[96] Mikko Villi and Janne Matikainen. 2016. Participation in social media: Studying explicit and implicit forms of participation in communicative social networks. *Media and communication* 4, 4 (2016), 109–117.

[97] Chong Wang, John Paisley, and David M Blei. 2011. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 752–760.

[98] Ingmar Weber, Venkata R Kiran Garimella, and Alaa Batayneh. 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.

[99] Jierui Xie, Boleslaw K Szymanski, and Xiaoming Liu. 2011. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *2011 ieee 11th international conference on data mining workshops*. IEEE, 344–349.

[100] Sarita Yardi and Danah Boyd. 2010. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of science, technology & society* 30, 5 (2010), 316–327.

[101] Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2018. Topic Modeling as a Method for Frame Analysis: Data Mining the Climate Change Debate in India and the USA. (2018).

[102] Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2020. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media and Communication* (2020), 17427665211023984.

[103] Renbo Zhao and Vincent YF Tan. 2016. Online nonnegative matrix factorization with outliers. *IEEE Transactions on Signal Processing* 65, 3 (2016), 555–570.

[104] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009